CLiC-it 2016

The Third Italian Conference on Computational Linguistics

Proceedings of the Conference

December 5-6, 2016

Copyright ©2016 for the individual papers by the papers' authors.
Copyright ©2016 for the volume as a collection by its editors.
This volume and its papers are published under the Creative Commons License Attribution 4.0 International (CC BY 4.0).

These proceedings have been published in the CEUR Workshop Proceedings series.
The original papers are available at: https://ceur-ws.org/Vol-1749.
The papers are mirrored in the ACL Anthology.

ISBN 978-88-99982-08-9

Table of Contents

Preface
Anna Corazza, Simonetta Montemagni and Giovanni Semeraro 1

Reflections on Universal Dependencies
Joakim Nivre 10

What's this movie about? Automatic Content Analysis and Summarization
Mirella Lapata 11

Determining the Compositionality of Noun-Adjective Pairs with Lexical Variants and Distributional Semantics
Marco S. G. Senaldi, Gianluca E. Lebani and Alessandro Lenci 12

Studio sull'Ordine dei Costituenti nel Confronto tra Generi e Complessità (Analysis of Constituents Order Across Textual Genres and Complexity)
Giulia Pieri, Dominique Brunato and Felice Dell'Orletta 18

Grounding the Lexical Sets of Causative-Inchoative Verbs with Word Embedding
Edoardo Maria Ponti, Elisabetta Jezek and Bernardo Magnini 23

Dieci sfumature di marcatezza sintattica: verso una nozione computazionale di complessità (Ten Shades of Syntactic Markedness: Towards a Computational Notion of Complexity)
Erica Tusa, Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi 29

Online Automatic Post-Editing across Domains
Rajen Chatterjee, Gebremedhen Gebremelak, Matteo Negri and Marco Turchi 35

Language Resources for Italian: towards the Development of a Corpus of Annotated Italian Multiword Expressions
Shiva Taslimipoor, Anna Desantis, Manuela Cherchi, Ruslan Mitkov and Johanna Monti 41

Emojitalianobot and EmojiWorldBot - New online Tools and Digital Environments for Translation into Emoji
Johanna Monti, Federico Sangati, Francesca Chiusaroli, Martin Benjamin and Sina Mansour . 47

IR Scientific Data: How to Semantically Represent and Enrich Them
Toine Bogers, Georgeta Bordea, Paul Buitelaar, Nicola Ferro and Gianmaria Silvello 52

Nyström Methods for Efficient Kernel-Based Methods for Community Question Answering
Danilo Croce, Simone Filice and Roberto Basili 58

Topic Modelling with Word Embeddings
Fabrizio Esposito, Anna Corazza and Francesco Cutugno 64

Imparare a Quantificare Guardando (Learning to Quantify by Watching)
Sandro Pezzelle, Ionut Sorodoc, Aurelie Herbelot and Raffaella Bernardi 70

Appetitoso: A Search Engine for Restaurant Retrieval based on Dishes
Gianni Barlacchi, Azad Abad, Emanuele Rossinelli and Alessandro Moschitti 76

Gamification for IR: The Query Aspects Game
Giorgio Maria Di Nunzio, Maria Maistro and Daniel Zilio 81

Semantic Interpretation of Events in Live Soccer Commentaries
Anne-Lyse Minard, Manuela Speranza, Bernardo Magnini and Mohammed R. H. Qwaider . . . 87

Linking IMAGACT Ontology to BabelNet through Action Videos
Lorenzo Gregori, Alessandro Panunzi and Andrea Amelio Ravelli 93

Relation Mining from Clinical Records
Anita Alicante, Anna Corazza, Francesco Isgrò and Stefano Silvestri 99

Subjective Well-Being and Social Media: A Semantically Annotated Twitter Corpus on Fertility and Parenthood
Emilio Sulis, Cristina Bosco, Viviana Patti, Mirko Lai, Delia Irazú Hernández Farías, Letizia Mencarini, Michele Mozzachiodi and Daniele Vignoli 104

An Extended Version of the KoKo German L1 Learner Corpus
Andrea Abel, Aivars Glaznieks, Lionel Nicolas and Egon W. Stemle 110

Annotating Content Zones in News Articles
Daniela Baiamonte, Tommaso Caselli and Irina Prodanof 116

LICO: A Lexicon of Italian Connectives
Anna Feltracco, Elisabetta Jezek, Bernardo Magnini and Manfred Stede 122

Formatio formosa est. Building a Word Formation Lexicon for Latin
Eleonora Litta, Marco Passarotti and Chris Culy 127

SIMPITIKI: a Simplification corpus for Italian
Sara Tonelli, Alessio Palmero Aprosio and Francesca Saltori 132

(Almost) Automatic Conversion of the Venice Italian Treebank into the Merged Italian Dependency Treebank Format
Linda Alfieri and Fabio Tamburini 138

Enrichring the Ita-TimeBank with Narrative Containers
Alice Bracchi, Tommaso Caselli and Irina Prodanof 143

Increasing Information Accessibility on the Web: a Rating System for Specialized Dictionaries
Valeria Caruso, Anna De Meo and Vincenzo Norman Vitale 149

Parsing di Corpora di Apprendenti di Italiano: un Primo Studio su VALICO (Parsing Italian Learner Corpora: a Case Study on VALICO)
Elisa Corino and Claudio Russo 154

Spammare senza pietà - Corpus based analysis of English, unacclimatised verb Loans in Italian and Creation of a Reference Lexicon
Anna Fantini 160

Taming Sense Sparsity: a Common-Sense Approach
Antonio Lieto, Enrico Mensa and Daniele P. Radicioni 166

Building a Computational Lexicon by using SQL
Alessandro Mazzei 172

FB-NEWS15: A Topic-Annotated Facebook Corpus for Emotion Detection and Sentiment Analysis
Lucia C. Passaro, Alessandro Bondielli and Alessandro Lenci 177

Sardinian on Facebook: Analysing Diatopic Varieties through Translated Lexical Lists
Irene Russo, Simone Pisano and Claudia Soria 182

A New Dataset for Source Code Comment Coherence
Anna Corazza, Valerio Maggio and Giuseppe Scanniello 187

Tracing Metaphors in Time through Self-Distance in Vector Spaces
Marco Del Tredici, Malvina Nissim and Andrea Zaninello 192

KD Strikes Back: from Keyphrases to Labelled Domains Using External Knowledge Sources
Giovanni Moretti, Rachele Sprugnoli and Sara Tonelli 198

Investigating the Application of Distributional Semantics to Stylometry
Giulia Benotto, Emiliano Giovannetti and Simone Marchi 204

Diachronic Analysis of the Italian Language exploiting Google Ngram
Pierpaolo Basile, Annalina Caputo, Roberta Luisi and Giovanni Semeraro 209

Hybrid Language Segmentation for Historical Documents
David Alfter and Yuri Bizzoni 214

La Leggibilità dei Testi di Ambito Medico rivolti al Paziente: il Caso dei Bugiardini di Farmaci Senza Obbligo di Prescrizione Medica (Readability of Medical Texts for Patients: a Case Study on Information Leaflets of Drugs without Obligation of Medical Prescription)
Franca Orletti, Felice Dell'Orletta and Rossella Iovino 219

La Modellazione Diacronica di Risorse Termino-Ontologiche nell'Ambito delle Digital Humanities: Esperimenti su Clavius (Diachronic Modeling of Terminological-Ontological Resources within Digital Humanities: Experiments on Clavius)
Silvia Piccini, Andrea Bellandi, Giulia Benotto and Emiliano Giovannetti 225

Panta Rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek
Martina A. Rodda, Marco S. G. Senaldi and Alessandro Lenci 230

Reassessing Inflectional Regularity in Modern Greek Conjugation
Stavros Bompolas, Marcello Ferro, Claudia Marzi, Franco Alberto Cardillo and Vito Pirrelli 235

Written Word Production and Lexical Self-Organisation: Evidence from English (Pseudo)Compounds
Marcello Ferro, Franco Alberto Cardillo, Vito Pirrelli, Christina L. Gagné and Thomas L. Spalding 241

Lexical Categories or Frequency Effects? A Feedback from Quantitative Methods Applied to Psycholinguistic Models in two Studies on Italian
Francesca Franzon, Giorgio Arcara and Chiara Zanini 247

Semantic Priming Effects in Italian Verbs Recognition: the Role of Grammatical Classes and Semantic Categories
Azzurra Mancuso, Maria De Martino and Alessandro Laudanna 252

Sentiment Analysis: Applicazione in un Dominio Psico-Forense
Francesca Guglielmi, Pierpaolo Basile, Antonietta Curci and Giovanni Semeraro 257

Stepping out of the Chinese Room: Word meaning with and without consciousness
Roberto Bottini, Daniel Casasanto, Andrea Nadalini and Davide Crepaldi 262

Context-aware Spoken Language Understanding for Human Robot Interaction
Andrea Vanzo, Danilo Croce, Roberto Basili and Daniele Nardi 267

Argument Mining on Italian News Blogs
Pierpaolo Basile, Valerio Basile, Elena Cabrio and Serena Villata 273

Twitter Sentiment Polarity Classification using Barrier Features
Anita Alicante, Anna Corazza and Antonio Pironti 278

The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts
Jennifer-Carmen Frey, Aivars Glaznieks and Egon W. Stemle 284

Sequenze N+pN (Nome Comune + Nome Proprio): Descrizione Linguistica da un Corpus dell'Italiano (N+pN (Common Noun + Proper Noun) Sequences: Linguistic Description from an Italian Corpus)
Felicia Logozzo 289

(Better than) State-of-the-Art PoS-tagging for Italian Texts
Fabio Tamburini 294

Tree Kernels-based Discriminative Reranker for Italian Constituency Parsers
Antonio Uva and Alessandro Moschitti 299

Comparing State-of-the-art Dependency Parsers on the Italian Stanford Dependency Treebank
Alberto Lavelli 304

May the Goddess of Hope Help Us. Homonymy in Latin Lexicon and Onomasticon
Marco Passarotti and Marco Budassi 310

Preface

Our very warm welcome to CLiC-it 2016 (http://clic-it2016.dieti.unina.it/), the 3rd edition of the *Italian Conference on Computational Linguistics*, held on December 5th and 6th, in Naples, Italy, co-located with Evalita 2016 (http://www.evalita.it/2016), hosted and locally organized by Università Federico II, one the oldest public and laic universities in the world. The organization of the conference is the result of a fruitful conjoint effort of different research groups (Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR, Università degli Studi di Bari Aldo Moro and Università degli Studi di Napoli Federico II) showing the nationwide spreading of Computational Linguistics in Italy. The CLiC-it conference series is organized by the *Italian Association for Computational Linguistics* (AILC) and has clearly established itself as the premier national forum for research and development in the fields of Computational Linguistics (CL) and Natural Language Processing (NLP), where leading researchers and practitioners from academia and industry meet to share their challenges, solutions, research results, and experiences.

CLiC-it covers all aspects of computational linguistics and natural language (both written and spoken) processing, and targets state-of-art theoretical results, experimental methodologies, technologies, as well as application perspectives, which may contribute to advance the field.

As in the previous editions, CLiC-it 2016 is organized around thematic areas, each chaired by two or more Area Chairs:

- *Cognitive modeling of language processing and psycholinguistics*. Area chairs: Davide Crepaldi, (Scuola Internazionale Superiore di Studi Avanzati, Trieste), Gianluca Lebani (Università degli Studi di Pisa), Vito Pirrelli (Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa)
- *NLP for Digital Humanities*. Area chairs: Marco Passarotti (Università Cattolica del Sacro Cuore, Milano), Sara Tonelli (Fondazione Bruno Kessler, Trento)
- *Information Retrieval and Question Answering*. Area chairs: Nicola Ferro (Università degli Studi di Padova), Alessandro Moschitti (Università degli Studi di Trento - Qatar Computing Research Institute)
- *Information Extraction, Entity Linking and (Linked) Open Data*. Area chairs: Valerio Basile (INRIA Sophia Antipolis Méditerranée), Roberto Navigli (Università degli Studi "La Sapienza" di Roma)
- *Linguistic Resources*. Area chairs: Cristina Bosco (Università degli Studi di Torino), Monica Monachini (Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR, Pisa), Simonetta Vietri (Università degli Studi di Salerno)
- *Machine Translation and Multilingual Applications*. Area chairs: Mauro Cettolo (Fondazione Bruno Kessler, Trento), Johanna Monti (Università degli Studi di Sassari)
- *Pragmatics, Creativity and Linguistic Games*. Area chairs: *Marco de Gemmis*, Università degli Studi di Bari "Aldo Moro", Marco Gori (Università degli Studi di Siena), Massimo Poesio (Sussex University, UK)
- *Semantics and Knowledge Acquisition*. Area chairs: Raffaella Bernardi (Università degli Studi di Trento), Aldo Gangemi (Istituto di Scienze e Tecnologie della Cognizione, CNR, Roma), Fabio Massimo Zanzotto (Università degli Studi di Roma Tor Vergata)
- *Spoken language processing and understanding*. Area chairs: Piero Cosi (Istituto di Scienze e Tecnologie della Cognizione, CNR, Padova), Antonio Origlia (Università degli Studi di Napoli Federico II)

- *Morphology and Syntax Processing*. Area chairs: Alberto Lavelli (Fondazione Bruno Kessler, Trento), Alessandro Mazzei (Università degli Studi di Torino)
- *NLP for Web and Social Media*. Area chairs: Danilo Croce (Università degli Studi di Roma Tor Vergata), Felice Dell'Orletta (Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa)
- *Linguistic Issues in CL and NLP*. Area chairs: Alessandro Lenci (Università degli Studi di Pisa), Paola Merlo (University of Geneva, CH)
- *Machine Learning for CL and NLP*. Area chairs: Giuseppe Attardi (Università degli Studi di Pisa), Roberto Basili (Università degli Studi di Roma Tor Vergata)

The maturity of the conference is clearly demonstrated by the high quality of the submitted research work. This year CLiC-it received 69 submissions by 212 authors from 14 countries and the Program Committee worked very hard to ensure that every paper got at least two careful and fair reviews, with the 74% of the papers which received three or even more reviews. This process finally led to the acceptance of 28 papers for oral presentation and 27 papers for poster presentation, with a global acceptance rate of 80% motivated by the inclusive spirit of the conference and which is in line with the previous editions (77% in 2014 and 81% in 2015). Regardless of the format of presentation, all accepted papers were allocated 6 pages in the proceedings, available as "open access" publication. We are therefore extremely grateful to our 152 Program Committee members and to our 3 additional reviewers for producing more than 206 detailed and insightful reviews, as well as to the Area Chairs, who assisted the Program Chairs in their work.

In 2016, the submissions were articulated into 13 thematic areas, some of which shared with the previous year(s), others which were introduced for the first time. The area of *Linguistic Resources* was confirmed as the track attracting the higher number of submissions: the papers in this area cover 32% of the submissions, against 29% in 2014 and 25% in 2015. In line with international computational linguistics conferences such as COLING 2014, this year two new tracks were introduced tackling transversal issues: namely, *Linguistic Issues in CL and NLP* and *Machine Learning for CL and NLP*, which attracted - together - almost 12% of the submitted papers. The track *Cognitive modeling of language processing and psycholinguistics*, new with respect to 2015 but already present in CLiC-it 2014, represents more than 8% of the submissions: interestingly, they originate also from research groups working in the area of cognitive psychology and neuroscience, thus complying with the conference goal of bridging the gap between the results emerging in different areas of computational linguistics and other related disciplines.

The remaining tracks represent, with minor adjustments, a continuation of the thematic areas of 2015 and 2014 editions. Here, the main difference lies at the level of the number of attracted submissions with respect to the total, which range from the 69% and 75% of 2014 and 2015 respectively, to the 48% of this year. This significant reduction of submissions can mainly be explained by the co-location of the conference with the Evalita final Workshop, whose challenges coincide with central topics of these areas: see, for instance, the track *NLP for Web and Social Media* and the SENTIPOLC (SENTIment POLarity Classification) Evalita shared task, or the track *Morphology and Syntax Processing* and the PoSTWITA (*POS tagging for Italian Social Media Texts*) task. The complementarity of the contributions of the two conferences emerges clearly by comparing the author keywords associated with individual CLiC-it and Evalita papers. It is interesting to note that key hot topics of these areas, such as Deep Learning, Named Entity Linking and Recognition, Part-of-speech tagging or Question Answering to mention just a few, are more frequently associated to papers in Evalita's rather than CLiC-it's proceedings: in some cases, they do not appear at all, as in the

case of Deep Learning which is not among the author keywords associated with CLiC-it papers.

The rich and articulated picture emerging from the wide variety of contributions to CLiC-it 2016 creates the prerequisites for a fruitful and stimulating conference. Therefore, we would like to thank the authors of all papers for submitting their work to the conference. Among them, many are young authors (PhD students and Postdocs). As in the previous editions, Young Author Best Paper Awards will be assigned to meritorious papers involving PhD students and Postdocs.

Besides the technical paper program consisting of oral presentations and two poster sessions, the core conference program also includes two keynote speeches, by Mirella Lapata (University of Edinburgh) and Joakim Nivre (Uppsala University). We are honoured they accepted to contribute to CLiC-it 2016 and thank them for agreeing to share their knowledge and expertise on key computational linguistics topics with the Italian community. Last but not least, the program also includes two panels, focusing on the impact to research results on the economy and society at large and aimed at creating a national and international network around the Italian computational linguistics community. Our thanks also go to the panelists who accepted to be involved in the event.

We also would like to take this opportunity to thank all our colleagues, who volunteered their time to contribute to the success of this conference, as well as to acknowledge the support from endorsing organizations and institutions and from our sponsors, who generously provided funds and services that are crucial for the organization of this event. At the time of writing this preface, CLiC-it was generously sponsored by the following companies and associations: Interactive Media s.p.a. (Gold Sponsor); CELI Language Technology and GruppoMeta (Silver Sponsors); the European Linguistic Resources Association (ELRA), NTT DATA and QuestionCube (Bronze Sponsors). We also would like to thank the following organizations for endorsing CLiC-it 2016: Associazione Italiana per l'Intelligenza Artificiale (AI*IA); Associazione Italiana Scienze della Voce (AISV); Associazione Italiana di Linguistica Applicata (AItLA); Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD); Società Italiana di Glottologia (SIG); Società di Linguistica Italiana (SLI). Last but not least, special thanks are also due to Università di Napoli Federico II and to the Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione of this university for the support in the organization of the event and hosting the conference.

This conference would also not be possible without the dedication, devotion and hard work of the members of our local organizing committee. Our special thanks extend to all of them. Finally, we want to acknowledge the EasyChair infrastructure for the management of the review process.

Please join us at CLiC-it 2016 to interact with experts from academia and industry on topics related to Computational Linguistics and Natural Language Processing and to experience and share new research findings, best practices, state-of-the-art systems and applications.

Anna Corazza, Simonetta Montemagni and Giovanni Semeraro
CLiC-it 2016 Conference Co-Chairs

November 2016

Committees

Conference Chairs

- Anna Corazza, Università degli Studi di Napoli Federico II
- Simonetta Montemagni, Istituto di Linguistica Computazionale "Antonio Zampolli", CNR
- Giovanni Semeraro, Università degli Studi di Bari "Aldo Moro"

Program Committee

Cognitive modeling of language processing and psycholinguistics.

Area chairs:

- Davide Crepaldi, Scuola Internazionale Superiore di Studi Avanzati, Trieste
- Gianluca Lebani, Università degli Studi di Pisa
- Vito Pirrelli, Istituto di Linguistica Computazionale "Antonio Zampolli", CNR

Reviewers:

- Valentina Bambini, Istituto Universitario di Studi Superiori di Pavia
- Anna Borghi, Università degli Studi di Bologna - Istituto di Scienze e Tecnologie della Cognizione, CNR
- Roberto Bottini, Università degli Studi di Trento
- Basilio Calderone, CNRS - Université de Toulouse-Jean Jaurès
- Cristiano Chesi, Istituto Universitario di Studi Superiori di Pavia
- Marcello Ferro, Istituto di Linguistica Computazionale "Antonio Zampolli", CNR
- Marco Marelli, Ghent University
- Claudia Marzi, Istituto di Linguistica Computazionale "Antonio Zampolli", CNR
- Claudio Mulatti, Università degli Studi di Padova
- Daniela Paolieri, Universidad de Granada
- Lucia Passaro, Università degli Studi di Pisa
- Giovanni Pezzulo, Istituto di Scienze e Tecnologie della Cognizione, CNR
- Valentina Schettino, Università degli Studi di Napoli Federico II
- Marco Senaldi, Scuola Normale Superiore, Pisa
- Simone Sulpizio, Università degli Studi di Trento
- Eva Maria Vecchi, University of Cambridge
- Francesco Vespignani, Università degli Studi di Trento
- Alessandra Zarcone, Universität des Saarlandes

Information Extraction, Entity Linking and (Linked) Open Data

Area chairs:

- Valerio Basile, INRIA
- Roberto Navigli, Università degli Studi "La Sapienza" di Roma

Reviewers:

- Gianni Barlacchi, Università degli Studi di Trento
- Pierpaolo Basile, Università degli Studi di Bari "Aldo Moro"
- Annalina Caputo, Trinity College Dublin
- Giuseppe Castellucci, Università degli Studi di Roma Tor Vergata
- Danilo Croce, Università degli Studi di Roma Tor Vergata
- Stefano Faralli, University of Mannheim

- Aldo Gangemi, Université Paris 13 - Istituto di Scienze e Tecnologie della Cognizione, CNR
- Pasquale Lops, Università degli Studi di Bari "Aldo Moro"
- Cataldo Musto, Università degli Studi di Bari "Aldo Moro"
- Viviana Patti, Università degli Studi di Torino
- Simone Paolo Ponzetto, University of Mannheim
- Giuseppe Rizzo, Istituto Superiore Mario Boella
- Paolo Rosso, Technical University of Valencia
- Fabrizio Silvestri, Facebook
- Marco Turchi, Fondazione Bruno Kessler, Trento

Information Retrieval and Question Answering

Area chairs:

- Nicola Ferro, Università degli Studi di Padova
- Alessandro Moschitti, Università degli Studi di Trento - Qatar Computing Research Institute

Reviewers:

- Gianluca Demartini, University of Sheffield
- Claudio Lucchese, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR
- Raffaele Perego, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR
- Paolo Rosso, Technical University of Valencia
- Fabrizio Silvestri, Facebook
- Guido Zuccon, Queensland University of Technology

Linguistic Issues in CL and NLP

Area chairs:

- Alessandro Lenci, Università degli Studi di Pisa
- Paola Merlo, University of Geneva, CH

Reviewers:

- Marco Baroni, Università degli Studi di Trento
- Marie Candito, Université Paris Diderot - INRIA
- Barbara Di Eugenio, University of Illinois at Chicago
- Francesca Masini, Università degli Studi di Bologna
- Suzanne Stevenson, University of Toronto
- Roberto Zamparelli, Università degli Studi di Trento

Linguistic Resources

Area chairs:

- Cristina Bosco, Università degli Studi di Torino
- Monica Monachini, Istituto di Linguistica Computazionale "Antonio Zampolli", CNR
- Simonetta Vietri, Università degli Studi di Salerno

Reviewers:

- Núria Bel, Universitat Pompeu Fabra
- Luisa Bentivogli, Fondazione Bruno Kessler, Trento
- Johan Bos, University of Groningen

- Nicoletta Calzolari, Istituto di Linguistica Computazionale “Antonio Zampolli”, CNR
- Tommaso Caselli, Vrije Universiteit Amsterdam
- Isabella Chiari, Università degli Studi “La Sapienza” di Roma
- Elisa Corino, Università degli Studi di Torino
- Thierry Declerck, DFKI GmbH
- Francesca Frontini, Istituto di Linguistica Computazionale “Antonio Zampolli”, CNR
- Aldo Gangemi, Université Paris 13 - Istituto di Scienze e Tecnologie della Cognizione, CNR
- Elisabetta Jezek, Università degli Studi di Pavia
- Fahad Khan, Istituto di Linguistica Computazionale “Antonio Zampolli”, CNR
- Borja Navarro-Colorado, Universidad de Alicante
- Malvina Nissim, University of Groningen
- Irene Russo, Istituto di Linguistica Computazionale “Antonio Zampolli”, CNR
- Maria Simi, Università degli Studi di Pisa
- Claudia Soria, Istituto di Linguistica Computazionale “Antonio Zampolli”, CNR
- Laure Vieu, Institut de Recherche en Informatique de Toulouse

Machine Learning for CL and NLP

Area chairs:

- Giuseppe Attardi, Università degli Studi di Pisa
- Roberto Basili, Università degli Studi di Roma Tor Vergata

Reviewers:

- Pierpaolo Basile, Dipartimento di Informatica - University of Bari
- Daniele Bonadiman, Università degli Studi di Trento
- Francesco Cutugno, Università degli Studi di Napoli Federico II
- Simone Filice, Università degli Studi di Roma Tor Vergata
- Bernardo Magnini, Fondazione Bruno Kessler, Trento
- Alessandro Moschitti, Università degli Studi di Trento - Qatar Computing Research Institute
- Malvina Nissim, University of Groningen
- Antonio Origlia, Università degli Studi di Napoli Federico II
- Maria Simi, Università degli Studi di Pisa
- Fabio Tamburini, Università degli Studi di Bologna

Machine Translation and Multilingual Applications

Area chairs:

- Mauro Cettolo, Fondazione Bruno Kessler, Trento
- Johanna Monti, Università degli Studi di Sassari

Reviewers:

- Anabela Barreiro, INESC ID
- Nicola Bertoldi, Fondazione Bruno Kessler, Trento
- José Guilherme Camargo de Souza, eBay Inc
- Marcello Federico, Fondazione Bruno Kessler, Trento
- Matteo Negri, Fondazione Bruno Kessler, Trento
- Bruno Pouliquen, World Intellectual Property Organization
- Marco Turchi, Fondazione Bruno Kessler, Trento

Morphology and Syntax Processing

Area chairs:

- Alberto Lavelli, Fondazione Bruno Kessler, Trento
- Alessandro Mazzei, Università degli Studi di Torino

Reviewers:

- Giuseppe Attardi, Università degli Studi di Pisa
- Bernd Bohnet, Google
- Cristiano Chesi, Istituto Universitario di Studi Superiori di Pavia
- Felice Dell'Orletta, Istituto di Linguistica Computazionale "Antonio Zampolli", CNR
- Albert Gatt, University of Malta
- Vincenzo Lombardo, Università degli Studi di Torino
- Daniele Radicioni, Università degli Studi di Torino
- Francesco Sartorio, Università degli Studi di Padova
- Giorgio Satta, Università degli Studi di Padova
- Fabio Tamburini, Università degli Studi di Bologna

NLP for Digital Humanities

Area chairs:

- Marco Passarotti, Università Cattolica del Sacro Cuore, Milano
- Sara Tonelli, Fondazione Bruno Kessler, Trento

Reviewers:

- Federico Boschetti, Istituto di Linguistica Computazionale "Antonio Zampolli", CNR
- Giovanni Colavizza, École Polytechnique Fédérale de Lausanne
- Antske Fokkens, Vrije Universiteit Amsterdam
- Emiliano Giovannetti, Istituto di Linguistica Computazionale "Antonio Zampolli", CNR
- Mike Kestemont, University of Antwerp
- John Lee City, University of Hong Kong
- Eleonora Litta Modignani, Università Cattolica del Sacro Cuore, Milano
- Francesco Mambrini, Deutsches Archäologisches Institut
- Stefano Menini, Fondazione Bruno Kessler - Università degli Studi di Trento
- Geoffrey Rockwell, University of Alberta
- Matteo Romanello, King's College London
- Rachele Sprugnoli, Fondazione Bruno Kessler - Università degli Studi di Trento

NLP for Web and Social Media

Area chairs:

- Danilo Croce, Università degli Studi di Roma Tor Vergata
- Felice Dell'Orletta, Istituto di Linguistica Computazionale "Antonio Zampolli", CNR

Reviewers:

- Costanza Asnaghi, Gent University - KU Leuven
- Francesco Barbieri, Universitat Pompeu Fabra
- Pierpaolo Basile, Università degli Studi di Bari "Aldo Moro"
- Valerio Basile, INRIA

- Dominique Brunato, Istituto di Linguistica Computazionale "Antonio Zampolli", CNR
- Annalina Caputo, Trinity College Dublin
- Giuseppe Castellucci, Università degli Studi di Roma Tor Vergata
- Francesca Chiusaroli, Università degli Studi di Macerata
- Andrea Cimino, Istituto di Linguistica Computazionale "Antonio Zampolli", CNR
- Stefano Cresci, Istituto di Informatica e Telematica, CNR
- Federica Da Milano, Università degli Studi di Milano-Bicocca
- Anna Feltracco, Fondazione Bruno Kessler, Trento
- Simone Filice, Università degli Studi di Roma Tor Vergata
- Lorenzo Gatti, Fondazione Bruno Kessler, Trento
- Nicola Grandi, Università degli Studi di Bologna
- Felicia Logozzo, Università degli Studi di Roma Tor Vergata
- Simone Magnolini, Fondazione Bruno Kessler, Trento
- Diego Marcheggiani, University of Amsterdam
- Alejandro Moreo Fernández, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR
- Cataldo Musto, Università degli Studi di Bari "Aldo Moro"
- Nicole Novielli, Università degli Studi di Bari "Aldo Moro"
- Viviana Patti, Università degli Studi di Torino
- Marco Pedicini, Università degli Studi Roma Tre
- Maria Laura Pierucci, Università degli Studi di Macerata
- Francesco Ronzano, Universitat Pompeu Fabra
- Maurizio Tesconi, Istituto di Informatica e Telematica, CNR
- Andrea Vanzo, Università degli Studi "La Sapienza" di Roma
- Giulia Venturi, Istituto di Linguistica Computazionale "Antonio Zampolli", CNR
- Torsten Zesch Sprachtechnologie, University of Duisburg-Essen

Pragmatics, Creativity and Linguistic Games

Area chairs:

- Marco de Gemmis, Università degli Studi di di Bari "Aldo Moro"
- Marco Gori, Università degli Studi di Siena
- Massimo Poesio, Sussex University, UK

Reviewers:

- Carla Bazzanella
- Gemma Boleda, Universitat Pompeu Fabra
- Francesca Bonin, Trinity College Dublin
- Rossana Damiano, Università degli Studi di Torino
- Rodolfo Delmonte
- Diana Inkpen, University of Ottawa
- Beata Klebanov
- Costanza Navarretta, University of Copenhagen
- Vincent Ng, University of Texas at Dallas
- Malvina Nissim, University of Groningen
- Viviana Patti, Università degli Studi di Torino
- Arndt Riester, University of Stuttgart
- Carlo Strapparava, Fondazione Bruno Kessler, Trento

Semantics and Knowledge Acquisition

Area chairs:

- Raffaella Bernardi, Università degli Studi di Trento
- Aldo Gangemi, Université Paris 13 - Istituto di Scienze e Tecnologie della Cognizione, CNR
- Fabio Massimo Zanzotto, Università degli Studi di Roma Tor Vergata

Reviewers:

- Maria Aloni, Universiteit van Amsterdam
- Elena Cabrio, University of Nice Sophia Antipolis
- Aurelie Herbelot, Università degli Studi di Trento
- Omer Levy, Bar-Ilan University
- Denis Paperno, Università degli Studi di Trento
- Diego Reforgiato Recupero, Università degli Studi di Cagliari

Spoken language processing and understanding

Area chairs:

- Piero Cosi, Istituto di Scienze e Tecnologie della Cognizione, CNR
- Antonio Origlia, Università degli Studi di Napoli Federico II

Reviewers:

- Francesco Cutugno, Università degli Studi di Napoli Federico II
- Bogdan Ludusan, LSCP - École Normale Supérieure - CNRS, Paris

Additional Reviewers

- Anna Lisa Gentile, Universität Mannheim
- Julia Bosque-Gil, Universidad Politécnica de Madrid
- Manuela Sanguinetti, Università degli Studi di Torino

Local Organizing Committee

- Anna Corazza
- Francesco Cutugno
- Francesco Isgrò
- Andrea Apicella
- Dario Di Mauro
- Fabrizio Esposito
- Antonio Origlia
- Valentina Schettino
- Giuseppe Vettigli

Invited speaker

Speaker: Joakim Nivre, Department of Linguistics and Philology - Uppsala University

Title: Reflections on Universal Dependencies

Abstract - Universal Dependencies is a framework for cross-linguistically consistent treebank annotation that has so far been applied to over 50 languages. It was developed primarily to support multilingual parsing research, but the resources have proven useful for a wide range of studies that were not foreseen originally, including research on language typology. In this talk, I will give a general introduction to Universal Dependencies and the resources developed in the project, and survey a number of representative examples of how the resources have been put to use in different scientific investigations. I will conclude with some reflections on how Universal Dependencies relate to linguistic theory, on the one hand, and parsing technology, on the other.

Invited speaker

Speaker: Mirella Lapata, Institute for Language, Cognition and Computation - School of Informatics - University of Edinburgh

Title: What's this movie about? Automatic Content Analysis and Summarization

Abstract - Movie analysis is an umbrella term for many tasks aiming to automatically interprete, extract, and summarize the content of a movie. Potential applications include generating shorter versions of scripts to help with the decision making process in a production company, enhancing movie recommendation engines by abstracting over specific keywords to more general concepts (e.g., thrillers with psychopaths), and notably generating movie previews. In this talk I will illustrate how NLP-based models together with video analysis can be used to facilitate various steps in the movie production pipeline. I will formalize the process of generating a shorter version of a movie as the task of finding an optimal chain of scenes and present a graph-based model that selects a chain by jointly optimizing its logical progression, diversity, and importance. I will then apply this framework to screenplay summarization, a task which could enhance script browsing and speed up reading time. I will also show that by aligning the screenplay to the movie, the model can generate movie previews with minimal modification. Finally, I will discuss how the computational analysis of movies can lead to tools that automatically create movie "profiles" which give a first impression of the movie by describing its plot, mood, location, or style.

Determining the Compositionality of Noun-Adjective Pairs with Lexical Variants and Distributional Semantics

Marco S. G. Senaldi
Laboratorio di Linguistica
Scuola Normale Superiore
Pisa, Italy
marco.senaldi@sns.it

Gianluca E. Lebani, Alessandro Lenci
Computational Linguistics Laboratory
Department of Philology, Literature, and Linguistics
University of Pisa, Italy
gianluca.lebani@for.unipi.it
alessandro.lenci@unipi.it

Abstract

English. In this work we employed a set of 26 Italian noun-adjective expressions to test compositionality indices that compare the distributional vector of an expression with the vectors of its lexical variants. These were obtained by replacing the components of the original expression with semantically related words. Our indices performed comparably or better than other compositionality measures reported in the distributional literature.

Italiano. *In questo lavoro si è utilizzato un set di 26 espressioni italiane nome-aggettivo per testare degli indici di composizionalità che confrontano il vettore distribuzionale di un'espressione con i vettori delle sue varianti lessicali. Queste sono state ottenute sostituendo i componenti dell'espressione di partenza con parole semanticamente correlate. La performance dei nostri indici si è dimostrata comparabile o superiore a quella di altri indici di composizionalità riportati nella letteratura distribuzionale.*

1 Introduction and previous research

While a *white car* is *white* and is a *car*, a *red herring* in a sentence like *I thought he was the culprit, but he was a red herring* is neither *red* nor a *herring*, but indicates something that distracts someone from a relevant issue. The former expression is compositional, since its meaning derives from the composition of the meanings of its subparts (Werning et al., 2012). The latter, by contrast, is an *idiom*, a non-compositional, figurative and proverbial word combination belonging to the wider class of *Multiword Expressions* (Nunberg et al., 1994; Cacciari, 2014). The compositionality of a given expression entails *salva-veritate*-interchangeability and systematicity (Fodor and Lepore, 2002). First of all, if we replace the constituents of a compositional expression with synonyms or similar words (e.g., from *white car* to *white automobile*), the whole meaning is not altered. Secondly, if we can understand the meaning of *white car* and *red herring* used in the literal sense, we can also understand what *white herring* and *red car* mean. Both these properties are not valid for idioms, which always exhibit lexical fixedness to some extent: variants of idiomatic *red herring* like *red fish* or *white herring* can just have a literal reading.

Computational studies to date have proposed several techniques to automatically measure idiomaticity. Of note, Lin (1999) and Fazly et al. (2009) label a given word combination as idiomatic if the Pointwise Mutual Information (PMI) (Church and Hanks, 1991) between its component words is higher than the PMIs between the components of a set of lexical variants of this combination. These variants are obtained by replacing the component words of the original expressions with semantically related words. Other researches have exploited Distributional Semantic Models (DSMs) (Sahlgren, 2008; Turney and Pantel, 2010), comparing the vector of a given phrase with the single vectors of its subparts (Baldwin et al., 2003; Venkatapathy and Joshi, 2005; Fazly and Stevenson, 2008) or comparing the vector of a phrase with the vector deriving from the sum or the products of their components (Mitchell and Lapata, 2010; Krčmář et al., 2013).

In a previous contribution (Senaldi et al., 2016),

we started from a set of Italian verbal idiomatic and non-idiomatic phrases (henceforth our *targets*) and generated lexical variants (simply *variants* henceforth) by replacing their components with semantic neighbours extracted from a linear DSM and Italian MultiWordNet (Pianta et al., 2002). Then, instead of measuring the associational scores between their subparts like in Lin (1999) and Fazly et al. (2009), we exploited Distributional Semantics to observe how different the context vectors of our targets were from the vectors of their variants. Our proposal stemmed from the consideration that a high PMI value does not necessarily imply the idiomatic or multiword status of an expression, but just that its components co-occur more frequently than expected by chance, as in the case of *read* and *book* or *solve* and *problem*, which are all instances of compositional pairings. By contrast, what watertightly distinguishes an idiomatic from a collocation-like yet still compositional expression is their context of use. Comparing the distributional contexts of the original expressions and their alternatives should therefore represent a more precise refinement of the PMI-based procedure. Actually, idiomatic expressions vectors were found to be less similar to their variants vectors with respect to compositional expressions vectors. In some of our models, we also kept track of the variants that were not attested in our corpus by representing them as orthogonal vectors to the vector of the original expression, still achieving considerable results. Noteworthily, most researches conducted so far have focused on verbal idioms, while the analysis of NP idioms like *red herring* or *second thoughts* has been usually left aside.

2 Applying variant-based distributional measures to noun-adjective pairs

In the present study, we firstly aimed to extend the variant-based method tested in Senaldi et al. (2016) on verbal idioms to noun-adjective expressions, which are mostly neglected in the idiom literature. In the second place, our former work lacked a comparison against conventional additive and multiplicative compositionality indices proposed in the distributional literature (Mitchell and Lapata, 2010; Krčmář et al., 2013). Finally, beside using a linear DSM and Italian MultiWordNet (Pianta et al., 2002) to extract our variants, we also experimented with a DSM (Padó and Lapata, 2007; Baroni and Lenci, 2010) which kept track of the syntactic dependency relations between a given target and its contexts.

3 Data extraction

3.1 Extracting the target expressions

All in all, our dataset was composed of 26 types of Italian noun-adjective and adjective-noun combinations. Of these, 13 were Italian idioms extracted from the itWaC corpus (Baroni et al., 2009), which totalizes about 1,909M tokens. The frequency of these targets varied from 21 (*alte sfere* 'high places', lit. 'high spheres') to 194 (*punto debole* 'weak point'). The remaining 13 items were compositional pairs of comparable frequencies (e.g., *nuova legge* 'new law').

3.2 Extracting lexical variants

Linear DSM variants. For both the noun and the adjective of our targets, we extracted its top cosine neighbors in a linear DSM created from the La Repubblica corpus (Baroni et al., 2004) (about 331M tokens). In Senaldi et al. (2016) we experimented with different thresholds of selected top neighbors (3, 4, 5 and 6). Since the number of top neighbors that were extracted for each constituent did not significantly affect our performances, for the present study we decided to use the maximum number (i.e., 6). All the content words occurring more than 100 times were represented as target vectors, ending up with 26,432 vectors, while the top 30,000 content words were used as dimensions. The co-occurrence counts were collected with a context window of $\pm$ 2 content words from each target word. The obtained matrix was then weighted by Positive Pointwise Mutual Information (PPMI) (Evert, 2008) and reduced to 300 latent dimensions via Singular Value Decomposition (SVD) (Deerwester et al., 1990). The variants were finally obtained by combining the adjective with each of the noun's top 6 neighbors, the noun with all the top 6 neighbors of the adjective and finally all the top 6 neighbors of the adjective and the noun with each other, ending up with 48 Linear DSM variants per target.

Structured DSM variants. While unstructured DSMs just record the words that linearly precede or follow a target lemma when collecting co-occurrence counts, structured DSMs conceive co-occurrences as $< w_1, r, w_2 >$ triples, where *r* rep-

resents the dependency relation between w_1 and w_2 (Padó and Lapata, 2007; Baroni and Lenci, 2010). Since we wanted to experiment with different kinds of distributional information to generate our variants, following the method described in Baroni and Lenci (2010) we created a structured DSM from La Repubblica (Baroni et al., 2004),where all the content words occurring more than 100 times were kept as targets and the co-occurrence matrix was once again weighted via PPMI and reduced to 300 latent dimensions. For each target, we generated 48 lexical variants with the same procedure described for the linear DSM variants.

iMWN variants. For each noun, we extracted the words occurring in the same synsets and its co-hyponyms in Italian MultiWordNet (iMWN) (Pianta et al., 2002). As for the adjectives, we experimented with two different approaches, extracting just their synonyms in the first case (iMWN$_{syn}$ variants) and adding also the antonyms in the second case (iMWN$_{ant}$ variants). The antonyms were translated from the English WordNet (Fellbaum, 1998). For each noun and adjective, we kept its top 6 iMWN neighbors in terms of cosine similarity in the same DSM used to acquire the linear DSM variants. Once again, this method provided us with 48 iMWN variants per target.

4 Gold standard idiomaticity judgments

To validate our computational indices, we presented 9 linguistics students with our 26 targets and asked them to rate how idiomatic each expression was on a 1-7 scale, with 1 standing for "totally compositional" and 7 for "totally idiomatic". The targets were presented in three different randomized orders, with three raters per order. The mean score given to our idioms was 6.10 (SD = 0.77), while the mean score given to compositional expressions was 2.03 (SD = 1.24). This difference was proven by a t-test to be statistically significant (t = 10.05, $p < 0.001$). Inter-coder reliability, measured via Krippendorff's α (Krippendorff, 2012) was 0.76. Following established practice, we took such value as a proof of reliability for the elicited ratings (Artstein and Poesio, 2008).

5 Calculating compositionality indices

For each of our 26 targets, we extracted from itWaC all the attested occurrences of the 48 linear DSM, structured DSM, iMWN$_{syn}$ and iMWN$_{ant}$ variants. We then computed two kinds of vector-based compositionality indices:

5.1 Variant-based indices

For every variant type (linear DSM, structured DSM, iMWN$_{syn}$ and iMWN$_{ant}$) we built a DSM from itWaC representing the 26 targets and their variants as vectors. While the dimension of the La Repubblica corpus seemed to be enough for the variants extraction procedure, we resorted to five-times bigger itWaC to represent the variants as vectors and compute the compositionality scores to avoid data sparseness and have a considerable number of variants frequently attested in our corpus. We also thought that using two different corpora had the additional advantage of showing the variants method to be generalizable to corpora of different text genres. Co-occurrence statistics recorded how many times each target or variant construction occurred in the same sentence with each of the 30,000 top content words in the corpus. The matrices were then weighted with PPMI and reduced to 150 dimensions via SVD. We finally calculated four different indices:

Mean. The mean cosine similarity between the vector of a target construction and the vectors of its variants.

Max. The maximum value among the cosine similarities between a target vector and its variants vectors.

Min. The minimum value among the cosine similarities between a target vector and its variants vectors.

Centroid. The cosine similarity between a target vector and the centroid of its variants vectors.

Since some of our targets had many variants that were not found in itWaC, each measure was computed twice: in the first case we simply did not consider the non-occurring variants (*no* models); in the second case, we conceived them as orthogonal vectors to the target vector (*orth* models). For the Mean, Max and Min indices, this meant to automatically set to 0.0 the target-variant cosine similarity. For the Centroid measure, we first computed the cosine similarity between the target vector and the centroid of its attested variants (cs_a). From this initial cosine value we then subtracted the product between the number of non-attested variants (n), cs_a and a costant factor k. This factor k, which was set to 0.01 in previous investigations,

represented the contribution of each zero variant in reducing the target-variants similarity towards 0.0. k was multiplied by the original cosine since we hypothesized that zero variants contributed differently in lowering the target-variants similarity, depending on the construction under consideration:

$$Centroid = cs_a - (cs_a \cdot k \cdot n)$$

5.2 Addition-based and multiplication-based indices

The indices in Section 5.1 were compared against two of the measures described in Krčmář et al. (2013). We trained a DSM on itWaC that represented all the content words with $tokenfrequency > 300$ and our 26 targets as row-vectors and the top 30,000 content words as contexts. The co-occurrence window was still the entire sentence and the weighting was still the PPMI. SVD was carried out to 300 final dimensions. Please note that the context vector of a given word did not include the co-occurrences of a target idiom or target compositional expression that was composed of that word (e.g. the vector for *punto* did not include the contexts of *punto debole*). We then computed the following measures:

Additive. The cosine similarity between a target vector and the vector resulting from the sum of the vectors of its components.

Multiplicative. The cosine similarity between a target vector and the vector resulting from the product of the vectors of its components.

6 Results and discussion

Our 26 targets were sorted in ascending order for each compositionality score. In each ranking, we predicted idioms (our positives) to be placed at the top and compositional phrases (our negatives) to be placed at the bottom, since we expected idiom vectors to be less similar to the vectors of their variants. First and foremost, we must notice that three idioms for every type of variants (Linear DSM, Structured DSM and iMWN) obtained a 0.0 score for all the variant-based indices since no variants were found in itWaC. Nevertheless, we kept this information in our ranking as an immediate proof of the idiomaticity of such expressions. These were *punto debole* 'weak point', *passo falso* 'false step' and *colpo basso* 'cheap shot' for the Structured DSM spaces, *punto debole*, *pecora nera* 'black sheep' and *faccia tosta* 'cheek' for the iMWN spaces and *punto debole*, *passo falso* and *zoccolo duro* 'hard core' for the Linear DSM spaces.

Top IAP Models	**IAP**	**F**	ρ
Additive	0.85	0.77	-0.62***
Structured DSM Mean$_{orth}$	0.84	0.85	-0.68***
iMWN$_{syn}$ Centroid$_{orth}$	0.83	0.85	-0.57**
iMWN$_{ant}$ Centroid$_{orth}$	0.83	0.77	-0.52**
iMWN$_{ant}$ Mean$_{orth}$	0.83	0.69	-0.64***
Top F-measure Models	**IAP**	**F**	ρ
Structured DSM Mean$_{orth}$	0.84	0.85	-0.68***
iMWN$_{syn}$ Centroid$_{orth}$	0.83	0.85	-0.57**
Additive	0.85	0.77	-0.62***
iMWN$_{ant}$ Centroid$_{orth}$	0.83	0.77	-0.52**
iMWN$_{syn}$ Centroid$_{no}$	0.82	0.77	-0.57**
Top ρ Models	**IAP**	**F**	ρ
Structured DSM Mean$_{orth}$	0.84	0.85	-0.68***
Linear DSM Mean$_{orth}$	0.75	0.69	-0.66***
iMWN$_{syn}$ Mean$_{orth}$	0.77	0.77	-0.65***
iMWN$_{syn}$ Mean$_{no}$	0.70	0.69	-0.65***
iMWN$_{ant}$ Mean$_{orth}$	0.83	0.69	-0.64***
Multiplicative	0.58	0.46	0.03
Random	0.50	0.31	0.05

Table 1: Best models ranked by IAP (top), F-measure at the median (middle) and Spearman's ρ correlation with the speakers' judgments (bottom) against the multiplicative model and the random baseline (** = $p < 0.01$, *** = $p < 0.001$).

Table 1 reports the 5 best models for Interpolated Average Precision (IAP), the F-measure at the median and Spearman's ρ correlation with our gold standard idiomaticity judgments respectively. Coherently with Fazly et al. (2009), IAP was computed as the average of the interpolated precisions at recall levels of 20%, 50% and 80%. Interestingly, while Additive was the model that best ranked idioms before non-idioms (IAP), closely followed by our variant-based measures, and figured among those with the best precision-recall trade-off (F-measure), Multiplicative performed comparably to the Random baseline. The best correlation with idiomaticity judgments was instead achieved by one of our variant-based measures (-0.68). Additive did not belong to the 5 models with top correlation, but still achieved a high significant ρ score (-0.62). It's worth noting that all these correlational indices are negative: the more the subjects perceived a target to be id-

iomatic, the less its vector was similar to its variants. Max and Min never appeared among the best performing measures, with all top models using Mean and Centroid. Moreover, the DSM models that worked the best for IAP and F-measure both used dependency-related distributional information, with linear DSM models not reaching the top 5 ranks. This difference was nonetheless ironed out when looking at the Top ρ models. Differently from what we observed for verbal idioms (Senaldi et al., 2016), the majority of our best models, and *de facto* all the Top ρ models, encoded zero variants as orthogonal vectors (*orth* models). Finally, the presence of antonymy-related information for iMWN models did not appear to influence the performances considerably.

7 Conclusions

In this contribution we applied to adjective-noun constructions the variant-based distributional measures we had previously tested on verbal idioms (Senaldi et al., 2016), obtaining effective performances. Interestingly, our measures performed comparably to or even better than the Additive method proposed in the distributional literature (Krčmář et al., 2013), while the Multiplicative one performed considerably worst than all our models, together with the Random baseline.

Future work will concern testing whether these variant-based measures can be succesfully exploited to predict psycholinguistic data about the processing of idiom compositionality and flexibility, together with other corpus-based indices of idiomaticity.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Marco Baroni, Silvia Bernardini, Federica Comastri, Lorenzo Piccioni, Alessandra Volpi, Guy Aston, and Marco Mazzoleni. 2004. Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-Compliant Corpus of Newspaper Italian. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1771–1774.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Cristina Cacciari. 2014. Processing multiword idiomatic strings: Many words in one? *The Mental Lexicon*, 9(2):267–293.

Kenneth W. Church and Patrick Hanks. 1991. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Stefan Evert. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, volume 2, pages 1212–1248. Mouton de Gruyter.

Afsaneh Fazly and Suzanne Stevenson. 2008. A distributional account of the semantics of multiword expressions. *Italian Journal of Linguistics*, 1(20):157–179.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 1(35):61–103.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database.* MIT Press.

Jerry A. Fodor and Ernest Lepore. 2002. *The compositionality papers*. Oxford University Press.

Klaus Krippendorff. 2012. *Content analysis: An introduction to its methodology*. Sage.

Lubomír Krčmář, Karel Ježek, and Pavel Pecina. 2013. Determining Compositionality of Expresssions Using Various Word Space Models and Measures. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 64–73.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324.

Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429.

Geoffrey Nunberg, Ivan Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: Developing and aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302.

Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54.

Marco S. G. Senaldi, Gianluca E. Lebani, and Alessandro Lenci. 2016. Lexical variability and compositionality: Investigating idiomaticity with distributional semantic models. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 21–31.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Sriram Venkatapathy and Aravid Joshi. 2005. Measuring the relative compositionality of verb-noun (V-N) collocations by integrating features. In *Proceedings of Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 899–906.

Markus Werning, Wolfram Hinzen, and Edouard Machery, editors. 2012. *The Oxford Handbook of Compositionality*. Oxford University Press.

Studio sull'ordine dei costituenti nel confronto tra generi e complessità

Giulia Pieri•, Dominique Brunato◇, Felice Dell'Orletta◇
• Università di Pisa, Emm&mmE Informatica
giulia.pieri@mminformatica.it
◇Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{dominique.brunato, felice.dellorletta}@ilc.cnr.it

Abstract

Italiano. In questo articolo presentiamo uno studio sull'ordine dei costituenti in italiano basato su corpora annotati in maniera automatica fino all'analisi sintattica a dipendenze. L'indagine comparativa ha permesso di valutare l'influenza sia del genere testuale sia della complessità linguistica nella distribuzione dei fenomeni di marcatezza sintattica.

English. *In this paper we present a study on the order of constituents in Italian based on automatically dependency–parsed corpora. The comparative investigation has allowed to evaluate the influence of the textual genre and the linguistic complexity on the distribution of phenonemena of syntactic markedeness.*

1 Introduzione

Sebbene non esista una metrica universalmente valida con la quale poter classificare le lingue secondo una scala di complessità (McWorther, 2001), esistono alcuni indicatori che, a diversi livelli linguistici, possono essere assunti come indici di complessità 'universalmente' validi (Fiorentino, 2009). Sul piano sintattico, uno di essi è rappresentato dall'ordine dei costituenti, per cui le lingue che ammettono un ordine libero sono considerate più complesse di quelle a ordine fisso. Nella letteratura linguistica e psicolinguistica la flessibilità dell'ordine viene ricondotta, a sua volta, a fattori diversi che tengono in considerazione, da un lato, i principi semantici e pragmatici determinati dalla struttura dell'informazione (Diessel, 2005), dall'altro i vincoli di *performances*, per cui le strutture non marcate sono quelle cognitivamente meno costose che permettono al parlante di elaborare l'informazione più velocemente (Hawkins, 1994; Gibson, 1998; Gibson, 2000). Esaminando in maniera comparativa due treebank del Latino e del Greco antico, lo studio di (Gulordava e Merlo, 2015) ha dimostrato come la flessibilità dell'ordine sintattico, misurata come la distanza tra l'effettiva lunghezza delle dipendenze di una frase e la sua lunghezza ottimale (Gildea e Temperley, 2010), sia un elemento di complessità che si può desumere tanto dalla minor precisione del parsing automatico nell'analisi di queste lingue, quanto dalla tendenza che si riscontra nel tempo verso modelli di ordine fisso dei costituenti.

Questo articolo propone uno studio quantitativo per l'italiano, lingua di tipo VO (o a testa iniziale) e relativamente poco flessibile, volto a indagare se, e in che misura, la disposizione naturale o *non marcata* dei costituenti nella frase sia influenzata dal genere testuale e dalla complessità della lingua usata nel testo. A questo scopo sono stati comparati due generi linguistici, narrativo e giornalistico, a loro volta distinti in due varietà linguistiche differenti per grado di complessità, dove tale grado è definito in relazione al lettore di riferimento. A differenza delle analisi tradizionali di tipo *corpus–based* sull'ordine dei costituenti in italiano, tutti i dati qui discussi sono ricavati da corpora annotati in maniera automatica fino al livello di analisi sintattica a dipendenze. Anche se la ricostruzione della struttura sintattica da parte di un parser statistico è soggetta inevitabilmente ad alcuni errori (Montemagni, 2013), che aumentano per i testi di un dominio distante da quello del *training* (Gildea, 2001), la varietà dei fenomeni che si possono monitorare con affidabilità a partire da un'analisi linguistica automatica è molto ampia e complessa. La prospettiva linguistico–computazionale apre dunque prospettive di ricerca promettenti per la costruzione e la validazione su larga scala di modelli teorici sul funzionamento dei sistemi linguistici sia in chiave tipologica sia rispetto ai tradizionali assi di variazione linguistica.

In quanto segue, verranno prima presentati i corpora utilizzati in questo studio e successivamente la metodologia di monitoraggio sui cui si è basata l'estrazione delle caratteristiche linguistiche oggetto di indagine (Paragrafo 2.1); nel Paragrafo 3 discuteremo i principali risultati ottenuti e infine trarremmo alcune conclusioni di questa ricerca.

2 I corpora

I corpora esaminati appartengono a due diversi generi testuali, narrativo e giornalistico. Per ciascun genere sono state selezionate due collezioni di testi rappresentative di due varietà di lingua che si possono collocare a due poli opposti per complessità linguistica, dove il grado di complessità è definito in base al destinatario previsto; ogni macroraccolta, dunque, contiene una collezione di testi "complessi" e una di testi "semplici".

I due corpora narrativi, *Terence* e *Teacher*, rappresentano la prima risorsa italiana per lo studio della semplificazione automatica e semi-automatica dei testi (Brunato et al., 2015). Entrambi sono costituiti da testi nella versione originale e nella rispettiva versione semplificata, allineate per ciascun corpus a livello di frase. Le versioni semplificate derivano da due differenti strategie di semplificazione manuale: la strategia "strutturale", che implica una semplificazione cumulativa (ovvero su diversi livelli linguistici) prodotta da esperti nel caso di *Terence*, e la strategia "intuitiva", che si avvale invece dell'intuizione e dell'esperienza dell'insegnante nel caso di *Teacher*. In particolare, *Terence* si compone di 32 racconti brevi per l'infanzia e delle rispettive versioni semplificate rivolte a bambini dai 7 agli 11 anni con deficit uditivi o con difficoltà nella comprensione dei testi [1]. *Teacher* è un corpus formato da 24 coppie di testi originali e semplificati raccolti da siti web educativi specializzati che forniscono risorse gratuite per gli insegnanti; in questo caso, il target della semplificazione sono principalmente studenti di lingua italiana L2.

Per il genere giornalistico, invece, il materiale analizzato è costituito da due corpora che raccolgono rispettivamente testi esemplificativi di una varietà complessa, *Repubblica*, e di una varietà semplice, *Due Parole*. Il primo (Rep) consiste in un ampio corpus di testi giornalistici (pari a 232.908 tokens) che include tutti gli articoli pubblicati dal 2000 al 2005 sul quotidiano *La Repubblica*, che si rivolge ad una platea di lettori con un profilo culturale medio-alto. Il secondo (2Par) è un corpus di 73.314 tokens che trae il nome dall'omonimo quotidiano *Due Parole*, un mensile di facile lettura curato da linguisti esperti in semplificazione dei testi che hanno utilizzato un linguaggio controllato per un pubblico adulto con un basso livello di alfabetizzazione o con lievi disabilità intellettuali (Piemontese, 1996). Il corpus qui analizzato comprende tutti gli articoli scritti tra il 2001 e il 2006. È importante sottolineare che, a differenza dei corpora di narrativa, il corpus giornalistico non è parallelo, in quanto i relativi testi "semplici" (quelli di *Due Parole*) non sono il risultato di un processo di semplificazione dei testi originali di *Repubblica*.

2.1 Analisi linguistica dei corpora

Come passo preliminare allo studio dei fenomeni di ordinamento sintattico riportati in Sezione 3, i corpora sono stati arricchiti automaticamente con annotazione morfo–sintattica e sintattica utilizzando la catena di analisi linguistica *LinguA*[2], che integra il Part–of–Speech tagger descritto in (Dell'Orletta, 2009) e il parser a dipendenze DeSR (Attardi et al., 2009). L'annotazione linguistica multi–livello ha permesso di analizzare gli stessi tramite MONITOR–IT: questo strumento, adottando la metodologia di monitoraggio descritta in Montemagni (2013), consente di ricavare la distribuzione di un'ampia gamma di caratteristiche lessicali, morfo–sintattiche e sintattiche rintracciate automaticamente in un corpus a partire dall'output dei diversi livelli di annotazione linguistica.

3 Analisi dei dati

Per gli scopi di questa indagine sono di interesse caratteristiche di ordine sintattico che fanno riferimento alla posizione lineare di un elemento rispetto alla "testa" da cui è retto in una rappresentazione sintattica a dipendenze. Gli elementi considerati sono stati: il soggetto, l'oggetto, l'avverbio, l'aggettivo e la clausola subordinata, di cui sono state calcolate: i) le occorrenze nella posizione "canonica" rispetto alla matrice prevalente SVO dell'italiano (preposta o posposta alla testa a seconda dell'elemento indagato) e nella posizione opposta, dunque "marcata" sintatticamente e/o

[1] Questo corpus deriva dall'omonimo progetto dell'Unione Europea (Terence Consortium, 2012).

[2] http://linguistic-annotation-tool.italianlp.it/

Corpus	Oggetto				Soggetto				Aggettivo				Avverbio			
	Pre-V		Post-V		Pre-V		Post-V		Pre-N		Post-N		Pre-V		Post-V	
	%	AvD	%	AvD	%	AvD	%	AvD	%	AvD	%	AvD	%	AvD	%	AvD
TT orig	9.18	1.93	90.82	2.52	85.38	2.56	14.62	2.88	53.91	1.11	46.09	1.2	55.49	2.4	44.51	1.61
Rep	8.37	2.43	91.63	2.72	80.14	3.87	19.86	3.45	41.87	1.19	58.13	1.32	56.11	2.66	43.89	1.47
TT sempl	7.87	1.93	92.13	2.43	84.28	2.23	15.72	2.63	56.53	1.12	43.47	1.16	56.24	2.19	43.76	1.47
2Par	3.47	1.6	96.53	2.56	89.11	3.07	10.89	3.5	24.97	1.09	75.03	1.12	56.69	3.84	43.31	1.4

Tabella 1: Ordine relativo dei costituenti (%) e distanza media (AvD) rispetto alla testa verbale (V) o nominale (N).

pragmaticamente; ii) la distanza (in numero di tokens) del dipendente dalla testa sintattica in entrambe le posizioni. Per ognuno di questi dati, il confronto tra i corpora è avvenuto su due livelli: la variazione di genere e il grado di complessità. Infatti, scopo dello studio è stato verificare quali sono gli ordini degli elementi che vengono condizionati dal genere testuale e quali dipendono dal grado di complessità: l'ipotesi di partenza era che fosse possibile ritrovare una somiglianza dell'ordine degli elementi in relazione al genere, ma soprattutto verificare che, indipendentemente dal genere, i testi semplici sono più fedeli a seguire l'ordine canonico degli elementi, mentre i testi complessi presentano una più alta percentuale di casi di ordine marcato.

La Tabella 1 mostra i risultati del monitoraggio relativi all'oggetto, al soggetto, all'aggettivo e all'avverbio[3]. Partiamo dall'analisi degli elementi che, nel confronto complessivo tra corpora, dimostrano una tendenza più netta a ricorrere nella posizione canonica: l'oggetto e il soggetto. Nel caso dell'oggetto, si osserva che i testi giornalistici si attengono maggiormente all'ordine canonico, mentre nei testi narrativi aumentano lievemente le occorrenze dell'oggetto in posizione preverbale. L'ordine marcato con anteposizione dell'oggetto alla testa verbale è inoltre influenzato dal grado di complessità della lingua: in ciascun genere infatti, quest'ordine ricorre in percentuale minore nei testi semplici e tale differenza è evidente soprattutto in *2Par* che registra poco più del 3% di oggetti in posizione preverbale. Anche rispetto alla posizione del soggetto, è possibile notare un'influenza sia del genere sia della complessità. In questo caso, però, sono i testi narrativi originali a rispettare maggiormente l'ordine canonico soggetto–verbo (85,38%) rispetto a quelli di *Repubblica* (80,14%). La variazione rispetto al grado di complessità produce invece risultati coerenti alle aspettative solo per la prosa giornalistica, dove lo scarto tra *Rep* e *2Par* è quasi di 10 punti percentuali in favore dell'ordine canonico (*2Par*: 89,11%). Al contrario, la semplificazione dei testi narrativi ha prodotto un aumento, seppure minimo, di soggetti postverbali (TT orig: 14,62%; TT semp: 15,72%). Pur considerando che i testi narrativi originali sono comunque più semplici di quelli di *Repubblica*, proprio perché rivolti a bambini, questo dato potrebbe segnalare che forme di marcatezza sintattica sono talvolta preferite come esito della semplificazione, perché permettono di ottenere un testo narrativo più coeso, mantenendone la progressione tematica. Interessanti sono anche i dati sulla distanza lineare tra soggetto e verbo che, in entrambi i generi della varietà semplice, aumenta quando il soggetto è in posizione postverbale. Si può ipotizzare, tuttavia, che la presenza dei tratti di accordo sul verbo in una lingua come l'italiano renda meno difficoltosa la ricostruzione della dipendenza soggetto–verbo, anche quando il soggetto è in posizione marcata.

A differenza del soggetto e dell'oggetto, l'aggettivo in italiano ha una posizione meno rigida nel sintagma nominale. Infatti, anche se la posizione non marcata è generalmente postnominale, essa varia in base alla funzione semantica che l'aggettivo svolge rispetto al nome (Cinque, 2010). Questa flessibilità trova conferme nell'analisi empirica, tuttavia con differenze rispetto al genere: i testi giornalistici, infatti, privilegiano l'ordine tendenzialmente non marcato mentre quelli narrativi mostrano la tendenza opposta. Anche in questo caso, sul piano della variazione testi complessi/testi semplici, l'effetto è marcato solo per il genere giornalistico (*Rep*: 58,13; *2Par*: 75,03).

Considerazioni analoghe possono essere avanzate per l'avverbio, la cui posizione in italiano, pur essendo tendenzialmente postverbale, gode di ampia flessibilità in relazione alla classe semantica di ap-

[3]Per rendere possibile il confronto tra gradi di complessità, i corpora Terence e Teacher sono stati uniti così da ottenere due corpora, l'uno composto di tutti i testi narrativi originali (TT orig), pari a 26.311 tokens, e l'altro di tutti i relativi testi semplificati (TT semp), pari a 24.083 tokens.

partenenza (Bonvino et al., 2008). In tutti e quattro i corpora è preferita la posizione preposta al verbo, che è anche quella a generare link sintattici mediamente più lunghi (si veda il dato riportato nella terzultima colonna). Si tratta di un dato significativo, soprattutto se si considera che il valore medio più elevato è riportato proprio dai testi di *2Par* (3.84 tokens). Come per il caso del soggetto, anche questo dato suggerisce la necessità di raffinare una nota misura di complessità sintattica quale la distanza dei link sintattici, tenendo in considerazione proprietà semantiche e morfologiche degli elementi coinvolti nella relazione di dipendenza.

Infine, abbiamo condotto uno studio più dettagliato sulla subordinazione (Tabella 2). Anche in questo caso sono state estratte sia le distribuzioni percentuali della subordinata in posizione preposta e posposta alla reggente sia la distanza (in numero di tokens) che separa la part–of–speech che introduce la subordinata [4] dal verbo della reggente. Inoltre, questo dato è stato ulteriormente raffinato andando a calcolare la lunghezza totale (in tokens) dell'intera clausola subordinata e la sua profondità media, quest'ultima computata come numero di relazioni di dipendenza che intercorrono tra la radice del sotto–albero della subordinata e una parola senza dipendenti (foglia).

Corpus	**Subordinata**			
	Pre-Principale			
	%	AvD	Length	Depth
TT orig	10.12	9.71	8.93	3.86
Rep	15.37	11.51	9.49	4.16
TT sempl	11.03	8.0	7.19	3.63
2Par	15.71	10.26	7.43	3.72
	Post-Principale			
TT orig	89.88	3.27	8.62	4.19
Rep	84.63	3.44	12.07	5.28
TT sempl	88.97	2.94	7.91	4.12
2Par	84.29	3.0	8.39	4.36

Tabella 2: Ordine della clausola subordinata rispetto alla principale. Per ciascuna posizione, vengono riportate la distribuzione percentuale (%), la distanza media dalla principale (AvD), la lunghezza media (Length) e la profondità media (Depth) dell'intera subordinata.

I risultati indicano una netta preferenza per la posizione posposta rispetto alla principale. Il dato è coerente con le previsioni dei modelli di *processing* secondo cui questo ordinamento comporta un impegno cognitivo minore da parte del parlante e dell'ascoltatore perché consente di minimizzare i domini di riconoscimento delle relazioni sintattiche (Hawkins, 1994). Anche se meno frequenti, i casi di anteposizione della subordinata si verificano maggiormente nel genere giornalistico, addirittura nella varietà semplice (*2Par*: 15.71% *Rep*: 15.37%). Questi dati sono riconducibili alle teorie che chiamano in causa l'interazione tra sintassi e fattori pragmatici e semantici, per cui il genere giornalistico sarebbe più propenso ad anteporre la subordinata alla principale poiché costituisce lo sfondo tematico dell'evento principale e conferisce la funzione di collegamento tematico e introduzione per l'informazione nuova (Diessel, 2005). Come prevedibile, l'anteposizione della subordinata determina dipendenze sintattiche mediamente più lunghe; la difficoltà di processing che ne deriva è compensata dall'uso di subordinate più semplici, non solo in termini di lunghezza totale ma soprattutto strutturalmente: in tutti i corpora, infatti, le "catene" subordinanti hanno una profondità media minore quando la subordinata precede la principale.

4 Conclusione

Questo articolo ha proposto uno studio comparativo su un particolare fenomeno relativo alla complessità sintattica, ovvero l'ordine dei costituenti in italiano. Il confronto è stato condotto su due livelli: la variazione di genere e il grado di complessità.

Per quanto riguarda il primo, è stato possibile constatare che i testi giornalistici sono quelli che maggiormente si attengono all'ordine canonico degli elementi, mentre i testi narrativi hanno riportato una frequenza superiore di ordini marcati. Dal punto di vista della complessità, è chiara la tendenza in entrambi i generi a utilizzare l'ordine canonico come esito della semplificazione, sia a seguito di un processo di semplificazione di un testo originale, sia quando il testo nativamente è concepito come testo semplice.

Indipendentemente dal genere, il fenomeno che è risultato più legato alla complessità riguarda l'uso delle subordinate. In entrambi i generi prevalgono nettamente subordinate posposte alla principale in quanto più facili da processare e quando questa posizione non è rispettata si registra una tendenza alla semplificazione della subordinata stessa sia in termine di numero di parole, ma soprattutto strutturalmente, in termini di profondità del sottoalbero sintattico.

[4] Sono state considerate sia le subordinate esplicite, introdotte da una congiunzione subordinante, sia quelle implicite, introdotte da un verbo di modo infinito o da una preposizione.

Va infine ricordato che tutte le osservazioni riportate in questo studio sono basate su testi linguisticamente annotati in maniera automatica, dunque soggetti a errore. Nonostante ciò, ci aspettiamo che almeno limitatamente all'analisi di testi dello stesso dominio e varietà di lingua, le distribuzioni degli errori siano simili, permettendo dunque un confronto interno rispetto ai parametri linguistici indagati. L'affidabilità dei dati discussi è inoltre corroborata dal fatto che sono stati considerati testi standard, linguisticamente vicini a quelli sui quali gli strumenti di annotazione automatica sono tipicamente addestrati. D'altra parte, proprio perché la distribuzione degli errori potrebbe variare al variare del dominio dei testi, tra gli sviluppi di questo lavoro intendiamo condurre delle analisi a campione per verificare l'impatto dell'errore sui confronti ottenuti rispetto alle diverse strutture esaminate.

References

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, Joseph Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*, Reggio Emilia, Italia, Dicembre 2009.

Elisabetta Bonvino, Mara Frascarelli, Paola Pietrandrea. 2008. Semantica, sintassi e prosodia di alcune espressioni avverbiali nel parlato spontaneo. *La comunicazione parlata*, Massimo Pettorino, Antonella Giannini, Marianna Vallone, Renata Savy (Eds), Napoli, Liguori, 565–607.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, Simonetta Montemagni. 2015. Design and annotation of the first italian corpus for text simplification. In *Proceedings of LAW IX - The 9th Linguistic Annotation Workshop.* Denver, Colorado, Giugno 2015.

Guglielmo Cinque. 2010. *The syntax of adjectives: A comparative study*. In *MIT Press*.

Felice Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. In *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*, Reggio Emilia, Italia, Dicembre 2009.

Holger Diessel. 2005. Competing motivations for the ordering of main and adverbial clauses. *Linguistics*, 43 (3): 449–470.

Giuliana Fiorentino. 2009. Complessità linguistica e variazione sintattica. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, (2), 281-312.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68:1–76.

Edward Gibson. 2000. The dependency Locality Theory: A distance–based theory of linguistic complexity. *Image, Language and Brain*, In W.O.A. Marants and Y. Miyashita (Eds.), Cambridge, MA: MIT Press, 95–126.

Daniel Gildea. 2001. Corpus variation and parser performance. *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2001)*, Pittsburgh, PA.

Daniel Gildea, David Temperley. 2010. Do Grammars Minimize Dependency Length? *Cognitive Science*, 34(2):286–310.

Kristina Gulordava, Paola Merlo. 2015. Diachronic Trends in Word Order Freedom and Dependency Length in Dependency-Annotated Corpora of Latin and Ancient Greek. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, Uppsala, Sweden, August 24–26 2015, pp. 121–130.

John A. Hawkins 1994. A performance theory of order and constituency. Cambridge studies in Linguistics. *Cambridge studies in Linguistics*, Cambridge University Press., Numero 73.

John. H. McWorther. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology*, 5, 125-166.

Simonetta Montemagni. 2013. Tecnologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica e Applicata (SILTA)*, (1), 145-172.

Maria Emanuela Piemontese. 1996. *Capire e farsi capire. Teorie e tecniche della scrittura controllata*. Napoli, Tecnodid.

Terence Consortium. 2012. Story simplification: User guide. Restricted Distribution.

Grounding the Lexical Sets of Causative-Inchoative Verbs with Word Embedding

Edoardo Maria Ponti
University of Cambridge
ep490@cam.ac.uk

Elisabetta Jezek
Università degli Studi di Pavia
jezek@unipv.it

Bernardo Magnini
Fondazione Bruno Kessler
magnini@fbk.eu

Abstract

English. Lexical sets contain the words filling the argument positions of a verb in one of its senses. They can be extracted from corpora automatically. The purpose of this paper is demonstrating that their vector representation based on word embedding provides insights onto many linguistic phenomena, such as causative-inchoative verbs. A first experiment aims at investigating the internal structure of the sets, which are known to be radial and continuous categories cognitively. A second experiment shows that the distance between the intransitive subject set and transitive object set is correlated with the spontaneity of the event expressed by the verb, defined according to morphological coding and frequency.

Italiano. *I set lessicali contengono le parole che occupano le posizioni argomentali di un verbo in una delle sue accezioni, e possono essere estratti in modo automatico dai corpora. L'obiettivo di questo articolo è dimostrare che la loro rappresentazione vettoriale illumina alcuni fenomeni linguistici, come i verbi ad alternanza causativo-incoativa. Un esperimento investiga la struttura interna degli insiemi, che a livello cognitivo sono ritenuti categorie radiali e continue. Inoltre, un secondo esperimento mostra che la distanza fra l'insieme dei soggetti intransitivi e l'insieme degli oggetti transitivi è correlata alla spontaneità dell'evento espresso dal verbo, definita secondo la marca morfologica e la frequenza.*

1 Introduction

Lexicographic attempts to cope with verb sense disambiguation often rely on "lexical sets" (Hanks, 1996), which represent the lists of corpus-derived words that appear as arguments for each distinct verb sense. The arguments are the "slots" that have to be filled to satisfy the valency of a verb (subject, object, etc.). For example, {*gun*, *bullet*, *shot*, *projectile*, *rifle*...} is the lexical set of the object for the sense 'to shoot' of *to fire*. In previous works, e.g. Montemagni et al. (1995), lexical sets were collected manually and were compared through set analysis. The measure of similarity between two sets was proportional to the extent of their intersection. We believe that possible improvements may stem from deriving the lexical sets automatically and from exploiting the semantic information of the fillers fully. In this work, we devise an extraction method from a huge corpus and use a distributional semantics approach to perform our analyses. More specifically, we represent fillers as word vectors and compare them with spatial distance measures. In order to test the relevance for linguistic theory of this approach, we focus on a case study, namely the properties of verbs undergoing the causative-inchoative alternation. Section 1.1. outlines a framework for word embeddings and section 1.2 introduces the case study. Section 2 presents the method and the data, whereas section 3 reports the results of a couple of experiments.

1.1 Word Embedding

The full exploitation of the semantic information inherent to argument fillers for verbs can take advantage from some recent developments in distributional semantics. Recently, efficient algorithms have been devised mapping each word of a vocab-

ulary into a corresponding vector of n real numbers, which can be thought as a sequence of coordinates in a n-dimensional space (Mikolov et al., 2013). This mapping is yielded by unsupervised machine learning, based on the assumption that the meaning of a word can be inferred by its context, i.e. its neighbouring words in texts. This model has some relevant properties: the geometric closeness of two vectors corresponds to the similarity in meaning of the corresponding words. Moreover, its dimensions have possibly a semantic interpretation.

1.2 Causative-Inchoative Alternation

A possible testbed for the usefulness of representing the argument fillers as vectors are the verbs showing the so called causative-inchoative alternation. These verbs appear either as transitive or intransitive. In the first case, an agent brings about a change of state; in the second, the change of a patient is presented as spontaneous (e.g. *to break*, as in "Mary broke the key" vs. "the key broke").

The two alternative forms of these verbs can be morphologically asymmetrical: in this case, one has a derivative affix and the other does not. The first is labelled here as "marked", the second as "basic". Italian verbs with an asymmetrical alternation derive from the phenomenon of anti-causativization. The intransitive form is marked since it is sometimes preceded by the clitic *si* (Cennamo and Jezek, 2011). Haspelmath (1993) maintain that verbs that show a preference for a marked causative form (and a basic inchoative form) cross-linguistically denote a more "spontaneous" situation. Spontaneity is intended by the author as the likelihood of the occurrence of the event without the intervention of an agent. This work is non-committal with respect to whether spontaneity be an actual semantic factor. Rather, it is considered a notion useful for labelling the observed variations in morphology and frequency.

In this way, a correlation between the form and the meaning of these verbs was demonstrated. Moreover, Samardzic and Merlo (2012) and Haspelmath et al. (2014) argue that verbs that appear more frequently (intra- and cross-linguistically) in the inchoative form tend to morphologically derive the causative form, too. This time, the correlation holds between form and frequency. Vice versa, situations entailing agentive participation prefer to mark the inchoative form and occur more frequently in the causative form.

2 Previous Work

In the literature, many methods are available for the automatic detection of verb classes, such as causative-inchoative verbs. They exploit features based on argument alternations, such as subcategorization frames (Joanis et al., 2008). The identification of verb classes displaying a diathesis alternation was also performed through the analysis of selectional preferences. Most notably, the lexical items were compared via distributional semantics (McCarthy, 2000).

These features were usually induced from automatic parses of heterogeneous and wide corpora (Schulte Im Walde, 2000). In particular, the extraction of subcategorization frames was refined including e.g. noise filters based on frequency (Korhonen et al., 2000). Our work is inspired by these attempts to automatically induce lexical information regarding verbs, but its direction of research is reversed. Indeed, rather than classifying verb classes given this information, it analyses this information given a verb class in order to shed light on its properties from the perspective of linguistic theory.

3 Data and Method

The data are sourced from a sample of ItWac, a wide Italian corpus gathered through web crawling (Baroni et al., 2009). This sample was further enriched with morpho-syntactic information through the MATE-tools parser (Bohnet, 2010)[1] and filtered by sentence length (< 100). Eventually, sentences in the sample amounted to 2,029,454 items. A target group of 20 causative-inchoative verbs was taken from Haspelmath et al. (2014): they are listed here based on the reported transitive/intransitive frequency ratio, from the highest to the lowest.

close > open > improve > break > fill > gather > connect > split > stop > go out > rise > rock > burn > freeze > turn > dry > wake > melt > boil > sink

The extraction step consisted in identifying their argument fillers inside the sentences in the sample. In particular, the arguments considered were the subjects of intransitives (S) and objects

[1] LAS scores for the relevant dependency relations: 0.751 with dobj (direct object), 0.719 with nsubj (subject), 0.691 with nsubjpass (subject of a passive verb).

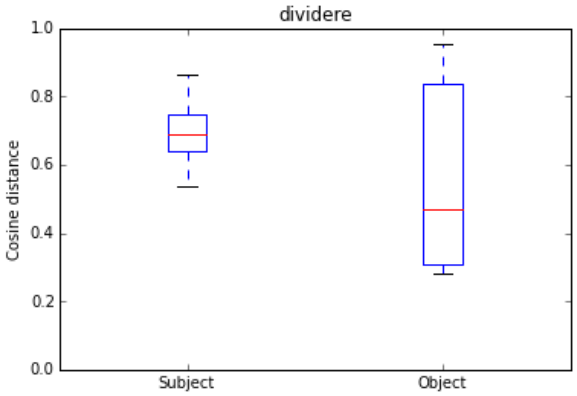


Figure 1: Distance of vectors from their centroid.

(O) (Dixon, 1994).[2] These arguments are relevant because they are deemed to share the same fillers (Pustejovsky, 1995).

These operations resulted in a database where each verb lemma had a single entry and was associated with a list of fillers, divided by argument type. With this procedure, lexical sets were extracted automatically, although they were not divided by verb sense. Afterwards, each of the argument fillers was mapped to a vector relying on a space model pre-trained through Word2Vec (Dinu et al., 2015).[3]

4 Experiments

In order to bring to light the linguistic information concealed in the automatically extracted lexical sets, we devised two experiments. One investigates the internal structure of lexical sets. In fact, previous works based on set theory treated them as categoric sets, of which a filler is either a member or not. Research in psychology, however, has long since demonstrated that the members of a linguistic set are found in a radial continuum where the most central one is the prototype for its category, and those at the periphery are less representative (Rosch, 1973; Lakoff, 1987).[4] Word vectors allow to capture this spatial continuum.

Once the fillers have been mapped to their respective vectors, a lexical set appears as a group of points in a multi-dimensional model. The centre of this group is the Euclidean mean among the vectors, which is a vector itself and is called centroid. In the first experiment, we calculated the coordinates of the centroid of the lexical sets S and O for any selected verb[5]. Then we evaluated the cosine similarity of every vector member of the sets from its centroid. The value of this metric goes from 0 (overlap) to 1 (maximum distance) and is useful to evaluate how far a filler is from its prototype. We obtained two sets of cosine similarity values for each verb: these can be plotted as boxes and whiskers, like in Figure 1. The example represents those of *dividere* 'to split'. The rectangles stand for the values in the second and third quartiles, whereas the horizontal line for the median[6]. From all these distance values, we picked the median value for each lexical set. The plot of these medians for the S set and the O set of each verb ordered according to Haspelmath's ranking is shown in Figure 2.

Two main results can be observed from these plots: the S lexical set lies in a more compact range of distances, whereas O is more scattered. On the other hand, the vectors of S tend farther from the centroid. This is demonstrated by the ranges where their distance values fall. Moreover, the averages of medians for the ten verbs on the left part of the scale (frequently transitive) and for the ten verbs on the right (frequently intransitive) were compared. The average median in S was 0.696567 for the former and 0.585263 for the latter. The average median in O was 0.556878 for the former and 0.522418 for the latter. This shows that the variation in O appears to be random. On the other hand, the median of the distances in S is normally lower for verbs that lie in the bottom half of the Haspelmath's scale.

The second experiment consisted in estimating the cosine distance between the centroid of S and the centroid of O for each verb. This operation was aimed at finding to which extent the lexical sets of S and O overlap. In fact, Montemagni et al. (1995) and McCarthy (2000) assessed in a corpus some asymmetries between these lexical sets, which in principle should share all their members.

[2] Subjects of forms with *si* were treated as intransitive subjects. Subjects of passive verbs were treated as objects.

[3] It was generated by a CBOW algorithm with negative sampling, 300 dimensions, a context window of 10 tokens, pruning of infrequent words and sub-sampling.

[4] For previous work on lexical sets considering prototypicality in the context of the notion of shimmering, see Jezek and Hanks (2010).

[5] Every filler was weighted proportionally to its absolute frequency.

[6] The median is the value separating the higher half of the ordered values from the lower half.

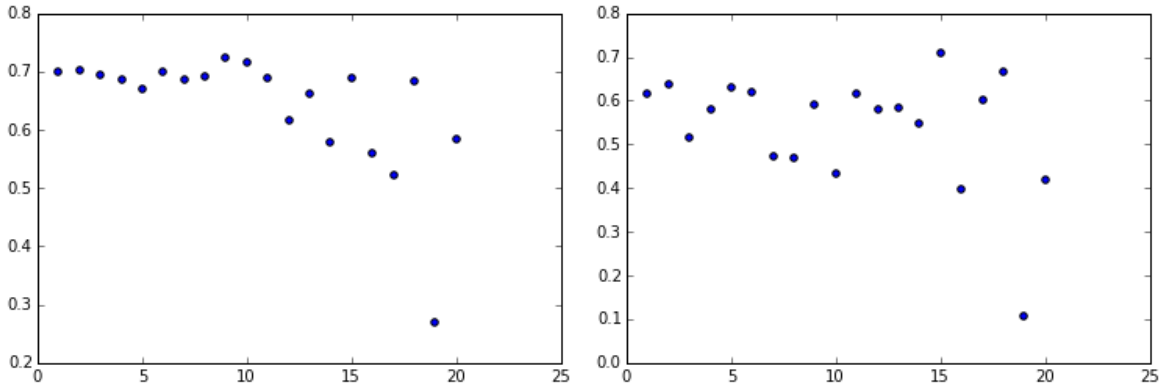


Figure 2: Medians of S (left) and O (right) distances for verbs ranked by position in Haspelmath's scale.

Inspecting our results, the distance between S and O seems to behave as a measure of spontaneity, intended as cross-linguistic frequency and morphological markedness of a verb: the more the centroids tend to be set apart, the more the verb tends to have a morphologically unmarked and more frequent intransitive form. In fact, we compared the ranking of 20 alternating verbs according to the ratio of their cross-linguistic frequency of transitive and intransitive forms (Haspelmath et al., 2014) and a ranking based on the centroid distances of the same verbs. Both these rankings are plotted in Figure 3: every verb is associated with its position in the two scales.

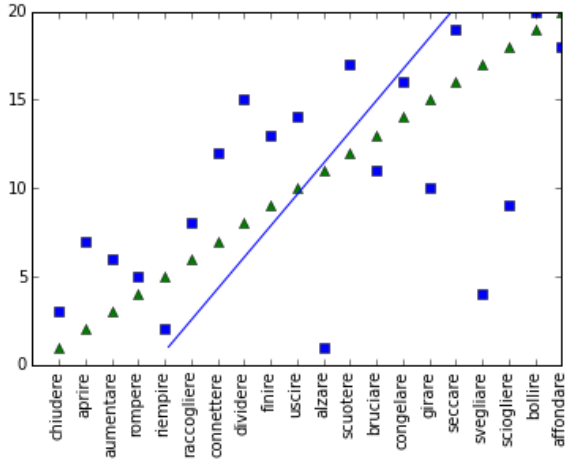


Figure 3: Ranking based on cross-linguistic form frequencies (green triangles) against ranking based on distance of the centroids of S and O in Italian (blue squares).

Both scales display a common tendency. In particular a Spearman's ranking test was performed over them, yielding a mild positive correlation of $\rho = 0.56391$ with a quite strong confidence, i.e. with $p < 0.01$.[7]

5 Discussion

The representation of lexical sets of Italian causative-inchoative verbs as vectors was demonstrated to provide insights into their internal structure and their relation with spontaneity defined according to morphological coding and frequency. The distances of the objects appeared to be distributed more uniformly, whereas those of the intransitive subjects more densely and remotely from the centroid. This difference cannot stem from the frequency of anaphoric fillers (contrary to transitive subjects), since both these argument positions share the discursive function of introducing new referents, and are hence occupied by fully referential fillers (Du Bois, 1985).

Moreover, the medians of the distances of the subject fillers from their centroid were shown to vary. An interpretation is that they are sensible to the frequency scale: this implies that frequently transitive (hence, non-spontaneous) verbs have semantically less homogeneous sets of referents, since they are farther from the prototype. Possibly this discovery can be related with the fact that non-spontaneous verbs impose less selectional restrictions on subjects (McKoon and Macfarland, 2000).

The lack of a perfect correlation between these vector distance and frequency measures is maybe due to errors in the automatic extraction and data sparseness for the former, or an insufficient sample

[7] An alternative measure was considered for the ranking: the cardinality of the S-O intersection weighted by the set union. In this case, Spearman correlation was $\rho = 0.42255$, but it was not significant because $p \approx 0.06$.

of languages in the typological survey of Haspelmath et al. (2014) for the latter. A possible interpretation of the correlation is that the entities capable of bringing about a change of state and those that undergo it are indiscernible only for non-spontaneous verbs. Studies on causer entities related them not only with the feature of agentivity, but also in general with the so-called 'teleological capability' (Higginbotham, 1997).

6 Conclusion

Our work provided evidence that lexical sets of Italian causative-inchoative verbs are continuous and radial categories, whose distribution around the prototype vary to a great extent. It is sensitive to the grammatical role and sometimes to the position of the verb in the so-called spontaneity scale. Moreover, a correlation was discovered between the distance between transitive object and intransitive subject lexical sets of a given verb and its cross-linguistic tendency to appear more frequently as intransitive or as transitive. Figure 4 is a synopsis of this result in the context of the correlations established in previous works.

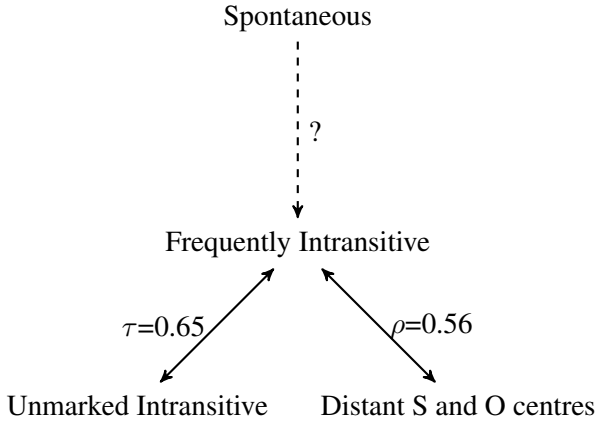


Figure 4: Synopsis of correlations among features of causative-inchoative verbs. The measures are based on Kendall Tau test (τ) and Spearman's ranking test (ρ).

In Figure 4, solid lines stand for correlations proven based on cross-linguistic evidence (frequency-form) and evidence from the Italian language (frequency-lexical sets). The dotted line, on the other hand, suggests the existence of and underlying motivation for the correlations, which nonetheless remains unproven and undetermined in its nature. Its possible validation is left to future research.

Future works should also choose different pretrained vector models, in order to try and replicate these results. In particular, the new vector models could be optimized for similarity through semantic lexica (Faruqui et al., 2015) or based on syntactic dependencies (Séaghdha, 2010). The experiments in this work may be extended to other languages, either individually or through a multi-lingual word embedding (Faruqui and Dyer, 2014).

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 89–97. Association for Computational Linguistics.

Michela Cennamo and Elisabetta Jezek. 2011. The anticausative alternation in italian. *I luoghi della traduzione*, pages 809–823.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. *workshop contribution at ICLR 2015*.

Robert MW Dixon. 1994. *Ergativity*. Cambridge University Press.

John W Du Bois. 1985. Competing motivations. *Iconicity in syntax*, pages 343–365.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation.

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL*.

Patrick Hanks. 1996. Contextual dependency and lexical sets. *International Journal of Corpus Linguistics*, 1(1):75–98.

Martin Haspelmath, Andreea Calude, Michael Spagnol, Heiko Narrog, and Elif Bamyaci. 2014. Coding causal–noncausal verb alternations: A form–frequency correspondence explanation. *Journal of Linguistics*, 50(03):587–625.

Martin Haspelmath. 1993. More on the typology of inchoative/causative verb alternations. *Causatives and transitivity*, 23:87.

James Higginbotham. 1997. Location and causation. *Ms., University of Oxford*.

Elisabetta Jezek and Patrick Hanks. 2010. What lexical sets tell us about conceptual categories. *Lexis*, 4(7):22.

Eric Joanis, Suzanne Stevenson, and David James. 2008. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(03):337–367.

Anna Korhonen, Genevieve Gorrell, and Diana McCarthy. 2000. Statistical filtering and subcategorization frame acquisition. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, pages 199–206. Association for Computational Linguistics.

George Lakoff. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. Cambridge University Press.

Diana McCarthy. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 256–263.

Gail McKoon and Talke Macfarland. 2000. Externally and internally caused change of state verbs. *Language*, pages 833–858.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop at ICLR*.

Simonetta Montemagni, Nilda Ruimy, and Vito Pirrelli. 1995. Ringing things which nobody can ring. a corpus-based study of the causative-inchoative alternation in italian. *Textus online only. 8 (1995), N. 2, 1995*, 8(2):1000–1020.

James Pustejovsky. 1995. *The generative lexicon*. The MIT Press.

Eleanor H Rosch. 1973. Natural categories. *Cognitive psychology*, 4(3):328–350.

Tanja Samardzic and Paola Merlo. 2012. The meaning of lexical causatives in cross-linguistic variation. *Linguistic Issues in Language Technology*, 7(12):1–14.

Sabine Schulte Im Walde. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 747–753.

Diarmuid O Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444.

Dieci sfumature di marcatezza sintattica: verso una nozione computazionale di complessità

Erica Tusa
Università di Pisa
ericatusa@hotmail.it

Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi
Istituto di Linguistica Computazionale
"Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{nome.cognome}@ilc.cnr.it

Abstract

English. In this work, we will investigate whether and to what extent algorithms typically used to assess the reliability of the output of syntactic parsers can be used to study the correlation between processing complexity and the linguistic notion of markedness. Although still preliminary, achieved results show the key role of features such as dependency direction and length in defining the markedness degrees of a given syntactic construction.

Italiano. *In questo lavoro indagheremo se e come algoritmi tipicamente utilizzati per valutare l'affidabilità dell'analisi prodotta da strumenti di annotazione sintattica automatica possono essere utilizzati per studiare la correlazione tra complessità computazionale e la nozione linguistica di marcatezza. I risultati raggiunti, sebbene ancora preliminari, mostrano il ruolo chiave di fattori quali l'orientamento della relazione e la lunghezza della dipendenza nel definire le varie "sfumature" di marcatezza di una stessa relazione.*

1 Introduzione

Fin dagli anni '80, è andata affermandosi la convinzione che metodi e tecniche sviluppate nell'ambito della linguistica computazionale potessero contribuire a far avanzare la ricerca fornendo nuova evidenza per lo studio di nozioni chiave della linguistica teorica. "Computational linguistics provides important potential tools for the testing of theoretical linguistic constructs and of their power to predict actual language use": così si apre il contributo di Kučera (1982), che fondandosi sulla correlazione tra marcatezza e frequenza propone i risultati di uno studio computazionale della nozione di marcatezza a livello lessicale e grammaticale condotto sul Brown corpus. Se da un lato l'evidenza raccolta conferma la correlazione attesa tra frequenza e marcatezza, dall'altro vengono registrati casi interessanti in cui "the statistical evidence from the Brown Corpus offers both greater problems and greater insight. [...] The frequency data is [...] the reverse of what one might have assumed under the markedness analysis".

Oggi, a più di 40 anni dallo studio pioniestico di Kučera (1982) che si fondava su un corpus annotato morfo–sintatticamente di circa un milione di parole, è possibile estrarre da corpora di ben maggiori dimensioni informazione linguistica accurata e variegata. L'affidabilità crescente degli strumenti di Trattamento Automatico del Linguaggio sta rendendo infatti possibile l'acquisizione di evidenza quantitativa e computazionale che spazia attraverso diversi livelli di descrizione linguistica, incluso quello sintattico, che può contribuire in modo significativo allo studio di questioni aperte della letteratura linguistica: Merlo (2016) rappresenta un importante esempio di questo rinnovato filone di studi.

All'interno del quadro delineato sopra, l'obiettivo del presente contributo è verificare se e in che misura algoritmi per l'identificazione dell'affidabilità e plausibilità dell'annotazione sintattica possano contribuire allo studio di un fenomeno linguistico quale la marcatezza. Con l'uso sempre più diffuso dell'annotazione sintattica a dipendenze come punto di partenza per una vasta gamma di applicazioni finalizzate all'estrazione di informazione da vaste collezioni documentali, tali algoritmi nascono dalla necessità di identificare al-

l'interno delle annotazioni prodotte in modo automatico quelle corrette o, più semplicemente, caratterizzate da un maggior grado di affidabilità e plausibilità. Questo tipo di valutazione può avvenire in relazione sia all'intero albero sintattico assegnato alla singola frase (cfr. ad esempio Dell'Orletta *et al.* (2011) e Reichart and Rappoport (2009b)), sia alla singola relazione di dipendenza (si vedano, tra gli altri, Dell'Orletta *et al.* (2013) e Che *et al.* (2014)). Se da un lato l'identificazione di alberi sintattici corretti rappresenta un ingrediente fondamentale all'interno di algoritmi di Active Learning (Settles, 2012) o di apprendimento automatico semi–supervisionato e non supervisionato (Goldwasser *et al.*, 2011), l'identificazione dell'affidabilità di singole relazioni di dipendenza e/o sotto–alberi sintattici diventa fondamentale, ad esempio, per fornire evidenza utile a migliorare le prestazioni di un sistema di analisi sintattica automatica (vanNoord, 2007; Chen *et al.*, 2009), oppure per l'estrazione di nuclei di informazione affidabili.

Nel presente studio, ci focalizzeremo sul secondo tipo di algoritmi, ovvero quelli che operano a livello della singola relazione di dipendenza, per verificarne le potenzialità nello studio della nozione di marcatezza linguistica. Per quanto questi algoritmi operino tipicamente su corpora annotati in modo automatico (Dickinson, 2010), sono attestati usi anche su corpora con annotazione validata manualmente (qualificata come "gold"): in questo caso, il fine consiste nell'identificazione di errori e incoerenze di annotazione (Dickinson, 2015). Il risultato di questi algoritmi varia da una classificazione binaria della dipendenza (corretta $vs.$ errata) come in Che *et al.* (2014), a un ordinamento delle relazioni secondo l'affidabilità e la plausibilità dell'analisi, come proposto da Dell'Orletta *et al.* (2013). Al di là di differenze a livello dell'algoritmo utilizzato e del tipo di risultato, in tutti i casi viene fatto uso di un esteso inventario di caratteristiche linguistiche selezionate come indicatori di complessità.

2 L'ipotesi di ricerca

Combinando la prospettiva linguistica e quella linguistico–computazionale, l'ipotesi che intendiamo esplorare è se il punteggio assegnato da algoritmi per la misura della plausibilità dell'annotazione possa essere utilizzato per ricostruire il passaggio graduale da costruzioni non marcate a costruzioni caratterizzate da gradi crescenti di marcatezza. L'assunto di base sottostante a tale ipotesi si fonda sulla correlazione, ampiamente adottata nella letteratura linguistica, tra marcatezza e complessità: se da un lato costruzioni non marcate saranno caratterizzate da un maggior livello di plausibilità di annotazione (dunque da un minore livello di complessità), dall'altro costruzioni caratterizzate da gradi crescenti di marcatezza saranno associate a punteggi di minore di plausibilità (equivalente a una maggiore complessità).

All'interno della letteratura linguistica, la "marcatezza" rappresenta una nozione ampiamente dibattuta e altamente polisemica. Secondo quanto affermato da Haspelmath (2006), a partire dalle prime accezioni delineate negli anni '30 (Jakobson, 1932) essa può essere ricondotta a dodici significati diversi, organizzati in quattro classi: "markedness as complexity, as difficulty, as abnormality, or as a multidimensional operation". Tra queste, il presente contributo intende focalizzarsi sulla definizione di "markedness as abnormality" e, in particolare, sull'idea che quando consideriamo "marcato" un determinato evento linguistico lo stiamo considerando "abnormal", ovvero deviante rispetto a strutture linguistiche riconosciute come basiche all'interno della "norma linguistica". In questa ottica, la marcatezza come devianza rispetto alla norma è strettamente connessa sia con la frequenza d'uso (cfr. "markedness as rarity in texts"), sia rispetto alla distribuzione di un evento linguistico all'interno di una varietà più o meno ampia di contesti linguistici (cfr. "markedness as restricted distribution").

In quanto segue, analizzeremo i risultati di un algoritmo per la valutazione della plausibilità di singole relazioni alla luce delle accezioni di marcatezza selezionate.

3 Metodologia di analisi e corpora

L'algoritmo che abbiamo utilizzato per la misura della plausibilità dell'annotazione sulla base della quale produrre l'ordinamento delle relazioni di dipendenza è costituito da LISCA (Dell'Orletta *et al.*, 2013). Tale algoritmo assegna a ogni relazione – definita come una tripla (d, h, t) dove d è il dipendente, h è la testa, e t è il tipo di dipendenza che connette d a h – un valore di *plausibilità*. LISCA opera in due fasi: 1) colleziona statistiche relative a un insieme di caratteristiche linguistica-

mente motivate estratte da un ampio corpus di alberi a dipendenze ottenuti attraverso un processo di annotazione sintattica automatica; 2) combina queste statistiche all'interno di una funzione descritta in Dell'Orletta *et al.* (2013) per ottenere il punteggio da associare all'arco sintattico in corso di valutazione. La combinazione viene calcolata come il prodotto dei pesi associati a ciascuna caratteristica identificata.

La Figura 1 descrive graficamente le caratteristiche prese in esame da LISCA per la misura della plausibilità dell'arco sintattico (d, h, t). Ai fini del presente studio, LISCA è stato utilizzato nella sua variante delessicalizzata per poter fare astrazione da variazioni di natura lessicale. In particolare, sono stati presi in considerazione due diversi tipi di caratteristiche, entrambe associate nella letteratura linguistica alla nozione di complessità sintattica:

- tratti *locali*, corrispondenti alle peculiarità dell'arco sintattico considerato, come ad esempio la distanza in termini di tokens all'interno della frase tra d e h, oppure la forza associativa che unisce le categorie grammaticali coinvolte (POS_d e POS_h), o la POS della testa di h e il tipo di relazione sintattica che li lega;

- tratti *globali*, volti a localizzare l'arco considerato all'interno della struttura sintattica della frase, ad esempio la distanza di d rispetto alla radice dell'albero, oppure rispetto alla foglia più vicina o a quella più lontana, oppure il numero di nodi "fratelli" e "figli" di d ricorrenti rispettivamente alla sua destra–sinistra nell'ordine lineare della frase.

In questo studio, per estrarre le statistiche rispetto alle caratteristiche linguistiche prese in esame, LISCA è stato applicato a un corpus di 1.104.237 frasi (22.830.739 tokens) estratte da articoli del quotidiano *La Repubblica*, parte del CLIC-ILC Corpus (Marinelli et al., 2003). Il corpus è stato annotato a livello morfosintattico con l'ILC–POS–Tagger (Dell'Orletta *et al.*, 2009) e a livello sintattico a dipendenze con DeSR (Attardi *et al.*, 2009). Gli strumenti di annotazione sono stati addestrati sulla *Italian Universal Dependecies Treebank*, in breve IUDT (Bosco *et al.*, 2013). Lo schema di annotazione utilizzato è quello delle "Universal Dependencies", concepito per massimizzare il parallelismo delle annotazioni in lingue diverse e che per questo motivo privilegia relazioni

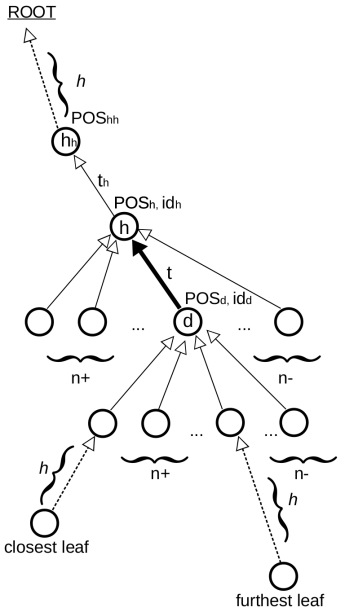


Figura 1: Caratteristiche utilizzate LISCA per il calcolo della plausibilità dell'arco (d, h, t).

di dipendenza tra parole lessicali trattando le parole grammaticali come dipendenti di parole semanticamente piene (Nivre, 2015). IUDT costituisce anche il corpus di indagine di questo lavoro: attraverso LISCA a ogni arco sintattico del corpus è stato assegnato un punteggio.

4 Analisi dei dati

I punteggi assegnati da LISCA alle relazioni di dipendenza della IUDT sono stati utilizzati per ordinare le relazioni in ordine descrescente di plausibilità. La lista ordinata così ottenuta è stata suddivisa in 10 fasce di 24,644 relazioni ciascuna (corrispondente al 10% del totale). Partendo dall'analisi della variazione della distribuzione dei tipi di relazioni di dipendenza attraverso le fasce, ci siamo poi focalizzati su singole relazioni, con l'intento di ricostruire il passaggio da non marcato (o prototipico) a marcato, e di identificare i fattori che contribuiscono a determinare il grado di marcatezza di una costruzione, definita in questo studio come una relazione di dipendenza all'interno del contesto sintattico di occorrenza. Nella consapevolezza che le relazioni sono distribuite all'interno delle fasce in virtù della combinazione di tutte le caratteristiche locali e globali prese in considerazione, ci siamo focalizzati su due parametri ampiamente indagati nella letteratura linguistica con l'intento di verificare se e in che misura l'ordinamento di

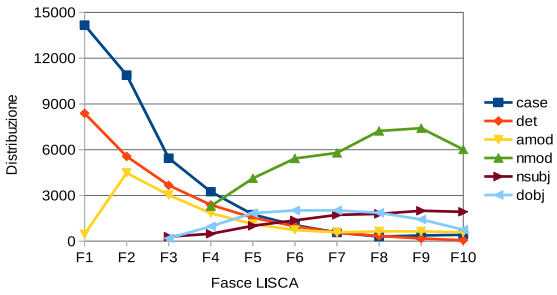


Figura 2: Distribuzione di una selezione di dipendenze.

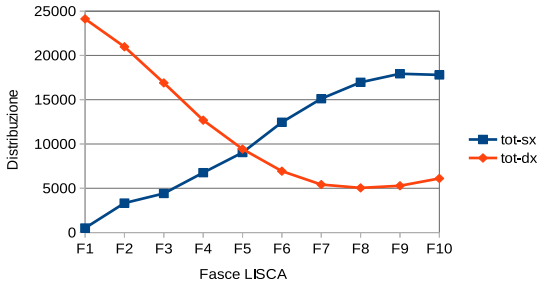


Figura 3: Orientamento generale delle dipendenze.

LISCA rifletta note gerarchie di marcatezza. In particolare, è stato analizzato il ruolo i) dell'orientamento della dipendenza, definito dalla direzione verso destra o sinistra dell'arco sintattico che lega d a h rispetto all'ordine lineare delle parole nella frase, e ii) della lunghezza della relazione di dipendenza, calcolata come la distanza in parole tra d e h.

4.1 Distribuzione delle dipendenze

Partiamo dall'analisi della distribuzione delle dipendenze nelle 10 fasce. Nelle prime fasce si osserva un insieme ristretto di tipi di dipendenze, la cui frequenza diminuisce proporzionalmente al decrescere dei punteggi assegnati da LISCA. Man mano che si prosegue verso le fasce intermedie, i tipi di dipendenza all'interno di ciascuna fascia si fanno più numerosi e variabili. Gli archi con i punteggi più alti, che si collocano nelle prime fasce, mettono in relazione parole grammaticali e parole lessicali, come *det*(erminer), *case* e *mark*(er), che collegano, rispettivamente, articoli, preposizioni, congiunzioni subordinanti o avverbi alla relativa testa. Proseguendo oltre le prime due fasce, troviamo relazioni come *advmod* (adverbial modifier), *nummod* (numerical modifier), *cop*(ula) e *aux*(iliary) che collegano, rispettivamente, avverbi, numerali, copule e ausiliari (verbi modali compresi) alla loro testa. A partire dalla quinta fascia si osserva un'incidenza sempre maggiore di relazioni che collegano parole lessicali come sostantivi e verbi alla loro testa, ad esempio *nsubj* (nominal subject), *dobj* (direct object), *ccomp* (clausal complement), *xcomp* (clausal complement with controlled subject) e *root*. Un caso a parte è rappresentato dalle relazioni *amod* e *nmod* che collegano modificatori aggettivali o nominali con la relativa testa: esse si distribuiscono in modo simile in tutte le fasce.

La Figura 2 riporta l'andamento della distribuzione di sei relazioni di dipendenza selezionate come segue: due relazioni concentrate principalmente nelle prime fasce (*det*, *case*), due relazioni caratterizzate da una distribuzione più diffusa (*amod*, *nmod*) e due relazioni maggiormente ricorrenti nelle ultime fasce (*nsubj*, *dobj*). La distribuzione delle relazioni che vedono una parola grammaticale come dipendente all'interno delle prime fasce può essere ricondotta alle strutture generalmente fisse o poco variabili, che le rendono facilmente trattabili computazionalmente. D'altro lato, le parole lessicali tendono ad inserirsi in costruzioni più complesse, caratterizzate da una maggiore flessibilità a livello dell'ordine lineare all'interno della frase, e potenzialmente soggette a condizionamenti di tipo pragmatico che portano alla formazione di strutture sintattiche più complesse. La presenza diffusa della relazione *amod* attraverso le fasce rappresenta un caso diverso, non riconducibile alla libertà di movimento ma piuttosto alla direzione degli archi sintattici che li collegano alla loro testa: mentre la distanza media tra d e h in *amod* rimane tendenzialmente costante attraverso le fasce, la direzione della relazione varia significativamente. In quanto segue, ci concentreremo su due dei parametri che sembrano svolgere un ruolo chiave nella distribuzione delle relazioni attraverso le fasce.

4.2 Orientamento delle dipendenze

La Figura 3 riporta la distribuzione attraverso le fasce di tutte le relazioni di dipendenza, facendo distinzione tra dipendenze con testa a destra ($d > h$) e dipendenze con testa a sinistra ($h < d$). Si osserva che i due tipi di orientamento, nonostante ricorrano con frequenza molto simile (112.886 $d > h$ vs 104.301 $h < d$), sono descritti da andamenti opposti: nelle prime fasce si concentrano le

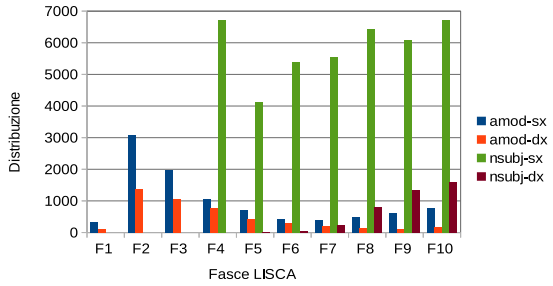


Figura 4: Orientamento di una selezione di dipendenze.

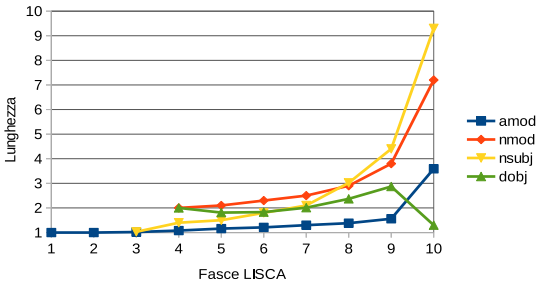


Figura 6: Andamento della lunghezza media di una selezione di dipendenze per fascia.

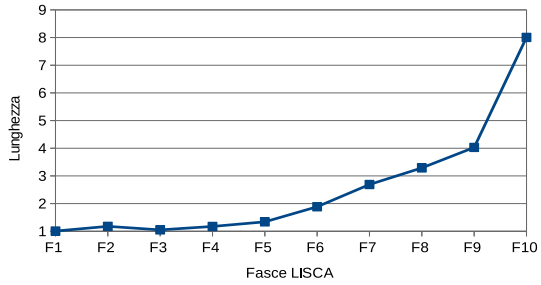


Figura 5: Andamento generale della lunghezza media delle dipendenze per fascia.

relazioni di tipo $d > h$, nelle ultime quelle $h < d$.

Nella Figura 4 è riportato l'andamento attraverso le fasce dell'orientamento di due relazioni, *amod* e *nsubj*. Nel caso di *amod* le teste dei modificatori aggettivali si trovano in netta maggioranza a sinistra, soprattutto nella seconda fascia: gli aggettivi postnominali, dunque in posizione non marcata, sono stati valutati come più plausibili e computazionalmente trattabili. Invece, *nsubj*, che collega il soggetto nominale alla testa verbale, nella maggior parte dei casi presenta la testa a destra: il soggetto preverbale corrisponde all'ordine non marcato in italiano. La sequenza verbo–soggetto è attestata, con andamento crescente, a partire dalla fascia 6. Le relazioni che sono state valutate con i punteggi più alti, ovvero *det* e *case*, mostrano la testa sempre a destra.

4.3 Lunghezza delle dipendenze

Nella Figura 5, per ciascuna fascia è riportata la media delle lunghezze delle relazioni di dipendenza: si osserva che il punteggio di LISCA decresce in maniera inversamente proporzionale al valore della lunghezza media. Nella Figura 6 sono riportate le medie all'interno di ogni fascia delle lunghezze di un gruppo selezionato di dipendenze: alcune delle relazioni che abbiamo visto concentrarsi nelle prime fasce, come *det* e *case*, ma anche relazioni dalla distribuzione più bilanciata come *amod* e *dobj* risultano essere quelle più brevi (la loro lunghezza in media non supera una distanza di 2 parole tra testa e dipendente); dipendenze come *nsubj* e *nmod*, che collegano unità dall'ordinamento più flessibile, cioè caratterizzate da una maggiore libertà di movimento rispetto alla loro testa, raggiungono in media, soprattutto verso le ultime fasce, le lunghezze maggiori. Considerando il ruolo ampiamente ascritto in letteratura, non solo linguistico–computazionale ma anche linguistica e psicolinguistica, alla lunghezza delle dipendenze come fattore di complessità linguistica, questi dati costituiscono una prova ulteriore che l'ordinamento prodotto da LISCA riflette la marcatezza della costruzione.

5 Conclusione

In questo studio abbiamo esplorato l'ipotesi che algoritmi sviluppati per valutare l'affidabilità e la plausibilità di annotazioni sintattiche a dipendenze possano fornire evidenza utile a una riflessione attorno al tema della complessità sintattica, e in particolare a ricostruire "sfumature" di marcatezza crescente in relazione alla stessa relazione di dipendenza. I primi risultati raggiunti sono incoraggianti: quanto osservato in relazione alla distribuzione delle dipendenze attraverso le fasce ci porta a ipotizzare una forte correlazione tra la complessità computazionale dell'analisi individuata da LISCA e la nozione di marcatezza sintattica. È stato indagato in particolare il ruolo di fattori quali l'orientamento della relazione e la lunghezza della dipendenza, con risultati che mostrano chiaramente che un algoritmo come LISCA può essere un valido strumento anche per analisi di tipo linguistico. Attraverso l'analisi della distribuzione delle

relazioni di dipendenza nelle fasce definite sulla base dell'ordinamento di LISCA è stato possibile non solo discriminare tra costruzioni marcate e non marcate (dato tipicamente recuperabile sulla base della frequenza), ma anche identificare – relazione per relazione – i fattori che hanno contribuito a renderla marcata. Se l'orientamento della relazione gioca un ruolo cruciale nel caso di *amod*, nel caso di *nsubj* è piuttosto la distanza tra la testa e il dipendente a determinare la marcatezza della costruzione. Ovviamente, questa metodologia di analisi dovrebbe essere estesa alla vasta tipologia di caratteristiche linguistiche considerate.

Gli sviluppi correnti di questo lavoro includono: l'estensione, al di là della lunghezza e l'orientamento della relazione, della tipologia di fattori linguistici esplorati, per arrivare anche allo studio dell'impatto di fattori lessicali; l'estensione della tipologia di costruzioni analizzate, che potrebbero anche includere combinazioni di dipendenze corrispondenti a sotto–alberi sintattici. Riteniamo che la metodologia dovrebbe anche essere applicata a treebank di lingue diverse, così come diversi generi testuali all'interno della stessa lingua.

References

Attardi G., Dell'Orletta F., Simi M., Turian J. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009*, Reggio Emilia, Italia, Dicembre 2009.

Bosco C., Montemagni, S., Simi, M. 2013. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *Proceedings of the ACL Linguistic Annotation Workshop & Interoperability with Discourse*, Sofia, Bulgaria, August 2013.

Che W., Guo J., Liu T. 2014. ReliAble dependency arc recognition. In *Expert Systems with Applications*. volume 41, number 4, pp. 17161722.

Chen W., Kazama J., Uchimoto K., Torisawa K. 2009. Improving Dependency Parsing with Subtrees from Auto-parsed Data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (EMNLP '09). Association for Computational Linguistics. volume 2, pp. 570–579.

Dell'Orletta F. 2009. Ensemble system for part-of-speech tagging. In *Proceedings of EVALITA 2009 - Evaluation of NLP and Speech Tools for Italian 2009* Reggio Emilia, Italia, Dicembre 2009.

Dell'Orletta F. 2011. ULISSE: an Unsupervised Algorithm for Detecting Reliable Dependency Parses. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL 2011, Portland, Oregon, USA, June 23-24, 2011*, pp. 115–124.

Dell'Orletta F., Venturi G., Montemagni S. 2013. Linguistically-driven Selection of Correct Arcs for Dependency Parsing. In *Computaciòn y Sistemas*. ISSN 1405-5546, vol. 17, No. 2, pp. 125-136.

Dickinson M. 2010. Detecting Errors in Automatically-Parsed Dependency Relations. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden, Association for Computational Linguistics, pp. 729–738.

Dickinson M. 2015. Detection of annotation errors in corpora. In *Language and Linguistics Compass*. ISSN 1749-818X, vol. 9, No. 3, pp. 119-138.

Goldwasser D., Reichart R., Clarke J., Roth D. 2011. Confidence Driven Unsupervised Semantic Parsing. In Proceedings of the *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA* (ACL–2011). pp. 1486–1495.

Haspelmath M. 2006. Against markedness (and what to replace it with). In *Journal of Linguistics*. ISSN 1469-7742, vol. 42, No. 01, pp. 25–70.

Jakobson R. 1932. Zur Struktur des russischen Verbums. In *Charisteria Gvilelmo Mathesio*.

Kučera H. 1982. Markedness and Frequency: a Computational Analysis. In *Proceedings of COLING 82*. pp. 167-173.

Marinelli R., L. Biagini, R. Bindi, S. Goggi, M. Monachini, P. Orsolini, E. Picchi, S. Rossi, N. Calzolari, A. Zampolli. 2003. *The Italian PAROLE corpus: an overview*. In Zampolli A. et al. (eds.), Computational Linguistics in Pisa, Special Issue, XVI–XVII, Pisa-Roma, IEPI. Tomo I, pp. 401-421.

Merlo P. 2016. Quantitative computational syntax: some initial results. In *Italian Journal of Computational Linguistics*. vol. 2.

Nivre J. 2015. Towards a Universal Grammar for Natural Language Processing. In *Proceedings of the 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Part I.* pp. 3–16.

Reichart Roi and Ari Rappoport. 2009b. *Sample Selection for Statistical Parsers: Cognitively Driven Algorithms and Evaluation Measures*. In Proceedings of CoNLL 2009, pp. 3–11.

Settles B. 2012. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers.

van Noord G. 2007. In *Proceedings of the 10th International Conference on Parsing Technologies* (IWPT–2007). Association for Computational Linguistics, pp. 1–10.

Online Automatic Post-Editing across Domains

Rajen Chatterjee, Gebremedhen Gebremelak, Matteo Negri, Marco Turchi
Fondazione Bruno Kessler, Italy
{chatterjee, gebremelak, negri, turchi}@fbk.eu

Abstract

English. Recent advances in automatic post-editing (APE) have shown that it is possible to automatically correct systematic errors made by machine translation systems. However, most of the current APE techniques have only been tested in controlled batch environments, where training and test data are sampled from the same distribution and the training set is fully available. In this paper, we propose an online APE system based on an instance selection mechanism that is able to efficiently work with a stream of data points belonging to different domains. Our results on a mix of two datasets show that our system is able to: *i)* outperform state-of-the-art online APE solutions and *ii)* significantly improve the quality of rough MT output.

Italiano. *Recenti miglioramenti dei sistemi automatici di post-editing hanno dimostrato la loro capacità di correggere errori ricorrenti commessi dalla traduzione automatica. Spesso, tuttavia, tali sistemi sono stati valutati in condizioni controllate dove i dati di training/test sono selezionati dalla stessa distribuzione e l'insieme di training è interamente disponibile. Questo articolo propone un sistema di post-editing online, basato su tecniche di selezione dei dati, capace di trattare sequenze di dati appartenenti a diversi dominii. I risultati su un insieme di dati misti mostrano che il sistema è in grado di ottenere risultati migliori rispetto i) allo stato dell'arte e ii) al sistema di traduzione.*

1 Introduction

Nowadays, machine translation (MT) is a core element in the computer-assisted translation (CAT) framework (Federico et al., 2014). The motivation for integrating MT in the CAT framework lies in its capability to provide useful suggestions for unseen segments, thus increasing translators productivity. However, it has been observed that MT is often prone to systematic errors that human post-editing has to correct before publication. The by-product of this "translation as post-editing" process is an increasing amount of parallel data consisting of MT output on one side and its corrected version on the other side. Besides being used to improve the MT system itself (Bentivogli et al., 2016), this data can be leveraged to develop automatic MT quality estimation tools (Mehdad et al., 2012; Turchi et al., 2013; C. de Souza et al., 2013; C. de Souza et al., 2014; C. de Souza et al., 2015) and automatic post-editing (APE) systems (Chatterjee et al., 2015b; Chatterjee et al., 2015a; Chatterjee et al., 2016). The APE components explored in this paper should be capable not only to spot recurring MT errors, but also to correct them. Thus, integrating an APE system inside the CAT framework can further improve the quality of the suggested segments, reduce the workload of human post-editors and increase the productivity of translation industries. In the last decade many studies on APE have shown that the quality of the machine translated text can be improved significantly by post-processing the translations with an APE system (Simard et al., 2007; Dugast et al., 2007; Terumasa, 2007; Pilevar, 2011; Béchara et al., 2011; Chatterjee et al., 2015b). These systems mainly follow the phrase-based machine translation approach where the MT outputs (with optionally the source sentence) are used as the source language corpus and the post-edits are used as the target language corpus. Although these standard

approaches showed promising results, they lack of the ability to continuously update their inner models by incorporating human feedback from a stream of data. To address this problem, several online systems have been proposed in MT, but only few of them have been applied to the APE scenario (Simard and Foster, 2013; Lagarda et al., 2015), only in a controlled working environment where they are trained and evaluated on homogeneous/coherent data sets.

In this paper, we propose a novel online APE system that is able to efficiently leverage data from different domains.[1] Our system is based on an instance selection technique that is able to retrieve the most relevant training instances from a pool of *multi-domain* data for each segment to post-edit. The selected data is then used to train and tune the APE system on-the-fly. The relevance of a training sample is measured by a similarity score that takes into account the context of the segment to be post-edited. This technique allows our online APE system to be flexible enough to decide if it has the correct knowledge for post-editing a sentence or if it is safer to keep the MT output untouched, avoiding possible damages. The results of our experiments over the combination of two data sets show that our approach is robust enough to work in a multi-domain environment and to generate reliable post-edits with significantly better performance than a state-of-the-art online APE system.

2 Online translation systems

Online translation systems aim to incorporate human post-editing feedback (or the corrected version of the MT output) into their models in real-time, as soon as it becomes available. This feedback helps the system to learn from the mistakes made in the past translations and avoid to repeat them in future translations. This continuous learning capability will eventually improve the quality of the translations and consequently increase the productivity of the translators/post-editors (Tatsumi, 2009) working with MT suggestions in a CAT environment. The basic workflow of an online translation system goes through the following steps repeatedly: *i)* the system receives an input segment; *ii)* the input segment is translated and provided to the post-editor to fix any errors in it; and *iii)* the human post-edited version of the translation is incorporated back into the system, by stepwise updating the underlying models and parameters. In the APE context, the input is a machine-translated segment (optionally with its corresponding source segment), which is processed by the online APE system to fix errors, and then verified by the post-editors. Several online translation systems have been proposed over the years (Hardt and Elming, 2010; Bertoldi et al., 2013; Mathur et al., 2013; Simard and Foster, 2013; Ortiz-Martïnez and Casacuberta, 2014; Denkowski et al., 2014; Wuebker et al., 2015).

The state-of-the-art online APE system is the Thot toolkit (Ortiz-Martïnez and Casacuberta, 2014) that has been previously developed to support fully automatic and interactive statistical machine translation and then used in the APE task (Lagarda et al., 2015). To update the inner models with the user feedback, a set of sufficient statistics was maintained and incrementally updated. In the case of language model, only the n-gram counts are required to maintain sufficient statistics. To update the translation model, an incremental version of EM algorithm is used to first obtain word alignment and then phrase pairs counts were extracted to update the sufficient statistics. Other features like source/target phrase-length models or distortion model are implemented by means of geometric distributions with fixed parameters. However, Thot differs from our approach because it does not embed any techniques for selecting the most relevant training data. In the long-run, when data points from different domain are continuously analysed, this system tends to become more and more generic, which may not be useful and even harmful for automatically post-editing domain-specific segments.

3 Instance Selection for online APE system

To preserve all the knowledge gained in the online learning process and at the same time being able to apply specific post-editing rules when needed, we propose an instance selection technique for online APE that has the ability to retrieve specific data points whose context is similar to the segment to be post-edited. These data points are then used to build reliable APE models. When there are no reliable data points in the knowledge base, the MT output is kept untouched, as opposed to the exist-

[1] A domain is made of segments belonging to the same text genre and the MT outputs are generated by the same MT system.

ing APE systems, which tends to always translate the given input segment independently from the reliability of the applicable correction rules.

Our proposed algorithm emulates an online APE system and assumes to have the following data to run the online experiments: *i)* source (*src*); *ii)* MT output (*mt*); and *iii)* human post-edits (*pe*) of the MT output. At the beginning the knowledge base of our online APE system is empty and it will be updated whenever an instance (a tuple containing parallel segments from all the above mentioned documents) is processed. When the system receives an input (*src*, *mt*), the most relevant training instances from a pool of *multi-domain* data stored in our knowledge base are retrieved. The similarity between the training instances and the input segment is measured by a score based on the term frequency−inverse document frequency (*tf-idf*), generally used in information retrieval. The larger the number of words in common between the training and the input sentences, the higher is the score. In our system, these scores are computed using the Lucene library.[2] Only those training instances that have similarity score above a certain threshold (decided over a held-out development set) are used to build: *i)* a tri-gram local language model over the target side of the training corpus with the IRSTLM toolkit (Federico et al., 2008); ii) the translation and reordering models using the Moses toolkit (Koehn et al., 2007) and the word alignment of each sentence pair is computed using the incremental GIZA++ software.[3] The log-linear model parameters are optimized over a part of the selected instances. To obtain reliably-tuned weights and a fast optimization process, multiple instances of MIRA (Chiang, 2012) are run in parallel on three small development sets randomly selected from the retrieved sentences. The obtained weights are then averaged. If a minimum value of retrieved sentences is not reached, the optimization step is skipped because having few sentences might not yield reliable weights. In this case, the weights computed on the previous input segment are used. The tuned weights and the models built on all the data are then used to post-edit the input sentences.

In a real translation workflow, the APE segment is then passed to the human translator that creates the post-edited segment. Once the post-edit is available it is added to the knowledge base along with the source and the mt sentences. In our experiments we emulate the post-edited sentence of the APE segment with the post-edit of the mt output.

4 Experimental setup

Data To examine the performance of the online APE systems in a *multi-domain* translation environment, we select two data sets for the English-German language pair belonging to information technology (IT). Although they come from the same category (IT), they feature variability in terms of vocabulary coverage, MT errors, and post-editing style. The two data sets are respectively a subset of the Autodesk Post-Editing Data corpus and the resources used at the second round of the APE shared task at the first conference on machine translation (WMT2016).[4] The data sets are pre-processed to obtain a joint-representation that links each source word with a MT word (*mt#src*). This representation has been proposed in the context-aware APE approach by (Béchara et al., 2011) and leverages the source information to disambiguate post-editing rules. Recently, (Chatterjee et al., 2015b) also confirmed this approach to work better than translating from raw MT segments over multiple language pairs. The joint-representation is used as a source corpus to train all the APE systems reported in this paper and it is obtained by first aligning the words of source (*src*) and MT (*mt*) segments using MGIZA++ (Gao and Vogel, 2008), and then each *mt* word is concatenated with its corresponding *src* words.

The Autodesk training, and development sets consist of 12,238, and 1,948 segments respectively, while the WMT2016 data contains 12,000, and 1,000 segments. To measure the diversity of the two data sets we compute the vocabulary overlap between the two joint-representations. This is performed internally to each data set (splitting the training data in two halves) and across them. As expected, in the first case the vocabulary overlap is much larger ($> 40\%$) than in the second one ($\sim 15\%$); this indicates that the two data sets are quite different and few information can be shared.

To emulate the multi-domain scenario, the two training data sets are first merged together and then shuffled. The same strategy is also used for the development sets. This represents the situation in

[2] https://lucene.apache.org/

[3] https://code.google.com/archive/p/inc-giza-pp/

[4] http://www.statmt.org/wmt16/ape-task.html

which an APE system serves two CAT tools that process documents from two domains and the sequence of points is random. Our approach and the competitors are run on all the shuffled training data and evaluated on the second half (12,100 points).

Evaluation metrics The performance of the different APE systems is evaluated using the Translation Error rate (TER) (Snover et al., 2006), BLEU (Papineni et al., 2002) and the precision (Chatterjee et al., 2015a). TER and BLEU measures the similarity between the MT outputs and their references by looking at the word/n-gram overlaps, while precision is the ratio of number of sentences an APE system improves (with respect to the MT output) over all the sentences it modifies.[5] Larger values indicate that the APE system is able to improve the quality of most of the sentences it changes. The statistical significance test for BLEU is computed using the paired bootstrap resampling technique (Koehn, 2004), and for TER using the stratified approximate randomization technique (Clark et al., 2011).

Terms of comparison We evaluate our online learning approach against the output produced by the *MT system*, the *batch APE system* that follows the approach proposed in (Chatterjee et al., 2015b), and the Thot toolkit.

5 Experiments and Results

The main goal of this research is to examine the performance of online APE methods in a multi-domain scenario, where the APE system receives a stream of data coming from different domains. The parameters of our approach (*i.e.* similarity score threshold and minimum number of selected sentence) are optimised following the grid search strategy. We set the threshold values to 1 and the minimum number of selected sentences to 20. The results of all the systems are reported in Table 1.

The *batch APE system* that is trained only on the first half of the data is able to slightly improve the performance of the *MT system*, but it damages most of the sentence it changes (precision smaller than 45%). Although Thot can learn from all the data, it is interesting to note that it does not significantly improve over the *MT system* and the *batch APE system*. This suggests that using all the data without considering the peculiarities of each domain does not allow an APE system to efficiently learn reliable correction rules and to improve the machine translation quality. Moreover, these results also show that few information can be shared between the two data sets. This is expected considering the limited overlap between the two corpora.

	BLEU	TER	Precision (%)
MT	52.31	34.52	N/A
Batch APE	52.52	34.45	42.67
Thot	52.51	34.37	42.22
Our approach	**53.97**†	**33.13**†	**64.82**

Table 1: Results on the mixed data. (†: statistically significant wrt. MT with $p<0.05$)

Our approach provides significant improvements in BLEU, TER and precision over all the competitors. In particular, it can obtain more than one TER and BLEU point improvement, and more than 20% precision points increment over the best APE system (the Thot toolkit). Such gains confirm that the instance selection mechanism allows our APE system to identify domain-specific data and to leverage it for extracting reliable correction rules. Further analysis of the performance of the online systems revealed that our approach modifies less segments compared with Thot, because it builds a model only if it finds relevant data, leaving the MT segment untouched otherwise. These untouched MT segments, when modified by Thot, often lead to deterioration. This suggests that, the output obtained with our solution has a higher potential for being useful to human translators. Such usefulness comes not only in terms of a more pleasant post-editing activity, but also in terms of time savings yield by overall better suggestions.

6 Conclusion

We addressed the problem of building a robust online APE system that is able to efficiently work on a stream of data points belonging to different domains. In this condition, our APE has shown its capability to continuously adapt to the dynamics of diverse data processed in real-time. In particular, the instance selection mechanism allows our APE method to reduce the number of wrong modifications, which result in significant improvements in precision over the state-of-the-art online APE system, and thus making it a viable solution to be deployed in a real-word CAT framework.

[5] For each sentence in the test set, if the TER score of APE system is different than the baseline then it is considered as a modified sentence

7 Acknowledgements

This work has been partially supported by the EC-funded H2020 project QT21 (grant agreement no. 645452).

References

Hanna Béchara, Yanjun Ma, and Josef van Genabith. 2011. Statistical post-editing for a statistical mt system. In *Proceedings of the XIII MT Summit*, pages 308–315.

Luisa Bentivogli, Nicola Bertoldi, Mauro Cettolo, Marcello Federico, Matteo Negri, and Marco Turchi. 2016. On the evaluation of adaptive machine translation for human post-editing. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 24(2):388–399.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. *Proceedings of the XIV MT Summit*, pages 35–42.

José G. C. de Souza, Christian Buck, Marco Turchi, and Matteo Negri. 2013. FBK-UEdin Participation to the WMT13 Quality Estimation Shared Task. In *Proc. of the Eighth Workshop on Statistical Machine Translation*, pages 352–358, Sofia, Bulgaria. Association for Computational Linguistics.

José G. C. de Souza, Jesús González-Rubio, Christian Buck, Marco Turchi, and Matteo Negri. 2014. FBK-UPV-UEdin Participation in the WMT14 Quality Estimation Shared-task. In *Proc. of the Ninth Workshop on Statistical Machine Translation*, pages 322–328, Baltimore, Maryland, USA.

José G. C. de Souza, Matteo Negri, Elisa Ricci, and Marco Turchi. 2015. Online multitask learning for machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 219–228, Beijing, China, July. Association for Computational Linguistics.

Rajen Chatterjee, Marco Turchi, and Matteo Negri. 2015a. The fbk participation in the wmt15 automatic post-editing shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 210–215.

Rajen Chatterjee, Marion Weller, Matteo Negri, and Marco Turchi. 2015b. Exploring the planet of the apes: a comparative study of state-of-the-art methods for mt automatic post-editing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 156–161, July.

Rajen Chatterjee, José G. C. de Souza, Matteo Negri, and Marco Turchi. 2016. The fbk participation in the wmt 2016 automatic post-editing shared task. In *Proceedings of the First Conference on Machine Translation*, pages 745–750, Berlin, Germany, August. Association for Computational Linguistics.

David Chiang. 2012. Hope and fear for discriminative training of statistical translation models. *Journal of Machine Learning Research*, 13(Apr):1159–1187.

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 176–181.

Michael Denkowski, Chris Dyer, and Alon Lavie. 2014. Learning from post-editing: Online model adaptation for statistical machine translation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 395–404, April.

Loïc Dugast, Jean Senellart, and Philipp Koehn. 2007. Statistical post-editing on systran's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220–223.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. Irstlm: an open source toolkit for handling large scale language models. In *Proceedings of Interspeech*, pages 1618–1621.

Marcello Federico, Nicola Bertoldi, Mauro Cettolo, Matteo Negri, Marco Turchi, Marco Trombetti, Alessandro Cattelan, Antonio Farina, Domenico Lupinetti, Andrea Martines, Alberto Massidda, Holger Schwenk, Loïc Barrault, Frederic Blain, Philipp Koehn, Christian Buck, and Ulrich Germann. 2014. The matecat tool. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 129–132, Dublin, Ireland, August.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Proceedings of Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

Daniel Hardt and Jakob Elming. 2010. Incremental re-training for post-editing smt. In *Proceedings of AMTA*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pages 177–180.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, pages 388–395.

Antonio L Lagarda, Daniel Ortiz-Martïnez, Vicent Alabau, and Francisco Casacuberta. 2015. Translating without in-domain corpus: Machine translation post-editing with online learning techniques. *Computer Speech & Language*, 32(1):109–134.

Prashant Mathur, Mauro Cettolo, Marcello Federico, and FBK-Fondazione Bruno Kessler. 2013. Online learning approaches in computer assisted translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation, ACL*, pages 301–308.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2012. Match without a Referee: Evaluating MT Adequacy without Reference Translations. In *Proceedings of the Machine Translation Workshop (WMT2012)*, pages 171–180, Montréal, Canada, June.

Daniel Ortiz-Martïnez and Francisco Casacuberta. 2014. The new thot toolkit for fully-automatic and interactive statistical machine translation. In *14th Annual Meeting of the European Association for Computational Linguistics: System Demonstrations*, pages 45–48.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Abdol Hamid Pilevar. 2011. Using statistical post-editing to improve the output of rule-based machine translation system. *IJCSC*.

Michel Simard and George Foster. 2013. Pepr: Post-edit propagation using phrase-based statistical machine translation. In *Proceedings of the XIV MT Summit*, pages 191–198.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical Phrase-Based Post-Editing. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 508–515.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, pages 223–231.

Midori Tatsumi. 2009. Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. In *Proceedings of the XII MT Summit*, pages 332–339.

Ehara Terumasa. 2007. Rule based machine translation combined with statistical post editor for japanese to english patent translation. In *Proceedings of the XI MT Summit*, pages 13–18.

Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the Subjectivity of Human Judgements in MT Quality Estimation. In *Proc. of the Eighth Workshop on Statistical Machine Translation*, pages 240–251, Sofia, Bulgaria. Association for Computational Linguistics.

Joern Wuebker, Spence Green, and John DeNero. 2015. Hierarchical incremental adaptation for statistical machine translation. In *Proceedings of EMNLP*, pages 1059–1065, September.

Language resources for Italian: towards the development of a corpus of annotated Italian multiword expressions

Shiva Taslimipoor
University of Wolverhampton, UK
shiva.taslimi@wlv.ac.uk

Anna Desantis, Manuela Cherchi
University of Sassari, Italy
annadesantis_91@libero.it,
manuealacherchi82@gmail.com

Ruslan Mitkov
University of Wolverhampton, UK
r.mitkov@wlv.ac.uk

Johanna Monti
"L'Orientale" University of Naples, Italy
jmonti@unior.it

Abstract

English. This paper describes the first resource annotated for multiword expressions (MWEs) in Italian. Two versions of this dataset have been prepared: the first with a fast markup list of out-of-context MWEs, and the second with an in-context annotation, where the MWEs are entered with their contexts. The paper also discusses annotation issues and reports the inter-annotator agreement for both types of annotations. Finally, the results of the first exploitation of the new resource, namely the automatic extraction of Italian MWEs, are presented.

Italiano. *Questo contributo descrive la prima risorsa italiana annotatata con polirematiche. Sono state preparate due versioni del dataset: la prima con una lista di polirematiche senza contesto, e la seconda con annotazione in contesto. Il contributo discute le problematiche emerse durante l'annotazione e riporta il grado di accordo tra annotatori per entrambi i tipi di annotazione. Infine vengono presentati i risultati del primo impiego della nuova risorsa, ovvero l'estrazione automatica di polirematiche per l'italiano.*

1 Rationale

Multiword expressions (MWEs) are a pervasive phenomenon in language with their computational treatment being crucial for users and NLP applications alike (Baldwin and Kim, 2010; Granger and Meunier, 2008; Monti et al., 2013; Monti and Todirascu, 2015; Seretan and Wehrli, 2013). However, despite being desiderata for linguistic analysis and language learning, as well as for training and evaluation of NLP tasks such as term extraction (and Machine Translation in multilingual scenarios), resources annotated with MWEs are a scarce commodity (Schneider et al., 2014b). The need for such types of resources is even greater for Italian which does not benefit from the variety and volume of resources as does English.

This paper outlines the development of a new language resource for Italian, namely a corpus annotated with Italian MWEs of a particular class: verb-noun expressions such as *fare riferimento*, *dare luogo* and *prendere atto*. Such collocations are reported to be the most frequent class of MWEs and of high practical importance both for automatic translation and language learning. To the best of our knowledge, this is the first resource of this kind in Italian.

The development of this corpus is part of a multilingual project addressing the challenge of computational treatment of MWEs. It covers English, Spanish, Italian and French and its goal is to develop a knowledge-poor methodology for automatically identifying MWEs and retrieving their translations (Taslimipoor et al., 2016) for any pair of languages. The developed methodology will be used for Machine Translation and multilingual dictionary compilation, and also in computer-aided tools to support the work of language learners and translators.

Two versions of the above resource have been produced. The first version consists of lists of MWEs annotated out-of-context with a view to performing fast evaluation of the developed methodology (out-of-context mark-up). The second version consists of annotated MWEs along with their concordances (in-context annotation).

The latter type of annotation is time-consuming, but provides the contexts for the MWEs annotated.

2 Annotation of MWEs: out-of-context mark-up and in-context annotation

After more than two decades of computational studies on MWEs, the lack of a proper gold standard is still an issue. Lexical resources like dictionaries have limited coverage of these expressions (Losnegaard et al., 2016) and there is also no proper tagged corpus of MWEs in any language (Schneider et al., 2014b).

Most previous studies on the computational treatment of MWEs have focused on extracting types (rather than tokens)[1] of MWEs from corpora (Ramisch et al., 2010; Villavicencio et al., 2007; Rondon et al., 2015; Salehi and Cook, 2013). The widely-used toolboxes of MWEToolkit (Ramisch et al., 2010) or Xtract (Smadja, 1993) extract expressions if their statistical occurrences represent the likelihood of them being MWEs. The evaluation for the type-based extraction of MWEs has been mostly performed against a dictionary (de Caseli et al., 2010), lexicon (Pichotta and DeNero, 2013) or list of human-annotated expressions (Villavicencio et al., 2007). However, there are some examples like the expression *have a baby*, which in exactly the same form and structure, might be an MWE (meaning *to give birth*) in some contexts and a literal expression in others.

As for the automatic identification of the tokens of MWEs, Fazly et al. (2009) make use of both linguistic properties and the local context, in determining the class of an MWE token. They report an unsupervised approach to identifying idiomatic and literal usages of an expression in context. Their method is evaluated on a very small sample of expressions in a small portion of the British National Corpus (BNC), which were annotated by humans. Schneider et al. (2014a) developed a supervised model whose purpose is to identify MWEs in context. Their methodology results in a corpus of automatically annotated MWEs. It is not clear, however, if the methodology is able to tag one specific expression as an MWE in one context and non-MWE in another. The PARSEME shared task[2] is also devoted to annotating verbal MWEs in several languages. The shared task, while having interesting discussions on the area, has embarked upon the labour-intensive annotation of verbal MWEs.

Since there is no list of verb-noun MWEs in Italian, we first automatically compile a list of such expressions, to be annotated by human experts. This is based on previous attempts at extracting a lexicon of MWEs (as in (Villavicencio, 2005)). Annotators are not provided with any context and hence the task is more feasible in terms of time. Human annotators are asked to label the expressions as MWEs only if they have sufficient degrees of idiomaticity. In other words, a Verb + Noun MWEs does not convey literal meaning in that the verb is delexicalised.

However, we believe that idiomaticity is not a binary property; rather it is known to fall on a continuum from completely semantically transparent, or literal, to entirely opaque, or idiomatic (Fazly et al., 2009). This makes the task of out-of-context marking-up of the expression more challenging for annotators, since they have to pick a value according to all the possible contexts of a target expression. This ambiguity and the fact that there are many expressions that in some contexts are MWEs and in some contexts not, prompted us to initiate a subsequent annotation where MWEs are tagged in their contexts. The idea is to extract the concordances around all the occurrences of a Verb + Noun expression and provide annotators with these concordances in order to be able to decide the degree of idiomaticity of the specific verb-noun expression. We compare the reliability of the in-context and out-of-context annotations by way of the agreement between annotators.

2.1 Experimental expressions

Highly polysemous verbs, such as *give* and *take* in English and *fare* and *dare* in Italian widely participate in Verb+Noun MWEs, in which they contribute a broad range of figurative meanings that must be recognised (Fazly et al., 2007). We focus on four mostly frequent Italian verbs: *fare*, *dare*, *prendere* and *trovare*. We extract all the occurrences of these verbs when followed by any noun, from the itWaC corpus (Baroni and Kilgarriff, 2006), using SketchEngine (Kilgarriff et al., 2004). For the first experiment all the Verb+Noun types are extracted when the verb is lemmatised; and for the second experiment all the concor-

[1] Type refers to the canonical form of an expression, while token refers to each instance (usage) of the expression in any morphological form in text.

[2] http://typo.uni-konstanz.de/parseme/index.php/2-general/142-parseme-shared-task-on-automatic-detection-of-verbal-mwes

dances of these verbs when followed by a noun are generated.

2.2 Out-of-context mark-up of Verb+Noun(s)

The extraction of Verb+Noun candidates of the four verbs in focus and the removal of the expressions with frequencies lower than 20, results in a dataset of $3,375$ expressions. Two native speakers annotated every candidate expression with 1 for an MWE if the expression was idiomatic and with 0 for a non-MWE if the expression was literal. We have also defined the tag 2 for the expressions that in some contexts behave as MWEs and in others do not, e.g. *dare frutti*, which has a literal usage that means *to produce fruits* but in some contexts means *to produce results* and is an MWE in these contexts. While this out-of-context 'fast track' annotation procedure saves time and yields a long list of marked-up expressions, annotators often feel uncomfortable due to the lack of context. The information about the agreements between annotators in terms of $Kappa$ is shown in Table 2 and is compared with the in-context annotation of MWEs as explained in Section 2.3.

2.3 Annotating Verb+Noun(s) in context

We design an annotation task, in which we provide a sample of all usages of any type of Verb+Noun expression to be annotated. For this purpose, we employ the SketchEngine to list all the concordances of each verb when it is followed by a noun. Concordances include the verb in focus with almost ten words before and ten words after that. The SketchEngine reports only $100,000$ concordances for each query. Among them, we filter out the concordances that include Verb+Noun expressions with frequencies lower than 50 and we randomly select 10% of the concordances for each verb. As a result, there are $30,094$ concordances to be annotated. The two annotators annotate all usages of Verb+Noun expressions in these concordances, considering the context that the expression occurred in, marking up MWEs with 1 and expressions which are not MWEs, with 0. Table 1 reports on the details of annotation tasks and Table 2 shows the agreement details for them.

2.4 Discussion

As seen in Table 2, the inter-annotator agreement is significantly higher when annotating the expressions in context. One of the main causes of disagreements in out-of-context annotation is concerned with abstract nouns. The annotation of expressions composed of a verb followed by a noun with an abstract meaning is a more complicated process as the candidate expression may carry a figurative meaning. Each annotator uses their intuition to annotate them and it leads to random tags for these expression (e.g. *fare notizia*, *dare identità*, *prendere possesso*) when they are out-of-context. However, in the case of in-context annotation, concordances composed of abstract nouns have been annotated in the majority of cases with 1 by both annotators.

In-context annotation is also very helpful for annotating expressions with both idiomatic and literal meanings. An interesting observation, reported in Table 3, is related to the number of expressions that are detected with the two different usages of idiomatic and non-idiomatic, in context.

Table 1: Annotation details (A: Annotator)

Annotation task	A	tag 0	tag 1 (MWE)	tag 2
Out-of-context	1st	2,491	792	92
	2nd	2,112	1,127	136
In-context	1st	10,478	19,616	-
	2nd	9,058	21,036	-

Table 2: Inter-annotator agreement

Annotation task	Kappa	Observed Agreement
Out-of-context	0.40	0.73
In-context	**0.65**	**0.85**

Table 3: Statistics on the in-context annotation

	0 tagged	1 tagged	context depending
1st annotator	924	195	530
2nd annotator	696	424	529

As can be seen in Table 3,[3] among the $1,649$ types of expressions in concordances, 530 (32%) of them could be MWEs in some context and non-MWEs in others (context-depending), according to the first annotator. This annotator has annotated only 3% of the expressions with tag '2' without context.

[3]Note that the numbers in Table 3 cannot be interpreted to validate agreement between annotators, i.e. no conclusion about agreement can be derived from 3.

3 First use of the MWE resource: comparative evaluation of the automatic extraction of Italian MWEs

In our multilingual project (see Section 1) we regard the automatic translation of MWEs as a two-stage process. The first stage is the extraction of MWEs in each of the languages; the second stage is a matching procedure for the extracted MWEs in each language which proposes translation equivalents. In this study the extraction of MWEs is based on statistical association measures (AMs).

These measures have been proposed to determine the degree of compositionality, and fixedness of expressions. The more compositional or fixed expressions are, the more likely it is that they are MWEs (Evert, 2008; Bannard, 2007). According to Evert (2008), there is no ideal association measure for all purposes. We aim to evaluate AMs as a baseline approach against the annotated data which we prepared. We focus on a selection of five AMs which have been more widely discussed to be the best measures to identify MWEs. These are: MI3 (Oakes, 1998), log-likelihood (Dunning, 1993), T-score (Krenn and Evert, 2001), log-Dice (Rychlý, 2008) and Salience (Kilgarriff et al., 2004) all as defined in SketchEngine. We compare the performance of these AMs and also frequency of occurrence (Freq) as the sixth measure to rank the candidate MWEs. We evaluate the effect of these measures in ranking MWEs on both kinds of datasets.

3.1 Experiments on type-based extraction of MWEs

In the first experiment, the list of all extracted Verb + Noun combinations (as explained in Section 2.1) are ranked according to the above measures that are computed from itWaC as a reference corpus. To perform the evaluation against the list of annotated expressions, we process all 2,415 expressions for which the annotators agreed on tags 0 or 1. After ranking the expressions by the measures, we examine the retrieval performance of each measure by computing the 11-point Interpolated Average Precision (11-p IAP). This reflects the goodness of a measure in ranking the relevant items (here, MWEs) before the irrelevant ones. To this end, the interpolated precision at the 11 recall values of 0, 10%, ..., 100% is calculated. As detailed in Manning et al. (2008), the interpolated precision at a certain recall level, r, is defined as the highest precision found for any recall level $r' \geq r$. The average of these 11 points is reported as 11-p IAP in Table 4.

Table 4: 11-p IAP for ranking MWEs using different AMs

AMs	11-p IAP
Freq	0.49
MI3	0.51
log-likelihood	0.49
Salience	0.49
log-dice	0.48
T-Score	0.49

Table 5: Accuracy of AMs in classifying usages of Verb+Noun(s).

AMs	Accuracy
Freq	0.72
MI3	0.68
log-likelihood	0.72
Salience	0.69
log-dice	0.67
T-Score	0.69

As can be seen in Table 4, the selected association measures generally perform with similar performance in ranking this type of MWEs, with $MI3$ performing slightly better than others.

3.2 Experiments on token-based identification of MWEs

In the second experiment, we seek to establish the effect of these measures on identifying the usages of MWEs in our dataset of in-context annotations. We set a threshold for each score that we have computed for Verb+Noun expression types. By setting thresholds we compute the classification accuracy of the measures to identify MWEs among the usages of Verb+Noun expressions in a corpus. Specifically, each candidate of a Verb+Noun in the concordances is automatically tagged as an MWE if its lemmatised form has a score higher than the threshold, and as a non-MWE, otherwise. For each measure, we compute the arithmetic mean (average) of all the values of that measure for all expressions, and set the resulted average value as a threshold.

The accuracies of classifying the candidate Verb+Noun expressions are computed based on the human annotations of the concordances and are shown in Table 5. The classification accuracies of AMs are also very close to each other (see Table 5); however, this time $Log\text{-}likelihood$ and $Freq$ fare slightly better than others in classifying tokens of Verb+Noun expressions.

3.3 Usage-related features

Our new resource of concordances contains useful linguistic information related to usages of expressions and as such important features can be

extracted from the resource to help identifying MWEs. One of these features can be obtained from the statistics of different possible inflections of the verb component of an expression. Based on the premise of the fixedness of MWEs, we expect that the verb component of a verb-noun MWE occurs only in a limited number of inflections. We implement this feature by dividing the frequency of occurrences of each expression by the number of inflections that the verb component occurs in. Note that to count the number of different inflections of the verb component, we rely on the subcorpus of concordances that we gathered.

We evaluate this approach only on 1,077 expressions that occur in concordances. We rank the expressions according to this newly computed score and we call this score, which depends on the inflection varieties, INF-VAR. For all verbs, the INF-VAR performs comparably to Frequency in ranking MWEs higher than non-MWEs, but for the verb *trovare*, we obtain better 11-p IAP using this score than by using Frequency (see Table 6).

Table 6: Performance of new scores in ranking MWEs in terms of 11-p IAP.

	total	*trovare*
Frequency	0.57	0.44
INF-VAR	0.58	**0.48**

4 Conclusions and future work

In this paper, we outline our work towards a gold-standard dataset which is tagged with Italian verb-noun MWEs along with their contexts. We show the reliability of this dataset by its considerable inter-annotator agreement compared to the moderate inter-annotator agreement on annotated verb-noun expressions presented without context. We also report the results of automatic extraction of MWEs using this dataset as a gold-standard. One of the advantages of this dataset is that it includes both 0-tagged and 1-tagged tokens of expressions and it can be used for classification and other statistical NLP approaches. In future work, we are interested in extracting context features from concordances in this resource to automatically recognise and classify the expressions that are MWEs in some contexts but not MWEs in others.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing, second edition.*, pages 267–292. CRC Press.

Colin Bannard. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pages 1–8. Association for Computational Linguistics.

Marco Baroni and Adam Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations*, EACL '06, pages 87–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Helena Medeiros de Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language resources and evaluation*, 44(1-2):59–77.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *COMPUTATIONAL LINGUISTICS*, 19(1):61–74.

Stefan Evert. 2008. Corpora and collocations. In *Corpus Linguistics. An International Handbook*, volume 2, pages 1212–1248.

Afsaneh Fazly, Suzanne Stevenson, and Ryan North. 2007. Automatically learning semantic knowledge about multiword predicates. *Language Resources and Evaluation*, 41(1):61–89.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Sylviane Granger and Fanny Meunier. 2008. *Phraseology: an interdisciplinary perspective*. John Benjamins Publishing Company.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz, and David Tugwell. 2004. The sketch engine. In *EURALEX 2004*, pages 105–116, Lorient, France.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? a case study on extracting pp-verb collocations. *Proceedings of the ACL Workshop on Collocations*, pages 39–46.

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. Parseme survey on mwe resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

Johanna Monti and Amalia Todirascu. 2015. Multiword units translation evaluation in machine translation: another pain in the neck? In *Proceedings of MUMTTT workshop, Corpas Pastor G, Monti J, Mitkov R, Seretan V (eds) (2015), Multi-word Units in Machine Translation and Translation Technology*.

Johanna Monti, Ruslan Mitkov, Gloria Corpas Pastor, and Violeta Seretan. 2013. Multi-word units in machine translation and translation technologies.

Michael P. Oakes. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Karl Pichotta and John DeNero. 2013. Identifying phrasal verbs using many bilingual corpora. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, Seattle, WA, October.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: a Framework for Multiword Expression Identification. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valetta, Malta, May. European Language Resources Association.

Alexandre Rondon, Helena Caseli, and Carlos Ramisch. 2015. Never-ending multiword expressions learning. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 45–53, Denver, Colorado, June. Association for Computational Linguistics.

Pavel Rychlý. 2008. A lexicographer-friendly association score. In *RASLAN 2008*, pages 6–9, Brno. Masarykova Univerzita.

Bahar Salehi and Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, 1:266–275.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014a. Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *TACL*, 2:193–206.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014b. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 455–461, Reykjavik, Iceland. European Language Resources Association (ELRA).

Violeta Seretan and Eric Wehrli. 2013. Syntactic concordancing and multi-word expression detection. *International Journal of Data Mining, Modelling and Management*, 5(2):158–181.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143–177.

Shiva Taslimipoor, Ruslan Mitkov, Gloria Corpas Pastor, and Afsaneh Fazly. 2016. Bilingual contexts from comparable corpora to mine for translations of collocations. In *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics*, CICLing'16. Springer.

Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *EMNLP-CoNLL*, pages 1034–1043.

Aline Villavicencio. 2005. The availability of verb–particle constructions in lexical resources: How much is enough? *Computer Speech & Language*, 19(4):415–432.

Emojitalianobot and EmojiWorldBot

New online tools and digital environments for translation into emoji

Johanna Monti
L'Orientale University
Naples, Italy
jmonti@unior.it

Federico Sangati
Independent Researcher
The Netherlands
federico.sangati@gmail.com

Francesca Chiusaroli
University of Macerata
Italy
f.chiusaroli@unimc.it

Martin Benjamin
EPFL
Lausanne, Switzerland
martin@kamusiproject.org

Sina Mansour
EPFL
Lausanne, Switzerland
mansour@ee.sharif.edu

Abstract

English. Emojitalianobot and EmojiWorldBot are two new online tools and digital environments for translation into emoji on Telegram, the popular instant messaging platform. Emojitalianobot is the first open and free Emoji-Italian and Emoji-English translation bot based on Unicode descriptions. The bot was designed to support the translation of Pinocchio into emoji carried out by the followers of the "Scritture brevi" blog on Twitter and contains a glossary with all the uses of emojis in the translation of the famous Italian novel. EmojiWorldBot, an off-spring project of Emojitalianobot, is a multilingual dictionary that uses Emoji as a pivot language from dozens of different languages. Currently the emoji-word and word-emoji functions are available for 72 languages imported from the Unicode tables and provide users with an easy search capability to map words in each of these languages to emojis, and vice versa. This paper presents the projects, the background and the main characteristics of these applications.

Italiano. *Emojitalianobot e EmojiWorldBot sono due applicazioni online per la traduzione in e da emoji su Telegram, la popolare piattaforma di messaggistica istantanea. Emojitalianobot è il primo bot aperto e gratuito di traduzione che contiene i dizionari Emoji-Italiano ed Emoji-Inglese basati sule descrizioni Unicode. Il bot è stato ideato per coadiuvare la traduzione di Pinocchio in emoji su Twitter da parte dei follower del blog Scritture brevi e contiene pertanto anche il glossario con tutti gli usi degli emoji nella traduzione del celebre romanzo per ragazzi. EmojiWorldBot, epigono di Emojitalianobot, è un dizionario multilingue che usa gli emoji come lingua pivot tra dozzine di lingue differenti. Attualmente le funzioni emoji-parola e parola-emoji sono disponibili per 72 lingue importate dalle tabelle Unicode e forniscono agli utenti delle semplici funzioni di ricerca per trovare le corrispondenze in emoji delle parole e viceversa per ciascuna di queste lingue. Questo contributo presenta i progetti, il background e le principali caratteristiche di queste applicazioni.*

1 Introduction

Emojitalianobot[1] and *EmojiWorldBot*[2] are two new translation bots[3] into and from emoji. These two bots were designed starting from the hypothesis of setting up an emoji multilingual dictionary and translator through a process of selection and assessment of conventional semantic values. Translation cases may show how images can convey common and universal meanings, beyond specific peculiarities, so as they can stand as models in the perspective of an interlanguage (Chiusaroli, 2015). The two

[1] https://telegram.me/emojitalianobot/
[2] https://telegram.me/emojiworldbot
[3] Computer programmes that carry out repetitive tasks and in their more sophisticated form can also simulate human behaviours.

bots ease the use of emojis but also collect, refine and make available valuable linguistic data by means of crowdsourcing and gamification approaches.

This contribution presents the state-of-the-art concerning the use of crowdsourcing and gamification approaches to linguistics in section 2, the *Emojitalianobot* and the *Pinocchio* project in section 3, the *EmojiWorldBot* in section 4 and finally conclusions and future work in section 5.

2 Crowdsourcing and gamification

Crowdsourcing, i.e., the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call (Howe, 2006) is becoming a widespread practice on the Internet to develop linguistic resources (dictionaries, glossaries, translation memories, etc.) or services (translation, localisation, fansubbing, etc.) (Monti, 2012, 2014). It allows the large scale involvement of users who contribute with their knowledge, their ideas, and their skills, in this way performing an active role in the achievement of a common goal. Crowdsourcing can be used for the creation, maintenance and sharing of lexical/terminological data such as: i. lexical resources for online dictionaries, e.g., Wiki platforms such as Wiktionary[4] and Omegawiki[5], and recent forays by more traditional dictionary publishing companies like Collins, Oxford, and Macmillan; ii. terminological resources for online terminological databases, like TermWiki[6] , the terminological counterpart of Wiktionary or TaaS[7]; iii. lexical and semantic resources for Natural Language Processing (NLP) tasks, such as Word Sense Disambiguation (WSA), Sentiment Analysis, Computer Aided Translation, Machine Translation and so on, using platforms for distributing parts of large development projects to professional or occasional lexicographers such as Mechanical Turk[8]. To the best of our knowledge only very few projects so far have been tailored to mobile devices to gather linguistic data in the field, (i) to collect dialect data as in Dialectbot[9], (ii) to document endangered languages as in Aikuma[10] and Ma Iwaidja[11], or (ii) to gather grammaticality judgments (Madnani et al., 2011). The social dimension of these types of activities is sometimes connected and fed by social communities, where users discuss problems, give suggestions, and exchange ideas (Brabham, 2012; McGonigal, 2011). In order to loyalize social communities and improve their engagement, gamification is used very often. The use of games is a very effective tool for active participation since it provides a strong motivational framework which pushes people to act for good. Some effective uses of games are to create new habits or modify wrong actions. Wang et al. (2013) list Games with a purpose (GWAPs)[12] among the different types of crowdsourcing. Some good examples of games with a purpose in the lexicographic field are Phrase Detectives[13] and JeuxDeMots[14]. The main advantage of GWAPs is their high attractiveness, because people love playing games and it is easier to obtain their contribution in this way in comparison to other forms of crowdsourcing. The difficulty in designing such games is to match attractiveness with usefulness, i.e. an attractive game which produces valuable data.

3 Emojitalianobot and the Pinocchio project

Emojitalianobot is the first open and free Emoji-Italian translation bot on Telegram. It was developed to support the translation project of *Pinocchio* in emoji [15] launched on Twitter in February 2016 by F. Chiusaroli, J. Monti and F. Sangati. The translation of the famous children's novel was carried out by the followers of the *Scritture brevi* blog[16] (by F. Chiusaroli and F.M. Zanzotto) and the first fifteen chapters have been translated, which correspond to the original novel published by Collodi in 1881. Every day tweets with sentences taken from the novel were posted on Twitter and the followers suggested their translations

[4] https://en.wiktionary.org/
[5] http://www.omegawiki.org/Meta:Main_Page
[6] http://it.termwiki.com/
[7] https://term.tilde.com/
[8] https://www.mturk.com/mturk/welcome
[9] https://telegram.me/dialectbot/

[10] http://www.aikuma.org/aikuma-app.html
[11] https://itunes.apple.com/au/app/ma-iwaidja/id557824618?mt=8
[12] When a player without any special knowledge is put into a gaming environment and has to make decisions to win the game under the pressure of time or any game mechanics' constraints.
[13] https://anawiki.essex.ac.uk/phrasedetectives/
[14] http://www.jeuxdemots.org/jdm-accueil.php
[15] http://www.treccani.it/lingua_italiana/speciali/ludolinguistica/Chiusaroli.html
[16] https://www.scritturebrevi.it/

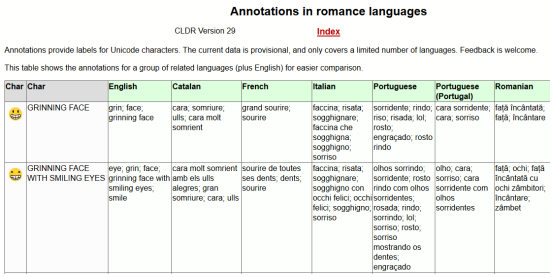
in emoji; at the end of each day, the official version of the translations was validated and published.[17] Translators used *Emojitalianobot* that contains (i) the Emoji-Italian dictionary, (ii) the Emoji-English descriptions based on Unicode and (iii) a glossary with all the uses of emoji in the translation of Pinocchio. The project was associated with the Emojitalia discussion group on Telegram, where users met to discuss problems, solutions, suggest improvements of the bot, in addition to the translation choices for *Pinocchio* and *communicate* in emoji. The *Pinocchio* translation project therefore allowed to crowdsource different linguistic data connected with the use of emojis as actual means of communication and not just simple graphics to express amusement or interest. In this respect the main findings of the project are twofold: the need to recur to compound multi-emoji expressions in order to express concepts which are not represented in the current set, as well as a related simple grammar to express syntactic relations among emojis, past and future tenses, etc. Unlike previous literary translation project in emojis, such as the translations of *Moby Dick* or *Alice in Wonderland*, this is the first attempt of a collective shared emoji code (vocabulary and grammar) based on a word for word translation totally in emojis. *Emojitalianobot* is an ideal test bench to experiment with new approaches like crowdsourcing and gamification in the field of Natural Language Processing (NLP). The *Pinocchio* project, games and features available in the bot to learn or guess the meaning of emoji are devised indeed both to enjoy the bot while using it and at the same time to give the opportunity to users to develop linguistic descriptions of emoji tailored on their actual perceptions. The most important reward for playing with the bot is the awareness of helping develop a linguistic resource for one's mother tongue, and the pride in contributing to it.

Since its release on Telegram, the project was an instant success, becoming a viral web phenomenon thanks to the *Scritture brevi* community and the *Pinocchio* translation in emojis, so that the bot has now almost 750 users. The Pinocchio translation project in emojis counts 611 tweets , 980 glossary entries which correspond to 2127 words, of which 185 are multi-emojis, i.e. compound emojis, such as 🔝👎 for the Italian word *peggio* (worst).

[17]The translation of *Pinocchio* in emoji can be followed on Twitter using #emojitaliano.

4 EmojiWorldBot

On the basis of (both linguistic and technological) experience with *Emojitalianobot*, the three Italian researchers together with Martin Benjamin and Sina Mansour of the Kamusi Project International[18] and EPFL (Switzerland) designed a new bot on Telegram in April 2016: *EmojiWorldBot*, a multilingual dictionary that uses Emoji as a pivot language from dozens of different languages. Currently the emoji-word and word-emoji functions are available for 70 languages imported from the Unicode tables [19] and provide users with an easy search capability to map words in each of these languages to emojis, and vice versa. Looking at the UNICODE descriptions (see Fig. 1) it is apparent that emojis are not annotated in a coherent way across languages, so some languages have more descriptions and some others, especially underrepresented languages, have less or in the most cases some languages are not represented at all.



Annotations in romance languages

CLDR Version 29 Index

Annotations provide labels for Unicode characters. The current data is provisional, and only covers a limited number of languages. Feedback is welcome.

This table shows the annotations for a group of related languages (plus English) for easier comparison.

Char	Char	English	Catalan	French	Italian	Portuguese	Portuguese (Portugal)	Romanian
😀	GRINNING FACE	grin; face; grinning face	cara; somriure; ulls; cara molt somrient	grand sourire; sourire	faccina; risata; sogghignare; faccina che sogghigna; sogghigno; sorriso	sorridente; rindo; riso; risada; lol; rosto; engraçado; rosto rindo	cara sorridente; cara; sorriso	față încântată; față; încântare
😂	GRINNING FACE WITH SMILING EYES	eye; grin; face; grinning face with smiling eyes; smile	cara molt somrient amb els ulls alegres; gran somriure; cara; ulls	sourire de toutes ses dents; dents; sourire	faccina; risata; sogghignare; sogghigno con occhi felici; occhi felici; sogghigno; sorriso	olhos sorrindo; sorridente; rosto rindo com olhos sorridentes; rosada; rindo; sorrindo; lol; sorriso; rosto; sorriso mostrando os dentes; engraçado	olho; cara; sorriso; cara sorridente com olhos sorridentes	față; ochi: față încântată cu ochi zâmbitori; încântare; zâmbet

Figure 1: Annotations in Romance languages

Our first goal with *EmojiWorldBot* is therefore to reach a uniform and comprehensive list of tags across multiple languages with a precise mapping between any language pair, which may serve to bootstrap a massive multilingual dictionary. The bot currently features:

- emoji-to-word and word-to-emoji translation for more than 70 languages
- *Eggs*, a tagging game for people to contribute to the expansion of these dictionaries or the creation of new ones for any additional language. Users can suggest additional tags for single emojis in any language (for example adding *egg* to the tag list for 🐣 in English).

[18]https://kamusi.org/

[19]http://www.unicode.org/cldr/charts/29/annotations/

- inline queries: type EmojiWorldBot and a word, and it will suggest a set of emojis for that word you can send in any Telegram conversation
- the possibility to add new languages.To date 56 new languages were added, such as Latin, Esperanto, Sardinian among others.

The basic idea of the *Eggs* game is to collect new tags to associate with emojis as shown in Fig. 2.

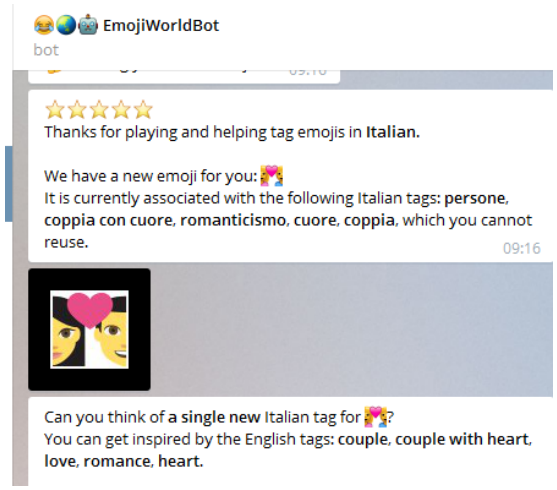


Figure 2: *Eggs* game

With fewer than 2000 official emojis, stretching the boundaries of their communicative potential makes them more useful. However, it also makes the dictionary more essential, so that someone who receives 🍆 in a chat in any language might look to see if it signifies something other than an *eggplant*. In the future, *Eggs* will experiment with multi-emoji terms (METs), building on the work of the *Pinocchio* translation project to Emoji, in an effort to build a larger pictorial vocabulary that is comprehensible across languages (Chiusaroli, 2015). A new version of the bot is already under development. It will feature *Ducks*, a second game where users are asked to map tags from a source language (e.g. English) to a target language (e.g. Swahili). In the example of Figure 1, several Romanian users would be shown the sense-specific definition of *grin* from Wordnet and all of the emojis that have been attached to that definition, and be asked which of the options among *față încântată față* and *încântare*, if any, is a good translation. The game would also be played for *face* and *grinning face*. In this way, all of the one-to-one relationships should be discovered, and all instances of a term that does not have a translation equivalent on the other side will be revealed. When it is known that no match from English exists, *Ducks* presents the definition, the emojis, and the English term, and asks the user to type in the best equivalent in their language. This is the method that will be most efficacious for new languages, bypassing the need to disentangle the many-to-many associations introduced through term clustering in the CLDR annotations. It should be noted that many terms will be removed from the game cycle through comparisons with Wordnets for available languages. For example, самолет appears in conjunction with English *airplane* in both the Emoji annotations and the Bulgarian Wordnet that are linked to the same English Princeton Wordnet (PWN) sense, which gives sufficient confirmation without needing a mass of human players. As of this writing, the project is in the process of importing and aligning Wordnet data for some 50 languages. In future work, terms from Wordnet synsets will be tested against the emojis with which they theoretically share a sense, e.g. asking crowd members whether 🚍 applies to other members of the PWN synset for *bus* (autobus, coach, jitney, motorbus, etc.), but the mechanism for doing so has not been finalized. In this way *EmojiWorldBot* employs crowd methods as part of an arsenal intended to conquer the walls of collecting data for numerous diverse languages. Data validation will be achieved via a consensus model through which answers are accepted as correct if the same result is provided by a threshold number of respondents. The new version of the bot will allow to:

- add new terms to the current languages (including the names of the countries for national flags)
- compare definitions across languages.

From the computational point of view, this project, as the *Emojitalianobot*, attempts to address the data chasm for natural language processing for most languages by distilling data collection to simple micro-tasks (Benjamin, 2015) using techniques adapted to least-common-denominator technology.

5 Conclusions

We described the *Emojitalianobot* and the *EmojiWorldBot* projects. Combining crowd-

sourcing, gamification and a smartphone app is a powerful strategy to collect, improve and refine valuable linguistic data easily and in a short time particularly for less-resourced languages (Benjamin and Radetzky, 2014).These may be the first crowdsourcing projects of this type to use bots for linguistic data collection and validation and are unique in their attempts at engaging participants for different languages.

References

Martin Benjamin. 2015. Crowdsourcing microdata for cost-effective and reliable lexicography. In *Proceedings of AsiaLex 2015 Hong Kong*, EPFL-CONF-215062, pages 213–221.

Martin Benjamin and Paula Radetzky. 2014. Multilingual lexicography with a focus on less-resourced languages: Data mining, expert input, crowdsourcing, and gamification. In *9th edition of the Language Resources and Evaluation Conference*, EPFL-CONF-200375.

Daren C Brabham. 2012. A model for leveraging online communities. *The participatory cultures handbook*, 120.

Francesca Chiusaroli. 2015. La scrittura in emoji tra dizionario e traduzione. *CLiC it*, page 88.

Jeff Howe. 2006. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

Nitin Madnani, Joel Tetreault, Martin Chodorow, and Alla Rozovskaya. 2011. They can help: Using crowdsourcing to improve the evaluation of grammatical error detection systems. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 508–513. Association for Computational Linguistics.

Jane McGonigal. 2011. *Reality is broken: Why games make us better and how they can change the world*. Penguin.

Johanna Monti. 2012. Translators' knowledge in the cloud: The new translation technologies. In *International Symposium on Language and Communication: Research Trends and Challenges(ISLC)*.

Johanna Monti. 2014. Dictionaries in the cloud: state of the art, trends and challenges. *Les Cahiers du dictionnaire*, (6):95–110.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language resources and evaluation*, 47(1):9–31.

IR Scientific Data: How to Semantically Represent and Enrich Them

Extended abstract of (Silvello et al., 2016)

Toine Bogers
Aalborg University Copenhagen, Denmark
toine@hum.aau.dk

Georgeta Bordea **Paul Buitelaar**
Insight Centre, National University of Ireland, Galway, Ireland
{georgeta.bordea, paul.buitelaar}@insight-centre.org

Nicola Ferro **Gianmaria Silvello**
University of Padua, Padua, Italy
{ferro, silvello}@dei.unipd.it

Abstract

English. Experimental evaluation carried out in international large-scale campaigns is a fundamental pillar of the scientific and technological advancement of Information Retrieval (IR) systems. Such evaluation activities produce a large quantity of scientific and experimental data, which are the foundation for all the subsequent scientific production and development of new systems. We discuss how to annotate and interlink this data, by proposing a method for exposing experimental data as Linked Open Data (LOD) on the Web and as a basis for enriching and automatically connecting this data with expertise topics and expert profiles. In this context, a topic-centric approach for expert search is proposed, addressing the extraction of expertise topics, their semantic grounding with the LOD cloud, and their connection to IR experimental data.

Italiano. *La valutazione sperimentale condotta mediante campagne internazionali su larga scala, è un pilastro fondante dello sviluppo scientifico e dell'avanzamento tecnologico dei sistemi di reperimento dell'informazione. Queste attività di valutazione producono una grande quantità di dati sperimentali che costituiscono la base per la conseguente produzione scientifica e lo sviluppo di nuovi sistemi. In questo lavoro, si discute come annotare e collegare questi dati, proponendo un metodo per esporre i dati sperimentali come LOD nel Web e per usare tali dati come base per arricchirli. In questo contesto, viene proposto un approccio centrato sui topic per la ricerca di esperti, che affronta il problema dell'estrazione dei topic e il collegamento di questi con la "LOD cloud" e con i dati sperimentali.*

1 Introduction

The importance of research data is widely recognized across all scientific fields as this data constitutes a fundamental building block of science. Recently, a great deal of attention was dedicated to the nature of research data (Borgman, 2015) and how to describe, share, cite, and re-use them in order to enable reproducibility in science and to ease the creation of advanced services based on them (Ferro et al., 2016; Silvello and Ferro, 2016).

Nevertheless, in the field of Information Retrieval (IR), where experimental evaluation based on shared data collections and experiments has always been central to the advancement of the field (Harman, 2011), the Linked Open Data (LOD) paradigm has not been adopted yet and no models or common ontologies for data sharing have been proposed. So despite the importance of data to IR, the field does not share any clear ways of exposing, enriching, and re-using experimental data as LOD with the research community.

Therefore, the main contributions of this paper are to:

- define an Resource Description Framework (RDF) model of the scientific IR data with the aim of enhancing their discoverability and easing their connections with the scientific production related to and based on them;

- provide a methodology for automatically enriching the data by exploiting relevant external entities from the LOD cloud.

2 Use Case: Discover, Understand and Re-use IR Experimental Data

In this section, we discuss an example of the outcomes of the semantic modeling and automatic enrichment processes applied to the use case of discovering, understanding and re-using the experimental data. Figure 1 shows an RDF graph, which provides a visual representation of how the experimental data are enriched. In particular, we can see the relationship between a contribution and an author enriched by expertise topics, expert profiles and connections to the LOD cloud, as supported by the Distributed Information Retrieval Evaluation Campaign Tool (DIRECT) system which provides the conceptual model for representing and enriching the data (Agosti and Ferro, 2009; Agosti et al., 2012).

In this instance, the author (*Jussi Karlgren*) and the contribution (*KarlgrenEtAl-CLEF2012*) are data derived from the evaluation workflow, whereas all the other information are automatically determined by the enrichment process. The adopted methodology for expertise topics extraction determined two main topics, "reputation management" and "information retrieval", which are related to the *KarlgrenEtAl-CLEF2012* contribution. We can see that *KarlgrenEtAl-CLEF2012* is featured by "reputation management" with a score of 0.53 and by "information retrieval" with 0.42, meaning that both these topics are subjects of the contribution; the scores (normalized in the interval $[0, 1]$) give a measure of how much this contribution is about a specific topic and we can see that in this case it is concerned a bit more with reputation management than with information retrieval. Furthermore, the backward-score gives us additional information by measuring how much a contribution is authoritative with respect to a scientific topic. In Figure 1, we can see that *KarlgrenEtAl-CLEF2012* is authoritative for reputation management (backward-score of 0.87), whereas it is not a very important reference for information retrieval (backward-score of 0.23). Summing up, we can say that if we consider the relation between a contribution and an expertise topic, the score indicates the pertinence of the expertise topic within the contribution; whereas the backward score indicates the pertinence of the contribution within the expertise topic. The higher the backward score, the more pertinent is the contribution for the given topic.

This information is confirmed by the expert profile data; indeed, looking at the upper-left part of Figure 1, the author *Jussi Karlgren* is considered "an expert in" reputation management (backward-score of 0.84), even if it is not his main field of expertise (score of 0.46).

All of this automatically extracted information enriches the experimental data enabling for a higher degree of re-usability and understandability of the data themselves. In this use case, we can see that the expertise topics are connected via an `owl:sameAs` property to external resources belonging to the DBPedia[1] linked open dataset. These connections are automatically defined via the semantic grounding methodology described below and enable the experimental data to be easily discovered on the Web. In the same way, authors and contributions are connected to the DBLP[2] linked open dataset.

In Figure 1 we can see how the contribution (*KarlgrenEtAl-CLEF2012*) is related to the experiment (*profiling_kthgavagai_1*) on which it is based. This experiment was submitted to the *RepLab 2012* of the evaluation campaign *CLEF 2012*. It is worthwhile to highlight that each evaluation campaign in DIRECT is defined by the name of the campaign (CLEF) and the year it took place (e.g., 2012 in this instance); each evaluation campaign is composed of one or more tasks identified by a name (e.g., RepLab 2012) and the experiments are treated as submissions to the tasks. Each experiment is described by a contribution which reports the main information about the research group which conducted the experiment, the system they adopted, developed and any other useful detail about the experiment.

We can see that most of the reported information are directly related to the contribution and they allow us to explicitly connect the research data with the scientific publications based on them. Furthermore, the experiment is evaluated from the "effectiveness" point of view by using the "accuracy" measurement which has 0.77 score. Retaining and exposing this information as LOD on the Web allow us to explicitly connect the

[1] `http://www.dbpedia.org/`
[2] `http://dblp.l3s.de/`

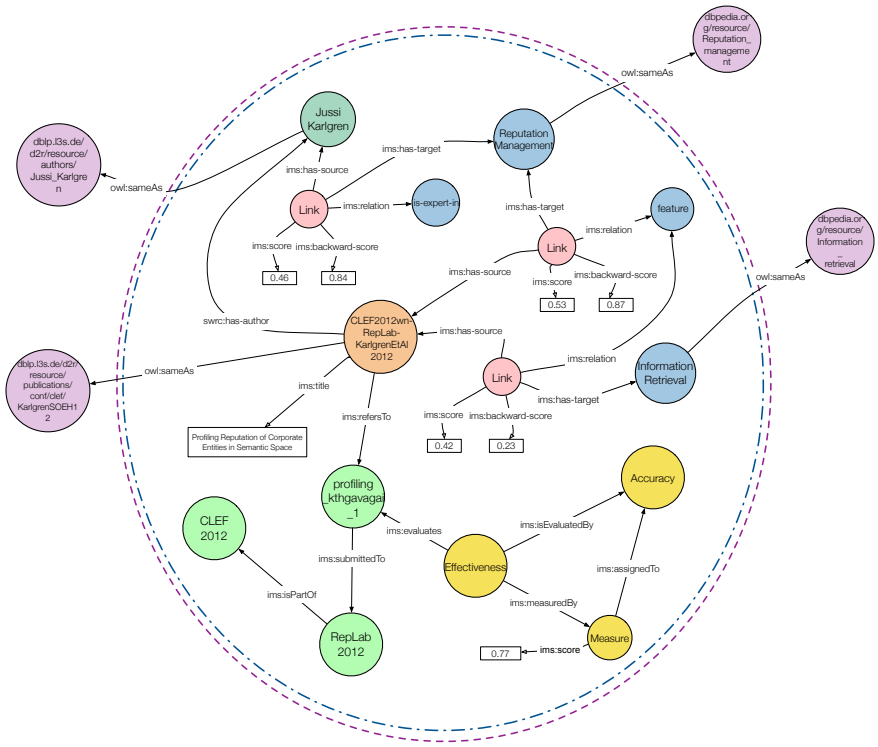


Figure 1: An example of RDF graph showing how expertise topics and expert profiles are used for enriching IR experimental data.

results of the evaluation activities to the claims reported by the contributions.

The details of the full RDF model are reported in (Silvello et al., 2016).

2.1 Accessing the Experimental Data

The described RDF model has been realized by the DIRECT system which allows for accessing the experimental evaluation data enriched by the expert profiles created by means of the techniques that will be described in the next sections. This system is called LOD-DIRECT and it is available at the URL: `http://lod-direct.dei.unipd.it/`.

The data currently available include the contributions produced by the Conference and Labs of the Evaluation Forum (CLEF) evaluation activities, the authors of the contributions, information about CLEF tracks and tasks, provenance events and the above described measures. Furthermore, this data has been enriched with expert profiles and expertise topics which are available as linked data as well.

At the time of writing, LOD-DIRECT allows access to $2,229$ contributions, $2,334$ author profiles and $2,120$ expertise topics. Overall, $1,659$ experts have been individuated and on average there are 8 experts per expertise topics (an expert can have more than one expertise of course).

The URIs of the resources are constructed following the pattern:

```
base-path/{resource-name}/
{id};{ns}
```

where,

- `base-path` is `http://lod-direct.dei.unipd.it`;
- `resource-name` is the name of the resource to be accessed as defined in the RDF model presented above;
- `id` is the identifier of the resource of interest;
- `ns` is the namespace of the resource of interest, this applies only for the namespace identifiable resources.

As an example, the URI corresponding to the contribution resource shown in Figure 1 with identifier `CLEF2012wn-RepLab-KarlgrenEt2012b` is:

```
http://lod-direct.dei.
unipd.it/contribution/
CLEF2012wn-RepLab-KarlgrenEt2012b
```

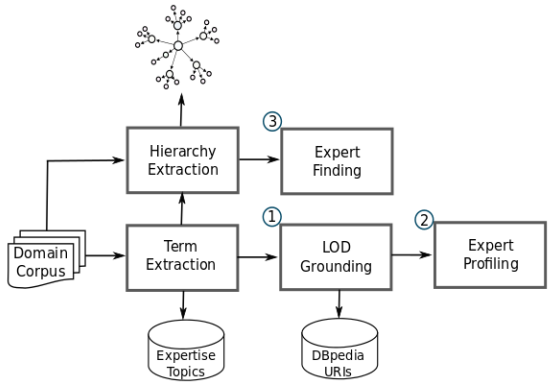


Figure 2: Data flow of the semantic enrichment approach

3 Semantic Enrichment

In this section we describe SOME methods for semantically enriching experimental IR data modelled as described above, by analysing unstructured data available in scientific publications. Figure 2 presents an overview of the semantic enrichment of documents and authors based on term and topical hierarchy extraction. First, we propose a method to automatically extract expertise topics from a domain-specific collection of publications using an approach for term extraction. Then, we present a preliminary approach for enriching expertise topics by grounding them in the LOD cloud.

Topic-centric approaches for expert search emphasize the extraction of keyphrases that can succinctly describe expertise areas, also called expertise topics, using term extraction techniques (Bordea et al., 2012). Expertise topics are extracted from a domain-specific corpus using the following approach. First, candidate expertise topics are discovered from text using a syntactic description for terms (i.e., nouns or noun phrases) and contextual patterns that ensure that the candidates are coherent within the domain. A domain model is constructed using the method proposed in (Bordea et al., 2013) and then noun phrases that include words from the domain model or that appear in their immediate context are selected as candidates.

These topics describe core concepts of the domain such as *search engine*, *IR system*, and *retrieval task*, as well as prominent subfields of the domain including *image retrieval*, *machine translation*, and *question answering*.

Only the best 20 expertise topics are stored for each document, ranking expertise topics based on their overall score. In this way, each document is enriched with keyphrases, taking into consideration the quality of a term for the whole corpus in combination with its relevance for a particular document.

Table 1: Precision and recall for DBpedia URI extraction

Approach	Precision	Recall	F-score
String Matching	0.96	**0.93**	0.94
Lemmatisation	**0.99**	0.90	0.94

Expertise topics can be used to provide links between IR experimental data and other data sources. These links play an important role in cross-ontology question answering, large-scale inference and data integration (Ngonga Ngomo, 2012). Additional background knowledge, as found on the LOD cloud, can inform expert search at different stages.

A first step in the direction of exploiting this potential is to provide an entry point in the LOD cloud through DBpedia[3]. Our goal is to associate as many terms as possible with a concept from the LOD cloud through DBpedia URIs—as shown in the use-case above. Where available, concept descriptions are collected as well and used in our system.

Two approaches for grounding expertise topics on DBpedia have been evaluated. The first approach matches a candidate DBpedia URI with an expertise topic, using the string as it appears in the corpus. The second approach makes use of the lemmatised form of the expertise topic. In order to evaluate our URI discovery approach, we build a small gold standard dataset by manually annotating 186 expertise topics with DBpedia URIs. First of all, we note that about half of the analysed expertise topics have a corresponding concept in DBpedia. One of the main reasons for the low coverage is that DBpedia is a general knowledge datasource that has a limited coverage of specialised technical domains.

Although both approaches achieve similar results in terms of F-score, the approach that makes use of lemmatisation (A2) achieves better precision, as can be seen in Table 1. Surprisingly, using lemmatization achieves a lower recall but higher precision but this might be due to the small size of the dataset.

Expert finding is the task of identifying the most

[3]DBpedia: `http://dbpedia.org/`

knowledgeable person for a given expertise topic. In this task, several competent people have to be ranked based on their relative expertise on a given expertise topic. We compare several topic-centric methods for expert finding with two language-modelling baselines.

The results for the expert finding task are presented in Table 2. The expert finding methods evaluated in this section include Experience (E), Relevance and Experience (RE) and Relevance, Experience and Area Coverage (REC).

Experience (E) is based on the idea that documents written by a person can be used as an indirect evidence of expertise, assuming that an expert often mentions his areas of interest. Relevance and Experience (RE) exploits the idea that expertise is closely related to the notion of experience. The assumption is that the more a person works on a topic, the more knowledgeable they are. We estimate the experience of a researcher on a given topic by counting the number of publications that have the topic assigned as a top ranked keyphrase. Relevance and expertise measure different aspects of expertise and can be combined to take advantage of both features. In the case that the subtopics of an expertise topic are known, we can evaluate the expertise of a person based on their knowledge of the more specialised fields. A previous study showed that experts have increased knowledge at more specific category levels than novices (Tanaka and Taylor, 1991). We introduce a novel measure for expertise called *Area Coverage* (REC) that measures whether an expert has in depth knowledge of an expertise topic, using an automatically constructed topical hierarchy.

The Area Coverage measure makes use of a topical hierarchy. Therefore we automatically construct a topical hierarchy for IR using the method proposed in (Hooper et al., 2012). Figure 3 shows a small extract from this hierarchy that correctly identifies "information retrieval" as the root of the taxonomy as well as several subfields including "digital libraries", "interactive information retrieval", and "cross language information retrieval".

The details on the algorithms and weighting schemes for topic extraction, expert profiling, and expert finding are reported in (Silvello et al., 2016).

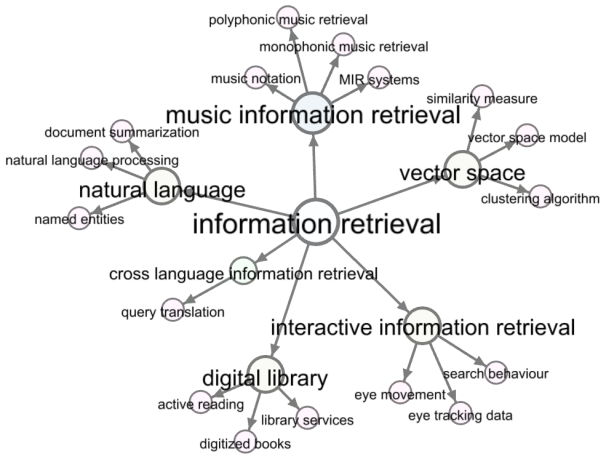


Figure 3: Sample hierarchical relations for the IR domain

4 Conclusion

In this paper we discussed the data modelling and the semantic enrichment of IR experimental data, as produced by large-scale evaluation campaigns.

In particular, the main results of the paper are:

- an accurate RDF data model for describing IR experimental data in detail, available at `http://ims.dei.unipd.it/data/rdf/direct.3.10.ttl`;
- a dataset about CLEF contributions, extracted expertise topics and related expert profiles;
- the online accessible LOD DIRECT system, available at `http://lod-direct.dei.unipd.it/`, to access the above data in different serialization formats, RDF+XML, Turtle, N3, XML and JSON.

Future work will concern the application of these semantic modeling and automatic enrichment techniques to other areas of the evaluation workflow. For example, expert profiling and topic extraction could be used to automatically improve and enhance the descriptions of the single experiments submitted to an evaluation campaign, which are typically not very rich and often cryptic—for example "second iteration with tuned parameters" as description—and to automatically link experiments to external resources, e.g., describing the used components, such as stemmers or stop lists, and systems. Finally, the RDF model defined within DIRECT opens up the possibility of integrating established Digital Library (DL) methodologies for data access and management which in-

Dataset	Measure	LM1	LM2	E	RE	REC
CL	MAP	0.0071	0.0056	0.0335	0.0335	**0.0340**
	MRR	0.0631	0.0562	0.2734	0.2738	**0.2754**
	P@5	0.0202	0.0173	0.1340	0.1339	**0.1347**
SW	MAP	0.0070	0.0067	**0.0327**	0.0305	0.0314
	MRR	0.0528	0.0522	**0.2262**	0.2115	0.2095
	P@5	0.0182	0.0188	**0.1065**	0.0967	0.0994
IR	MAP	0.0599	0.0402	0.1592	**0.1669**	0.1657
	MRR	0.1454	0.1231	0.4056	**0.4141**	0.4120
	P@5	0.0614	0.0485	0.1771	0.1771	**0.1783**
UvT	MAP	**0.2009**	0.1994	0.1155	0.1151	0.1158
	MRR	0.3551	**0.3571**	0.2298	0.2266	0.2281
	P@5	**0.1357**	0.1347	0.0850	0.0846	0.0841

Table 2: Expert finding results for the language modelling approach (LM), Experience (E), Relevance and Experience (RE), and Relevance, Experience and Area Coverage (REC)

creasingly exploit the LOD paradigm (Hennicke et al., 2011; Di Buccio et al., 2013). This would enable broadening the scope and the connections between IR evaluation and other related fields, providing new paths for semantic enrichment of the experimental data.

References

M. Agosti and N. Ferro. 2009. Towards an Evaluation Infrastructure for DL Performance Evaluation. In G. Tsakonas and C. Papatheodorou, editors, *Evaluation of Digital Libraries: An insight into useful applications and methods*, pages 93–120. Chandos Publishing, Oxford, UK.

M. Agosti, E. Di Buccio, N. Ferro, I. Masiero, S. Peruzzo, and G. Silvello. 2012. DIRECTions: Design and Specification of an IR Evaluation Infrastructure. In T. Catarci, P. Forner, D. Hiemstra, A. Peñas, and G. Santucci, editors, *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics. Proceedings of the Third International Conference of the CLEF Initiative (CLEF 2012)*, pages 88–99. Lecture Notes in Computer Science (LNCS) 7488, Springer, Heidelberg, Germany.

Georgeta Bordea, Sabrina Kirrane, Paul Buitelaar, and Bianca O Pereira. 2012. Expertise Mining for Enterprise Content Management. In N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proc. of the Eighth Int. Conference on Language Resources and Evaluation (LREC-2012)*, pages 3495–3498. European Language Resources Association (ELRA).

G. Bordea, T. Polajnar, and P. Buitelaar. 2013. Domain-Independent Term Extraction Through Domain Modelling. In *10th International Conference on Terminology and Artificial Intelligence*.

C. L. Borgman. 2015. *Big Data, Little Data, No Data*. MIT Press.

E. Di Buccio, G. M. Di Nunzio, and G. Silvello. 2013. A Curated and Evolving Linguistic Linked Dataset. *Semantic Web*, 4(3):265–270.

N. Ferro, N. Fuhr, K. Järvelin, N. Kando, M. Lippold, and J. Zobel. 2016. Increasing Reproducibility in IR: Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science". *SIGIR Forum*, 50(1):68–82, June.

D. K. Harman. 2011. *Information Retrieval Evaluation*. Morgan & Claypool Publishers, USA.

S. Hennicke, M. Olensky, V. de Boer, A. Isaac, and J. Wielemaker. 2011. Conversion of EAD into EDM Linked Data. In L. Prediu, S. Hennicke, A. Nürnberger, A. Mitschick, and S. Ross, editors, *Proc. 1st International Workshop on Semantic Digital Archives (SDA 2011)* `http://ceur-ws.org/Vol-801/`, pages 82–88.

Clare J. Hooper, Nicolas Marie, and Evangelos Kalampokis. 2012. Dissecting the butterfly: representation of disciplines publishing at the web science conference series. In Noshir S. Contractor, Brian Uzzi, Michael W. Macy, and Wolfgang Nejdl, editors, *WebSci*, pages 137–140. ACM.

Axel-Cyrille Ngonga Ngomo. 2012. On link discovery using a hybrid approach. *Journal on Data Semantics*, 1(4):203–217.

G. Silvello and N. Ferro. 2016. "Data Citation is Coming". Introduction to the Special Issue on Data Citation. *Bulletin of IEEE Technical Committee on Digital Libraries (IEEE-TCDL)*, 12(1):1–5, May.

G. Silvello, G. Bordea, N. Ferro, P. Buitelaar, and T. Bogers. 2016. Semantic Representation and Enrichment of Information Retrieval Experimental Data. *International Journal on Digital Libraries (IJDL)*.

James W. Tanaka and Marjorie Taylor. 1991. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23(3):457–482, July.

Nyström Methods for Efficient Kernel-Based Methods for Community Question Answering

Danilo Croce[1], **Simone Filice**[2] and **Roberto Basili**[1]
[1] Dept. of Enterprise Engineering
[2] Dept. of Civil Engineering and Computer Science Engineering
University of Roma, Tor Vergata, Italy
{croce,filice,basili}@info.uniroma2.it

Abstract

English. Expressive but complex kernel functions, such as Sequence or Tree kernels, are usually underemployed in NLP tasks, e.g., in community Question Answering (cQA), as for their significant complexity in both learning and classification stages. Recently, the Nyström methodology for data embedding has been proposed as a viable solution to scalability problems. By mapping data into low-dimensional approximations of kernel spaces, it positively increases scalability through compact linear representations for highly structured data. In this paper, we show that Nyström methodology can be effectively used to apply a kernel-based method in the cQA task, achieving state-of-the-art results by reducing the computational cost of orders of magnitude.

Italiano. *Metodi di apprendimento automatico basato su funzioni kernel complesse, come Sequence o Tree Kernel, rischiano di non poter essere adeguatamente utilizzati in problemi legati all'elaborazione del linguaggio naturale (come ad esempio in Community Question Answering) a causa degli alti costi computazionali per l'addestramento e la classificazione. Recentemente é stata proposta una metodologia, basata sul metodo di Nyström, per poter far fronte a questi problemi di scalabilitá: essa permette di proiettare gli esempi, osservabili in fase di addestramento e classificazione, all'interno di spazi a bassa dimensionalitá che approssimano lo spazio sottostante la funzione kernel. Queste rappresentazioni compatte permettono di applicare algoritmi di apprendimento automatico estremamente efficienti e scalabili. In questo lavoro si dimostra che é possibile applicare metodi kernel al problema di Community Question Answering, ottenendo risultati che sono lo stato dell'arte, riducendo di ordini di grandezza i costi computazionali.*

1 Introduction

Kernel methods (Shawe-Taylor and Cristianini, 2004) have been employed in several Machine Learning algorithms (Crammer et al., 2006; Vapnik, 1998) achieving state-of-the-art performances in many classification tasks. Recently, the kernel based approach presented in (Filice et al., 2016) has been applied in the community Question Answering (cQA) challenge at SemEval 2016 (Nakov et al., 2016) obtaining state-of-the-art results.

Unfortunately, when large data volumes are involved, time and space complexity required in learning and classification may prevent the adoption of expressive but complex kernel functions, such as Sequence (Cancedda et al., 2003) or Tree kernels (Collins and Duffy, 2001). In particular, the classification cost required by a kernel-based model crucially depends on its number of support vectors: classifying a new instance requires a kernel computation against all support vectors. This scalability issue is evident in many NLP and IR applications, such as in re-ranking answers in question answering (Moschitti et al., 2007; Severyn et al., 2013; Filice et al., 2016), where the number of support vectors is typically very large.

Some approaches have been defined to bound the complexity of kernel-based methods, such as (Wang and Vucetic, 2010; Vedaldi and Zisserman, 2012; Filice et al., 2014), but they are still specific to kernel formulations and learning algorithms.

In (Croce and Basili, 2016) it has been shown that a viable and more general solution to the

above scalability issues is the Nyström methodology, a dimensionality reduction technique that has been applied also in kernel-based methods since (Williams and Seeger, 2001). This methodology has been designed to approximate the Gram Matrix derivable by a kernel function, enabling the projections of examples into low-dimensional spaces. The Nyström projection function is generated by using some examples called *landmarks*, whose number directly impacts on the embeddings quality; dually, costs of projecting a new example in the embedding space rise linearly with the number of landmarks, that is usually of orders of magnitude lower with respect of the number of possible support vectors that can be derived from a learning process. Once each example is projected in the dense low-dimensional space, the application of efficient linear learning methods is enabled, such as (Hsieh et al., 2008), preserving at the same time the expressiveness and effectiveness of kernel methods. This approach is highly applicable to different input data as well as to different kernels or learning algorithms, as discussed in (Croce and Basili, 2016).

In this paper we show that the Nyström method can be effectively used in the cQA task, by adopting the same kernel functions proposed in (Filice et al., 2016) and obtaining the same results w.r.t. the metrics adopted in the SemEval task, by reducing the computational cost of orders of magnitude.

In Section 2, we demonstrate the viability of the Nyström method to reduce the computational costs of kernel machines. Experimental results (obtained by adopting efficient SVM learning over the cQA task) are discussed in Section 3. Finally, Section 4 describes related work, while in Section 5 conclusions are derived.

2 Linearizing linguistic properties through Nyström Approach

Given an input training dataset $o_i \in \mathcal{D}$, a kernel function $K(o_i, o_j)$ is a similarity function that corresponds to a dot product in the implicit kernel space, i.e., $K(o_i, o_j) = \Phi(o_i) \cdot \Phi(o_j)$. The advantage of kernels is that the projection function $\Phi(o_i) = \vec{x}_i \in \mathbb{R}^n$ is never explicitly computed (Shawe-Taylor and Cristianini, 2004). In fact, this operation may be prohibitive when the dimensionality n of the underlying kernel space is extremely large. For example, Tree Kernels (Collins and Duffy, 2001) give rise to spaces whose number of dimensions is proportional to the number of possible sub-trees in a Natural Language. Kernel functions are exploited by kernel-based learning algorithms, such as SVM (Vapnik, 1998), to operate on the implicit kernel space without its explicit definition.

Let us assume that, given a kernel K, its explicit projection function ϕ over $\mathcal{D}$ is available to derive new representations $\vec{x}_i$ being the rows of the resulting matrix X. We define the Gram Matrix as $G = XX^\top$, with each single element corresponding to $G_{ij} = \Phi(o_i)\Phi(o_j) = K(o_i, o_j)$. The aim of the Nyström method is to derive a new low-dimensional embedding in a l-dimensional space, with $l \ll n$ so that $G \approx \tilde{G} = \tilde{X}\tilde{X}^\top$. This is obtained by generating an approximation of G using a subset of l columns of the matrix. This corresponds to selecting a subset L of the available examples, called *landmarks*. Suppose we randomly sample l columns of G, and let C be the $n \times l$ matrix of these sampled columns. Then, we can rearrange the columns and rows of G and define $X = [X_1 \;\; X_2]$ such that:

$$G = XX^\top = \begin{bmatrix} W & X_1^\top X_2 \\ X_2^\top X_1 & X_2^\top X_2 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} W \\ X_2^\top X_1 \end{bmatrix} \tag{1}$$

where $W = X_1^\top X_1$, i.e., the subset of G that only considers landmarks. The Nyström approximation can be defined as:

$$G \approx \tilde{G} = CW^\dagger C^\top \tag{2}$$

where $W^\dagger$ denotes the Moore-Penrose inverse of W. The Singular Value Decomposition (SVD) is used to obtain $W^\dagger$ as it follows. First W is decomposed so that $W = USV^\top$ where U and V are both orthogonal matrices, and S is a diagonal matrix containing the (non-zero) singular values of W on its diagonal. Since W is symmetric and positive definite $W = USU^\top$. Then $W^\dagger = US^{-1}U^\top = US^{-\frac{1}{2}}S^{-\frac{1}{2}}U^\top$ and the Equation 2 can be rewritten as

$$\begin{aligned} G \approx \tilde{G} &= CUS^{-\frac{1}{2}}S^{-\frac{1}{2}}U^\top C^\top \\ &= (CUS^{-\frac{1}{2}})(CUS^{-\frac{1}{2}})^\top = \tilde{X}\tilde{X}^\top \end{aligned} \tag{3}$$

Given an input example $o_i \in \mathcal{D}$, a new low-dimensional representation $\tilde{\vec{x}}_i$ can be thus determined by considering the corresponding i-th item of C as

$$\tilde{\vec{x}}_i = \Theta(o_i) = \vec{c}_i U S^{-\frac{1}{2}} \tag{4}$$

where $\vec{c_i}$ corresponds to a vector whose dimensions contain the evaluation of the kernel function between o_i and each landmark $o_j \in L$. The method produces l-dimensional vectors, and no restriction is applied to the input dataset as long as a valid $K(o_i, o_j)$ is used.

Several policies have been defined to determine the best selection of landmarks to reduce the Gram Matrix approximation error. In this work the uniform sampling without replacement is adopted, as suggested by (Kumar et al., 2012), where this policy has been theoretically and empirically shown to achieve results comparable with other (more complex) selection policies.

Assuming that k is the computational cost[1] of a single kernel operation, the runtime cost of the Nyström method is $\mathcal{O}(knl + l^3 + nl^2)$ as it depends on (*i*) the computation of the $n \times l$ matrix C, i.e., $\mathcal{O}(knl)$; (*ii*) the SVD evaluation on W, which is $\mathcal{O}(l^3)$; and (*iii*) the projection of the entire dataset through the multiplication by C, i.e., $\mathcal{O}(nl^2)$. For several classes of kernels, such as Tree or Sequence Kernels (Collins and Duffy, 2001), the kernel computation cost is extremely high. Therefore, the computational cost for the construction of the matrix C dominates the overall expression.

Once an example is projected in the l-dimensional space, efficient and large-scale learning algorithm can be applied. To further control the computational cost of the training step, we addressed a class of algorithms that bounds the number of times a single instance is re-used during training. In particular, we investigated the Dual Coordinate Descent algorithm (Hsieh et al., 2008): it is a batch learning algorithm whose achievable accuracy is made inversely dependent on the number of iterations T over a training dataset. Its training time cost on a dataset of n examples in $\mathbb{R}^l$ is $\mathcal{O}(Tnl)$. Being fixed the number of iterations required to obtain an accurate model[2], such cost is negligible w.r.t. the projection cost. Therefore, a complete training process exploiting the Nyström method is simply $\mathcal{O}\big(kln\big)$, that should be compared with a traditional kernel-based SVM learning algorithm, e.g., (Chang and Lin, 2011), whose computational cost is almost $\mathcal{O}\big(kn^2\big)$, with $l \ll n$.

The computational cost of a classification step only depends on the projection of the example in the new space, i.e., $\mathcal{O}\big(kl\big)$. In fact, once a test example is projected, the final decision requires a dot product between the low-dimensional representation $\tilde{\vec{x}_i}$ and the hyperplane underlying the classification function: again, this is negligible with respect to the cost of the single kernel operations. Such cost is typically extremely lower than the cost of a pure kernel-based classification, which requires a kernel operation againts all the support vectors selected during the training process, which is usually far larger than the number of landmarks.

[1] Expressed in terms of basic operations, such as products.

[2] In (Croce and Basili, 2016) a number of iterations $T = 30$ obtained stable and accurate results in several tasks.

3 Empirical Investigation: the Community QA task

The proposed stratified Nyström method has been applied in the SemEval-2016 community Question Answering (cQA) task. In this task, participants are asked to automatically provide good answers in a community question answering setting (Nakov et al., 2016).

In particular, we focused on the Subtask A: given a question and a large collection of question-comment threads created by a user community, the task consists in (re-)ranking comments that are most useful for answering the question. This task is interesting as kernel methods achieved the highest results in the cQA task, as demonstrated by the KeLP team (Filice et al., 2016). In particular, Subtask A is modeled as a binary classification problem, where examples are generated by considering (question,comment) pairs. Each pair generates an example for a binary SVM, where the positive label is associated with a *good* comment and the negative label includes the *potential* and *bad* comments. The classification score is used to sort the instances and produce the final ranking. According to the above setting, a train and test dataset made of 20,340 and 3,270 examples are generated. In (Filice et al., 2016), a Kernel-based SVM classifier achieved state-of-the-art results by adopting a kernel combination that exploited (*i*) feature vectors containing linguistic similarities between the texts in a pair; (*ii*) shallow syntactic trees that encode the lexical and morpho-syntactic information shared between text pairs; (*iii*) feature vectors capturing task-specific information.

First, a batch kernel-based SVM (Chang and Lin, 2011) learning algorithm operating on the kernel function proposed in (Filice et al., 2016) is adopted to determine the upper bound in terms

of classification quality (but with higher computational costs). Then, multiple standard Nyström methods are used to linearize the dataset by sampling different numbers of landmarks: 10 configurations have been investigated by starting from 100 landmarks and incrementally adding 100 landmarks at a time. The higher is the number of used landmarks, the higher is the quality of the approximated low-dimensional space (Drineas and Mahoney, 2005), but the higher is also the computational cost. The most complex projection function is thus based on 1,000 landmarks. Landmarks have been selected by applying a random selection without replacement, as suggested in (Kumar et al., 2012). An efficient linear SVM (Hsieh et al., 2008) is adopted on the resulting embedding space. Experiments have been carried out by using the KeLP framework[3] (Filice et al., 2015a).

Results are reported in Table 1 in terms of Mean Average Precision (MAP, the official rank of the competition), F_1 on the *good* class, and computational saving, i.e., percentage of avoided kernel operations in classification. The standard SVM model contains 11,322 Support Vectors, thus requiring more than 37M kernel operations for the complete classification of the 3,270 test instances[4]. By adopting the Nyström methodology with only 1,000 landmarks the same F_1 score (i.e., 64.4) is obtained. Moreover, a comparable MAP (i.e., 78.2%) achieved by the KeLP team is replicated with a 91.2% of saving. The speed up is impressive also when fewer landmarks are used: with 300 landmarks, 77.7 MAP is obtained by saving more that 97% of kernel computations. These results are straightforward, considering that results comparable with the state-of-the-art can be obtained by reducing of almost two orders of magnitude the computational costs. Overall, the MAP obtained by the proposed approach is still higher than the one achieved by all the other systems of the challenge, including ConvKN (Barrón-Cedeño et al., 2016) and SemanticZ (Mihaylov and Nakov, 2016), i.e., the second and third best systems, respectively.

Table 1: Results in CQA. Upperbound is achieved by a SVM with more than 37M kernel operations.

Landmarks	MAP	F1	Saving
100	76.0	58.6	99.1%
200	77.0	60.8	98.2%
300	77.5	62.2	97.4%
400	77.7	62.4	96.5%
500	77.9	63.1	95.6%
600	78.0	63.6	94.7%
700	78.1	63.7	93.8%
800	78.0	63.8	92.9%
900	78.1	64.2	92.1%
1000	78.2	64.4	91.2%
standard SVM	79.2	64.4	-
ConvKN	77.7	66.2	-
SemanticZ	77.6	61.8	-

4 Related Work

Improving the efficiency of kernel-based methods is a largely studied topic. The reduction of computational costs has been early designed by imposing a budget in the number of support vectors (Cesa-Bianchi and Gentile, 2006; Dekel and Singer, 2006; Orabona et al., 2008; Wang and Vucetic, 2010; Filice et al., 2014). However, in complicated tasks, such methods still require large budgets that systematically rely on many kernel computations. They are thus less efficient than Nyström: a classifier based on the Nyström method with l landmarks has approximately the same computational complexity of its budgeted counterpart with a budget set to l, but its accuracy is typically higher, as shown in (Croce and Basili, 2016). Alternatively, Zanzotto and Dell'Arciprete (2012) proposed Distributed Tree Kernels that approximate tree kernels (Collins and Duffy, 2001) through the explicit mapping of trees into vectors. DTKs focus on specific tree kernel functions, while the approach proposed here can be effectively applied to any kernel function. An alternative strategy is presented in (Filice et al., 2015b), where a cascade of kernel-based classifiers is proposed according to the computational cost of their kernel functions, so that more complex classifiers are invoked only on difficult instances. Their solution is strictly connected to the availability of multiple kernels that have to be sorted according to their complexity and expressiveness. Usually, it is hard to define many kernels for a given task, and consequently only few layers can be set.

5 Conclusion

This paper discussed the application of Nyström method for a significant reduction of computa-

[3] https://github.com/SAG-KeLP

[4] The classification time was more than 3 hours and a half on a standard machine with 4 cores i7-2600 3.4 GHz.

tional costs in kernel-based classifications in the cQA task. By projecting examples into low-dimensional embeddings, Nyström enables the adoption of efficient linear classifier, and drastically reduces the overall computational cost. Experimental results demonstrate that the proposed approach leads to a cost reduction higher than 90%, with a negligible performance drop. Future research will be devoted to the definition of a principled strategy to estimate the optimal number of layers, as well as the size of embeddings at each layer.

References

Alberto Barrón-Cedeño, Giovanni Da San Martino, Shafiq Joty, Alessandro Moschitti, Fahad Al-Obaidli, Salvatore Romeo, Kateryna Tymoshenko, and Antonio Uva. 2016. Convkn at semeval-2016 task 3: Answer and question selection for question answering on arabic and english fora. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 896–903, San Diego, California, June. Association for Computational Linguistics.

Nicola Cancedda, Éric Gaussier, Cyril Goutte, and Jean-Michel Renders. 2003. Word-sequence kernels. *Journal of Machine Learning Research*, 3:1059–1082.

Nicolo' Cesa-Bianchi and Claudio Gentile. 2006. Tracking the best hyperplane with a simple budget perceptron. In *In proc. of the nineteenth annual conference on Computational Learning Theory*, pages 483–498. Springer-Verlag.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May.

Michael Collins and Nigel Duffy. 2001. Convolution kernels for natural language. In *Proceedings of Neural Information Processing Systems (NIPS'2001)*, pages 625–632.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, December.

Danilo Croce and Roberto Basili. 2016. Large-scale kernel-based language learning through the ensemble nystrom methods. In *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, pages 100–112.

Ofer Dekel and Yoram Singer. 2006. Support vector machines on a budget. In Bernhard Schlkopf, John Platt, and Thomas Hoffman, editors, *NIPS*, pages 345–352. MIT Press.

Petros Drineas and Michael W. Mahoney. 2005. On the nystrm method for approximating a gram matrix for improved kernel-based learning. *Journal of ML Research*, 6.

Simone Filice, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2014. Effective kernelized online learning in language processing tasks. In *Proceedings of ECIR 2014*, pages 347–358.

Simone Filice, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2015a. Kelp: a kernel-based learning platform for natural language processing. In *Proceedings of ACL: System Demonstrations*, Beijing, China, July.

Simone Filice, Danilo Croce, and Roberto Basili. 2015b. A Stratified Strategy for Efficient Kernel-based Learning. In *AAAI Conference on Artificial Intelligence*.

Simone Filice, Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2016. Kelp at semeval-2016 task 3: Learning semantic relations between questions and answers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1116–1123, San Diego, California, June. Association for Computational Linguistics.

Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathiya Keerthi, and S. Sundararajan. 2008. A dual coordinate descent method for large-scale linear svm. In *Proceedings of the ICML 2008*, pages 408–415, New York, NY, USA. ACM.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. 2012. Sampling methods for the nyström method. *J. Mach. Learn. Res.*, 13:981–1006, April.

Todor Mihaylov and Preslav Nakov. 2016. Semanticz at semeval-2016 task 3: Ranking relevant answers in community question answering using semantic similarity based on fine-tuned word embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 879–886, San Diego, California, June. Association for Computational Linguistics.

Alessandro Moschitti, Silvia Quarteroni, Roberto Basili, and Suresh Manandhar. 2007. Exploiting syntactic and shallow semantic kernels for question/answer classification. In *Proceedings of ACL'07*.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, June. Association for Computational Linguistics.

Francesco Orabona, Joseph Keshet, and Barbara Caputo. 2008. The projectron: a bounded kernel-based perceptron. In *Proceedings of ICML '08*, pages 720–727, USA. ACM.

Aliaksei Severyn, Massimo Nicosia, and Alessandro Moschitti. 2013. Building structures from classifiers for passage reranking. In *Proceedings of the 22nd ACM international Conference on Information and Knowledge Management*, CIKM '13, pages 969–978, New York, NY, USA. ACM.

John Shawe-Taylor and Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA.

Vladimir N. Vapnik. 1998. *Statistical Learning Theory*. Wiley-Interscience.

Andrea Vedaldi and Andrew Zisserman. 2012. Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3).

Zhuang Wang and Slobodan Vucetic. 2010. Online passive-aggressive algorithms on a budget. *Journal of Machine Learning Research - Proceedings Track*, 9:908–915.

Christopher K. I. Williams and Matthias Seeger. 2001. Using the nyström method to speed up kernel machines. In *Proceedings of NIPS 2000*.

Fabio Massimo Zanzotto and Lorenzo Dell'Arciprete. 2012. Distributed tree kernels. In *Proceedings of ICML 2012*.

Topic Modelling with Word Embeddings

Fabrizio Esposito
Dept. of Humanities
Univ. of Napoli Federico II
fabrizio.esposito3
@unina.it

Anna Corazza, Francesco Cutugno
DIETI
Univ. of Napoli Federico II
anna.corazza|francesco.cutugno
@unina.it

Abstract

English. This work aims at evaluating and comparing two different frameworks for the unsupervised topic modelling of the CompWHoB Corpus, namely our political-linguistic dataset. The first approach is represented by the application of the latent DirichLet Allocation (henceforth LDA), defining the evaluation of this model as baseline of comparison. The second framework employs Word2Vec technique to learn the word vector representations to be later used to topic-model our data. Compared to the previously defined LDA baseline, results show that the use of Word2Vec word embeddings significantly improves topic modelling performance but only when an accurate and task-oriented linguistic pre-processing step is carried out.

Italiano. *L'obiettivo di questo contributo è di valutare e confrontare due differenti framework per l'apprendimento automatico del topic sul CompWHoB Corpus, la nostra risorsa testuale. Dopo aver implementato il modello della latent DirichLet Allocation, abbiamo definito come standard di riferimento la valutazione di questo stesso approccio. Come secondo framework, abbiamo utilizzato il modello Word2Vec per apprendere le rappresentazioni vettoriali dei termini successivamente impiegati come input per la fase di apprendimento automatico del topic. I risulati mostrano che utilizzando i 'word embeddings' generati da Word2Vec, le prestazioni del modello aumentano significativamente ma solo se supportati da una accurata fase di 'pre-processing' linguistico.*

1 Introduction

Over recent years, the development of political corpora (Guerini et al., 2013; Osenova and Simov, 2012) has represented one of the major trends in the fields of corpus and computational linguistics. Being carriers of specific content features, these textual resources have met the interest of researchers and practitioners in the study of topic detection. Unfortunately, not only has this task turned out to be hard and challenging even for human evaluators but it must be borne in mind that manual annotation often comes with a price. Hence, the aid provided by unsupervised machine learning techniques proves to be fundamental in addressing the topic detection issue.

Topic models are a family of algorithms that allow to analyse unlabelled large collections of documents in order to discover and identify hidden topic patterns in the form of cluster of words. While LDA (Blei et al., 2003) has become the most influential topic model (Hall et al., 2008), different extensions have been proposed so far: Rosen-Zvi et al. (Rosen-Zvi et al., 2004) developed an author-topic generative model to include also authorship information; Chang et al. (Chang et al., 2009a) presented a probabilist topic model to infer descriptions of entities from corpora identifying also the relationships between them; Yi Yang et al. (Yang et al., 2015) proposed a factor graph framework for incorporating prior knowledge into LDA.

In the present paper we aim at topic modelling the CompWHoB Corpus (Esposito et al., 2015), a political corpus collecting the transcripts of the White House Press Briefings. The main characteristic of our dataset is represented by its dialogical structure: since the briefing consists of a question-answer sequence between the US press secretary and the news media, the topic under discussion may change from one answer to the fol-

lowing question, and vice versa. Our purpose was to address this main feature of the CompWHoB Corpus associating at each answer/question only one topic. In order to reach our goal, we propose an evaluative comparison of two different frameworks: in the first one, we employed the LDA approach by extracting from each answer/question document only the topic with the highest probability; in the second framework, we applied the word embeddings generated from the Word2Vec model (Mikolov and Dean, 2013) to our data in order to test how dense high-quality vectors represent our data, finally comparing this approach with the previously defined LDA baseline. The evaluation was performed using a set of gold-standard annotations developed by human experts in political science and linguistics. In Section 2 we present the dataset used in this work. In Section 3, the linguistic pre-processing is detailed. Section 4 shows the methodology employed to topic-model our data. In Section 5 we present the results of our work.

2 The dataset

2.1 The CompWHoB Corpus

The textual resource used in the present contribution is the CompWHoB (Computational White House press Briefings) Corpus, a political corpus collecting the transcripts of the White House Press Briefings extracted from the American Presidency Project website, annotated and formatted into XML encoding according to TEI Guidelines (Consortium et al., 2008). The CompWHoB Corpus spans from January 27, 1993 to December 18, 2014. Each briefing is characterised by a turn-taking between the podium and the journalists, signalled in the XML files by the use of a *u* tag for each utterance. At the time of writing, 5,239 briefings have been collected, comprising 25,251,572 tokens and a total number of 512,651 utterances (from now on, utterances will be referred to as 'documents'). The document average length has been measured to 49.25 tokens, while its length variability is comprised within a range of a minimum of 0 and a maximum of 4724 tokens. The dataset used in the present contribution was built and divided into training and test set by randomly selecting documents from the CompWHoB Corpus in order to vary as much as possible the topics dealt with by US administration.

2.2 Gold-Standard Annotation

Two hundred documents of the test set were manually annotated by scholars with expertise in linguistics and political science using a set of thirteen categories. Seven macro-categories were created taking into account the US major federal executive departments so as not to excessively narrow the topic representation, accounting for 28.5% of the labelled documents. Six more categories were designed in order to take into account the informal nature of the press briefings that makes them an atypical political-media genre (Venuti and Spinzi, 2013), accounting for the remaining 71.5% (Table 1). The labelled documents represent the gold-standard to be used in the evaluation stage. This choice is motivated by the fact that even if metrics such as perplexity or held-out likelihood prove to be useful in the evaluation of topic models, they often fail in qualitatively measuring the coherence of the generated topics (Chang et al., 2009b). Thus, more formally our gold-standard can be defined as the set $G = \{g_1, g_2, ..., g_S\}$ where g_i is the ith category in a range $\{1, S\}$ with $S = 13$ as the total number of categories.

Crime and justice	Culture and Education
Economy and welfare	Foreign Affairs
Greetings	Health
Internal Politics	Legislation & Reforms
Military & Defense	President Updates
Presidential News	Press issues
Unknown topic	

Table 1: Gold-Standard Topics

3 Linguistic Pre-Processing

In order to improve the quality of our textual data, special attention was paid to the linguistic pre-processing step. In particular, since LDA represents documents as mixtures of topics in forms of words probability, we wanted these topics to make sense also to human judges. Being press briefings actual conversations where the talk moves from one social register to another (e.g. switch from the reading of an official statement to an informal interaction between the podium and the journalists) (Partington, 2003), the first step was to design an *ad-hoc* stoplist able to take into account the main features of this linguistic genre. Indeed, not only were words with a low frequency discarded,

but also high frequency ones were removed in order not to overpower the rest of the documents. More importantly, we included in our stoplist all the personal and indefinite pronouns as well as the most commonly used honorifics (e.g. Mr., Ms., etc.), given their predominant role in addressing the speakers in both informal and formal settings (e.g. "Mr. Secretary, you said oil production is up, [...]"). Moreover, the list of the first names of the press secretaries in office during the years covered by the CompWHoB Corpus was extracted from Wikipedia and added to the stoplist, since most of the time used only as nouns of address (Brown et al., 1960). As regards the proper NLP pipeline implemented in this work, the Natural Language ToolKit[1] (NLTK) platform (Bird et al., 2009) was employed: word tokenization, POS-tagging, using the Penn Treebank tag set (Marcus et al., 1993) and lemmatization were carried out to refine our data. When pre-processing is not applied to the dataset, only punctuation is removed from the documents.

4 Methodology

This section deals with the two techniques employed in this work to topic-model our data. We first discuss the LDA approach and then focus on the use of the word embeddings learnt employing Word2Vec model. Both the techniques were implemented in Python (version 3.4) using the Gensim[2] library (Rehurek and Sojka, 2010).

4.1 Latent DirichLet Allocation

In our first experiment we ran LDA, a generative probabilistic model that allows to infer latent topics in a collection of documents. In this unsupervised machine learning technique the topic structure represents the underlying *hidden* variable (Blei, 2012) to be discovered given the *observed* variables, i.e. documents' items from a fixed vocabulary, be them textual or not. More formally, LDA describes each document d as multinomial distribution θ_d over topics, while each topic t is defined as a multinomial distribution ϕ_t over words in a fixed vocabulary where $i_{d,n}$ is the nth item in the document d.

4.1.1 Topic modelling with LDA

Data were linguistically pre-processed prior to training LDA model and only words pos-tagged as nouns ('NN') were kept in both the training and test sets' documents. This choice was motivated by the necessity of generating topics that could be semantically meaningful. After having carried out the pre-processing step, we trained LDA model on our training corpus by employing the online variational Bayes (VB) algorithm (Hoffman et al., 2010) provided by the Gensim library. Based on online stochastic optimization with a natural gradient step, LDA online proves to converge to a local optimum of the VB objective function. It can be applied to large streaming document collections being able to make better predictions and find better topic models with respect to those found with batch VB. As parameters of our model, we set the k number of topics to thirteen as the numbers of classes in our gold-standard, updating the model every 150 documents and giving two passes over the corpus in order to generate accurate data. Once the model was trained, we inferred topic distributions on the unseen documents of the test set. For each document d_i, the topic $t_{max(i)}$ with the highest probability in the multinomial distribution was selected and associated to it. The cluster ω_k corresponds then to the set of documents associated to the topic t_k. Due to the presence of a gold-standard, the *external criterion* of *purity* was chosen as evaluation measure of this approach. *Purity* is formally defined as:

$$\textit{purity}(\Omega, G) = \frac{1}{N}\sum_{k}\max_{j}|w_k \cup g_j|$$

$\Omega = \{\omega_1, \omega_2, ..., \omega_K\}$ is the set of clusters and $G = \{g_1, g_2, ..., g_S\}$ is the set of gold-standard classes. The *purity* computed for the LDA approach is:

$$purity \approx 0.46$$

This measure constituted the baseline of comparison with the Word2Vec word embeddings approach.

4.2 Word2Vec

Word2Vec (Mikolov et al., 2013a) is probably the most popular software providing learning models for the generation of dense embeddings. Based on Zelig Harris' *Ditributional Hypothesis* (Harris, 1954) stating that words occurring in similar contexts tend to have similar meanings, Word2Vec model allows to learn vector representations of words referred to as *word embeddings*. Differently from techniques such as LSA (Dumais, 2004),

[1] http://www.nltk.org

[2] https://radimrehurek.com/gensim/

LDA and other topic models that use *documents* as context, Word2Vec learns the distributed representation for each target word by defining the context as the terms surrounding it. The main advantage of this model is that each dimension of the embedding represents a latent feature of the word (Turian et al., 2010), encoding in each word vector essential syntactic and semantic properties (Mikolov et al., 2013c). In this way, simple vector similarity operations can be computed using *cosine similarity*. Moreover, it must not be forgotten that one of Word2Vec's secrets lies in its efficient implementation that allows a very robust and fast training.

4.2.1 Topic modelling with Word2Vec

Training data were linguistically pre-processed beforehand according to the *ad-hoc* pipeline implemented in this work. The model was initialised setting a minimum count for the input words: terms whose frequency was lower than 20 were discarded. In addition, we set the default threshold at $1 \exp -3$ for configuring the high-frequency words to be randomly downsampled in order to improve word embeddings quality (Mikolov and Dean, 2013). Moreover, as highlighted by Goldberg and Levy (Goldberg and Levy, 2014), both sub-sampling and rare-pruning seem to increase the effective size of the window making the similarities more topical. Finally, based on the recommendation of Mikolov et al. (Mikolov et al., 2013b) and Baroni et al. (Baroni et al., 2014), in this work we trained our model using the CBOW algorithm since more suitable for larger datasets. The dimensionality of our feature vectors was fixed at 200. Once constructed the vocabulary and trained the input data, we used the learnt word vector representations on our unseen test set documents. Then, we calculated the centroid c for each document d, where $e_{d,i}$ is the ith embedding in d, so as to obtain a meaningful topic representation for each document (Mikolov and Dean, 2013). Finally, we clustered our data using the k-means algorithm. In order to compare our approach with the baseline previously defined, the external criterion of *purity* was computed also in this experiment to evaluate how well the k-means clustering matched the gold-standard classes:

$$purity \approx 0.54$$

This technique proved to outperform the LDA topic model approach presented in this work. Surprisingly, notwithstanding the fact that Word2Vec relies on a broad context to produce high-quality embeddings, this framework showed to perform better using a linguistically pre-processed dataset where only nouns are kept. Table 2 shows the results obtained in the two experiments.

Topic Models Results	
Framework	Results
LDA without pre-processing	**0.45**
LDA with pre-processing	**0.46**
Word2Vec without pre-processing	**0.44**
Word2Vec with pre-processing	**0.54**

Table 2: Results of the two frameworks. When pre-preprocessing is not applied, only punctuation is removed.

5 Conclusions

In this contribution we have presented a comparative evaluation of two unsupervised learning approaches to topic modelling. Two experiments were carried out: in the first one, we applied a classical LDA model to our dataset; in the second one, we trained our model using Word2Vec so as to generate the word embeddings for topic-modelling our test set. After clustering the output of the two approaches, we evaluated them using the external criterion of purity. Results show that the use of word embeddings outperforms the LDA approach but only if a linguistic task-oriented pre-processing stage is carried out. As at the moment no comprehensive explanation can be provided, we can only suggest that the main reason for these results may lie in the fluctuating length of each document in our dataset. In fact, we hypothesise that the use of word embeddings may prove to be the boosting factor of Word2Vec topic model since encoding information about the close context of the target term. As part of future work, we aim to further investigate this aspect and design a topic model framework that could take into account the main structural and linguistic features of the CompWHoB Corpus.

Acknowledgments

The authors would like to thank Antonio Origlia for the useful and thoughtful discussions and insights.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

David M Blei. 2012. Probabilistic Topic Models. *Communications of the ACM*, 55(4):77–84.

Roger Brown, Albert Gilman, et al. 1960. The Pronouns of Power and Solidarity. *Style in language*, pages 253–276.

Jonathan Chang, Jordan Boyd-Graber, and David M. Blei. 2009a. Connections between the Lines: Augmenting Social Networks with Text. In *Knowledge Discovery and Data Mining*.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009b. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in neural information processing systems*, pages 288–296.

TEI Consortium, Lou Burnard, Syd Bauman, et al. 2008. *TEI P5: Guidelines for electronic text encoding and interchange*. TEI Consortium.

Susan T Dumais. 2004. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230.

Fabrizio Esposito, Pierpaolo Basile, Francesco Cutugno, and Marco Venuti. 2015. The CompWHoB Corpus: Computational Construction, Annotation and Linguistic Analysis of the White House Press Briefings Corpus. *CLiC it*, page 120.

Yoav Goldberg and Omer Levy. 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Marco Guerini, Danilo Giampiccolo, Giovanni Moretti, Rachele Sprugnoli, and Carlo Strapparava. 2013. The New Release of CORPS: A Corpus of Political Speeches Annotated with Audience Reactions. In *Multimodal Communication in Political Speech. Shaping Minds and Social Action*, pages 86–98. Springer.

David Hall, Daniel Jurafsky, and Christopher D Manning. 2008. Studying the History of Ideas Using Topic Models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 363–371. Association for Computational Linguistics.

Zellig S Harris. 1954. Distributional Structure. *Word*, 10(2-3).

Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online Learning for latent Dirichlet Allocation. In *advances in neural information processing systems*, pages 856–864.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.

T Mikolov and J Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in neural information processing systems*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting Similarities among Languages for Machine Translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic Regularities in Continuous Space Word Representations. In *HLT-NAACL*, volume 13, pages 746–751.

Petya Osenova and Kiril Simov. 2012. The Political Speech Corpus of Bulgarian. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Alan Partington. 2003. *The Linguistics of Political Argument: The Spin-Doctor and the Wolf-Pack at the White House*. Routledge.

Radim Rehurek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The Author-Topic Model for Authors and Documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics.

M Venuti and C Spinzi. 2013. Tracking the change in institutional genre: a diachronic corpus-based study of White House Press Briefings. *The three waves of globalization: winds of change in Professional, Institutional and Academic Genres*.

Yi Yang, Doug Downey, Jordan Boyd-Graber, and Jordan Boyd Graber. 2015. Efficient Methods for Incorporating Knowledge into Topic Models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Imparare a quantificare guardando

Sandro Pezzelle CIMeC
Ionut Sorodoc EM LCT
Aurelie Herbelot CIMeC
Raffaella Bernardi CIMeC, DISI
Università degli Studi di Trento
{nome.cognome@unitn.it}

Abstract

English. In this paper, we focus on linguistic questions over images which may be answered with a quantifier (e.g. *How many dogs are black? Some/most/all of them*, etc.). We show that in order to learn to quantify, a multimodal model has to obtain a genuine understanding of linguistic and visual inputs and of their interaction. We propose a model that extracts a fuzzy representation of the set of the queried objects (e.g. *dogs*) and of the queried property in relation to that set (e.g. *black* with respect to *dogs*), outputting the appropriate quantifier for that relation.

Italiano. *In questo lavoro studiamo le domande del tipo "Quanti cani sono neri?", la cui risposta è un quantificatore (es. "alcuni"/"tutti"/"nessuno"). Mostriamo che al fine di imparare a quantificare, un modello multimodale deve ottenere una rappresentazione profonda della domanda linguistica, dell'immagine e della loro interazione. Proponiamo un modello che estrae una rappresentazione approssimativa dell'insieme degli oggetti e della proprietà sui quali verte la domanda.*

1 Introduzione

La linguistica computazionale ed i sistemi di visione artificiale stanno attraversando un momento particolarmente favorevole, ben rappresentato dallo sviluppo di modelli multimodali capaci di svolgere compiti che sembravano fuori portata fino a pochissimi anni fa, e che adesso si trovano integrati in molte applicazioni destinate all'utilizzo da parte degli utenti. Il compito di "Visual Question Answering" (VQA), finalizzato a rispondere a domande a partire da input visivi, e la generazione automatica di didascalie ("caption generation") sono solo alcuni dei compiti in cui maggiormente si è assistito a un rapido e inarrestabile miglioramento. Oltre a sfociare in dirette applicazioni nel mondo reale, dove i contenuti visivi giocano un ruolo cruciale, la performance di questi modelli di lingua e visione rappresenta anche un indicatore della misura in cui essi riescono a catturare il significato, fornendoci importanti intuizioni teoriche.

I recenti miglioramenti nel campo delle applicazioni VQA sono da attribuire principalmente alla nuova generazione di reti neurali artificiali di apprendimento profondo ("deep learning neural networks"), insieme alla crescente disponibilità di grandi dataset di immagini. Questi modelli hanno dimostrato che anche reti neurali (RN) molto semplici possono catturare interazioni complesse tra le proprietà di un dataset, spesso superando in performance modelli molto più complessi. Ad esempio, (Zhou et al., 2015) ha recentemente dimostrato che un semplice modello bag-of-words (BoW) può ottenere performance all'avanguardia in uno dei più importanti dataset di VQA (Antol et al., 2015). Gli autori dello studio, tuttavia, obiettano che la performance del modello è da attribuire più alla eccellente abilità della rete di memorizzare correlazioni che a una reale capacità di ragionamento e comprensione, e concludono sostenendo la necessità di passare al secondo obiettivo. Il loro studio dimostra anche quanto difficile sia valutare se un modello sia in grado di afferrare realmente il significato di immagini e parole. Nel presente lavoro proponiamo un compito che riteniamo essere utile per la comunità ai fini del raggiungimento di questo obiettivo.

Nel compito di VQA, le valutazioni dei modelli si basano su domande che riguardano le proprietà di oggetti specifici, come la posizione, il colore, ecc., mentre nel compito di generazione di dida-

scalie viene presa in considerazione l'intera immagine. In entrambi i casi, è ovviamente possibile imparare correlazioni tra le parole presenti nella domanda, le parole presenti nella risposta e l'immagine stessa (o parti di essa). Presentata a un modello RN insieme alla domanda *Cosa sta mangiando?*, ad esempio, l'immagine di una persona che mangia una torta attiverà sicuramente l'associazione tra il verbo *mangiare* e proprietà (sia linguistiche che visive) collegate al cibo, producendo la risposta corretta. Per testare un modello al di là di questo semplice meccanismo, emerge la necessità di proporre un compito in cui una particolare associazione tra domanda e immagine non dia come risultato sempre la stessa risposta.

Le domande riguardanti il "numero" sono state studiate solo marginalmente in letteratura, e la performance dei sistemi di VQA su queste domande si è dimostrata abbastanza scarsa (Ren et al., 2015; Antol et al., 2015). Inoltre, i lavori precedenti si sono concentrati quasi esclusivamente sulla modellizzazione di cardinali esatti, analizzando quindi solo parzialmente il fenomeno della quantificazione. Nel presente articolo investighiamo un nuovo tipo di domanda che richiede un certo grado di comprensione dei quantificatori generalizzati (*few, most, all*, ecc.). Il motivo per cui siamo interessati a queste domande è che, per rispondervi, non è sufficiente identificare l'area di un'immagine correlata alla domanda. La domanda *In che proporzione i cani sono neri?*, ad esempio, richiede qualcosa di più dell'identificazione delle proprietà *cane* e *nero* nell'immagine: la rete deve essere in grado di riflettere sulla relazione tra queste proprietà, e di generare uno dei quantificatori, che sono potenzialmente simili tra di loro. L'abilità di identificare i membri di un insieme e le loro proprietà condivise richiede un certo grado di ragionamento più profondo che, sosteniamo, non può essere ottenuto con un semplice meccanismo di memorizzazione.

In ciò che segue, consideriamo il compito di VQA come un problema di "fill in the blank" (riempire uno spazio vuoto con la parola corretta), e poniamo la domanda *In che proporzione i cani sono neri?* nella seguente forma: ___ *cani sono neri*. Le possibili risposte sono selezionate all'interno di un insieme di quantificatori linguistici, ovvero *no, few, some, most* e *all*. Per assegnare il quantificatore corretto, il modello deve essere in grado di porre l'attenzione sugli oggetti rilevanti (il restrittore del quantificatore) e di quantificare gli oggetti che, all'interno di questo dominio ristretto, possiedono la proprietà richiesta (la portata del quantificatore). Mostriamo che un semplice modello BoW non è sufficiente per compiere efficacemente questo compito, e proponiamo un modello RN alternativo e linguisticamente motivato, la cui performance risulta essere superiore al modello di (Zhou et al., 2015).

2 Dati

Abbiamo usato scenari contenenti ognuno sedici immagini estratte da ImageNet. ImageNet ci fornisce immagini annotate manualmente con un'etichetta identificativa dell'oggetto (e.s., *dog*, *wine*, *pizza*, ecc.) e diverse proprietà associate ad esso (e.s., *black*, *furry*, ecc.) Abbiamo selezionato tutte le immagini di ImageNet annotate con almeno una proprietà per un totale di 9600 immagini rappresentanti 203 tipi di oggetti e 24 proprietà. Abbiamo poi selezionato le immagini utili per la costruzione del nostro dataset, mantenendo solo gli oggetti che compaiono in un numero significativo di immagini e che vengono menzionati frequentemente nel corpus UkWaC.[1] Questo ci permette di ottenere rappresentazioni visive e linguistiche adeguate. Applicando questa selezione, abbiamo ottenuto 161 oggetti e 7124 immagini,[2] che sono state successivamente assemblate in scenari "plausibili". A tal fine abbiamo calcolato la probablità con cui due oggetti possano far parte di una stessa scena utilizzando la loro co-occorrenza nelle didascalie disponibili in MS-COCO (Lin et al., 2014) – ad esempio, un cane e un divano hanno più possibilità di co-occorrere in una stessa scena rispetto a un elefante e un divano. Abbiamo usato quindi la seguente misura:

$$PMI(o1, o2) = log \frac{f(o1, o2) * N}{f(o1) * f(o2)} \quad (1)$$

[1] Abbiamo scelto oggetti con almeno 16 immagini e che occorrono nel corpus almeno 150 volte.

[2] Ognuno dei 161 oggetti è rappresentato da una media di 48 immagini uniche (sd=99), con una distribuzione che va da 13 (*pasta*) a 1104 (*dog*). Ogni oggetto è associato a un numero di proprietà che varia da un minimo di 2 (es. *lion*) ad un massimo di 18 (es. *dog*) con una media pari a 8 (sd=3.4). All'interno delle 7124 immagini uniche, la coppia oggetto-proprietà più frequente è *furry dog* con 769 occorrenze, mentre le meno frequenti (es. *pink salmon*) occorrono in una sola immagine (media=13.5, sd=37). Infine, la proprietà più frequente, *furry*, compare in 2936 immagini uniche, seguita da *brown* (2782) e *smooth* (2266). La meno frequente è *violet*, che occorre in 24 immagini. La frequenza media è 801 (sd=837).

in cui $o1$ e $o2$ sono due oggetti, $f(o1, o2)$ conta quante volte le etichette di $o1$ and $o2$ appaiono nelle didascalie delle stesse immagini, $f(o)$ conta quante volte o appare nella didascalia di MS-COCO in totale, e N è il numero delle parole in tutte le didascalie. Le etichette che non sono usate nelle didascalie ricevono un valore uniformemente distribuito in relazione a tutti gli altri oggetti. 10,000 scenari sono stati generati secondo la procedura seguente:

- scegliamo un'etichetta a caso dall'insieme di 161 oggetti (e.s., *dog*);
- scegliamo una proprietà p dall'insieme delle 24 proprietà (e.s., *black*);
- selezioniamo n_1 immagini che contengono oggetti con l'etichetta l e n_2 immagini che contengono oggetti con l'etichetta l e proprietà p, cosi che $0 \leq n_1 \leq 16$ and $0 \leq n_2 \leq n_1$;
- riempiamo le rimanenti celle dello scenario $(16 - n_1)$ con le immagini non etichettate con l ed usando la distanza PMI per scegliere oggetti che possano plausibilmente co-occorrere con l'oggetto target;
- usando la proporzione tra n_1 e n_2, calcoliamo quale quantificatore assegnare allo scenario assemblato, seguendo regole pre-definite: *no* e *all* sono assegnati quando, rispettivamente, nessun n_1 o tutti gli n_1 hanno la proprietà p. Per *most* e *few* usiamo le stime riportate in (Khemlani et al., 2009), e assegniamo il primo quando la proporzione ùguale o superiore al 70%, e il secondo quando la proporzione è al piú 17%. Tutte le proporzioni che cadono tra questi due valori sono assegnate a *some*. Ad esempio, se $n_1 = 6$ oggetti con l'etichetta l, assegniamo *no* a casi in cui $n_2 = 0$, *few* quando $n_2 = 1$ (1/6=0.1667), *some* quando $2 \leq n_2 \leq 4$, *most* quando $n_2 = 5$ $(5/6 = 0.833)$, e *all* quando $n_2 = 6$;
- da l e p generiamo la domanda (es. *How many dogs are black?*).

Come mostrato nell'esempio sopra descritto, nel generare gli scenari abbiamo posto come restrizione che il numero di immagini n_1 etichettate con il restrittore sia uguale o maggiore a 6. Sulla base delle proporzioni che abbiamo usato per definire i quantificatori, infatti, tale numero rappresenta il valore minimo per coprire tutti i 5 casi di quantificazione. Questo significa che i modelli non possono migliorare la loro accuratezza semplicemente imparando la correlazione tra valori bassi di n_1 e la non plausibilità di *few/most* negli scenari associati. Inoltre, gli scenari sono uniformemente distribuiti tra i 5 quantificatori. Di conseguenza, ogni quantificatore descrive circa 2000 scenari.[3]

Rappresentazioni visive Per ogni immagine in ciascuno scenario abbiamo estratto una rappresentazione visiva usando una tecnica che si basa su reti neurali convoluzionali (CNN) (Simonyan and Zisserman, 2014). In particolare, abbiamo usato il modello VGG-19 preaddestrato sui dati di ImageNet ILSVRC (Russakovsky et al., 2015) e il pacchetto MatConvNet (Vedaldi and Lenc, 2015) per l'estrazione delle features. Ogni immagine è rappresentata da un vetttore di 4096 dimensioni estratto dal settimo layer totalmente connesso ("fully-connected layer") e successivamente ridotto a un vettore di 400 dimensioni usando la tecnica SVD. Questi vettori di 400 dimensioni rappresentano l'input visuale dei nostri modelli.

Rappresentazioni linguistiche La domanda è espressa mediante due parole: $parola_1$, il restrittore del quantificatore, e $parola_2$, la sua portata. Ogni parola è rappresentata da un vettore di 400 dimensioni costruito usando l'architettura CBOW del pacchetto word2vec (Mikolov et al., 2013) e i migliori parametri di (Baroni et al., 2014). Il corpus usato per costruire lo spazio semantico, contentente circa 2.8 miliardi di tokens, è una concatenzione di UKWaC, un ampio estratto in inglese di Wikipedia 2009 e il British National Corpus (BNC).

3 Modelli

iBOWIMG è il modello di domanda e risposta su immagini (VQA) di (Zhou et al., 2015) adattato al nostro compito, cioè l'apprendimento dei quantificatori. Le domande sono rappresentate da un vettore che mette insieme tutte le parole del vocabolario indicando con "1" quelle presenti ("one-hot bag-of-words") ed elaborando un vettore basato su proprietà salienti delle parole ("word feature embedding"). La rappresentazione linguistica è

[3]Sia i codici che il dataset usati nel presente lavoro saranno resi disponibili per studi successivi.

concatenata con quella visiva. Il vettore è quindi usato come input da un classificatore (*softmax*) che sceglie la risposta considerando tutto il vocabolario. L'ultimo passaggio puó esser visto come un modello di regressione logistica multi-classe. Per adattarlo al nostro scopo, abbiamo modificato il modello originario in due modi. Abbiamo convertito lo scenario in una singola immagine, concatenando i vettori delle sedici immagini ottenendo un vettore di 6400 dimensioni. Inoltre, la risposta deve essere scelta tra cinque casi (i cinque quantificatori), per cui il vocabolario in output consiste di cinque nodi.

Rete neurale dei quantificatori (RNQ) Questo modello sfrutta i vantaggi delle reti neurali e riesce ad imaparare anche dai propri errori tramite *backpropagation*. Inoltre, disponendo di celle in cui archiviare le rappresentazioni vettoriali delle singole entità, riesce ad ottenere un'astrazione dello scenario alla quale contribuiscono tutte gli oggetti dello scenario. I passi della rete nurale sono i seguenti: (1) i vettori visuali e linguistici sono transformati in uno spazio di trecento dimensioni ($V1$). (2) I vettori visuali delle sedici immagini sono archiviati nelle celle della "memoria": per ogni cella calcoliamo la somiglianza tra ciascun vettore visuale e il vettore linguistico rappresentante il nome della domanda (e.s. *cane*); a tal scopo utilizziamo la norma euclidea. Il risultato è un vettore "di somiglianza" di sedici dimensioni ($S1$). (3) Calcoliamo quindi i vettori pesati di ogni entità moltiplicando le celle della memoria $V1$ con i valori corrispondenti di somiglianza in $S1$. Questo ci dà la rappresentazione di quanta "caninità" ci sia in ognuno degli oggetti. (4) `Sunto 1` è calcolato sommando le celle della memoria con i vettori pesati e rappresenta quanta "caninità" ci sia in un certo scenario. (5) Calcoliamo il prodotto scalare tra i vettori pesati ($W1$) e il vettore linguistico della proprietà (e.s. *nero*), e normalizziamo i valori con la norma euclidea. Il risultato è un vettore di somiglianza di sedici dimensioni ($S2$). Un secondo vettore pesato $W2$ è ottenuto moltiplicanto $W1$ e $S2$. Questo ci da la quantità di "caninità nera" in ogni oggetto. (6) `Sunto 2` è ottenuto sommando i nuovi vettori pesati che sono archiviati nelle celle della memoria e rappresenta quanta "caninità nera" ci sia in un certo scenario. (7) `Sunto 1` e `Sunto 2` sono concatenati in un singolo vettore di seicento dimensioni che viene poi trasformato linearmente in un vettore di cinque dimensioni. (8) Applichiamo quindi un classificatore (*softmax*) che prende come input il vettore di cinque dimensioni e restituisce per ogni quantificatore la probabilità che esso sia la risposta giusta. La rete neurale impara la proporzione tra il nome e la proprietà che caratterizza la relazione espressa dal quantificatore.

4 Risultati

Il modello a rete neurale `RNQ` ottiene un'accuratezza del 38.9% contro il 30.8% del modello `iBOWIMG`. Il modello `RNQ` va molto bene con *no* e *all* e va abbastanza bene con *few* e *most*. Ma la sua accuratezza diminuisce per *some* (18.16%), incidendo negativamente sul risultato complessivo. Una diversa analisi emerge invece dai risultati dell'altro modello, che va abbastanza bene nel predire *no* e *all*, ma va molto male negli altri casi. Per esso, *few*, *most* e *some* sono ugualmente difficili da imparare. Sono da rimarcare i suoi errori in particolare nell'apprendere *few* e *most* (si veda la Tabella 1). Questi quantificatori richiedono una comprensione più precisa delle proporzioni tra nome e proprietà di quanta non ne richiedano *no* e *all*. Pensiamo, quindi, che ciò dimostri i limiti di un modello che non impara ad astrarre ed elaborare le proporzioni. Il modello `RNQ` compie errori comprensibili, confondendo ad esempio *few* con *no* ma non con *some*, *most* e *all*. Ciò dimostra che questo modello ha imparato a quantificare e non semplicemente a tener traccia delle correlazioni nel dataset. Al contrario, l'altro modello confonde *few* con *no* e con *all* in numero uguale.

In generale, il modello `RNQ` risulta essere piut-

RNQ					
	some	all	no	few	most
some	73	88	57	89	95
all	29	**211**	20	19	125
no	32	28	**240**	70	32
few	46	53	104	**129**	68
most	49	148	31	38	126
iBOWIMG					
	some	all	no	few	most
some	89	77	50	108	78
all	45	**163**	63	46	87
no	30	69	**199**	59	52
few	82	81	100	85	52
most	75	110	63	64	80

Tabella 1: Matrice di confusione.

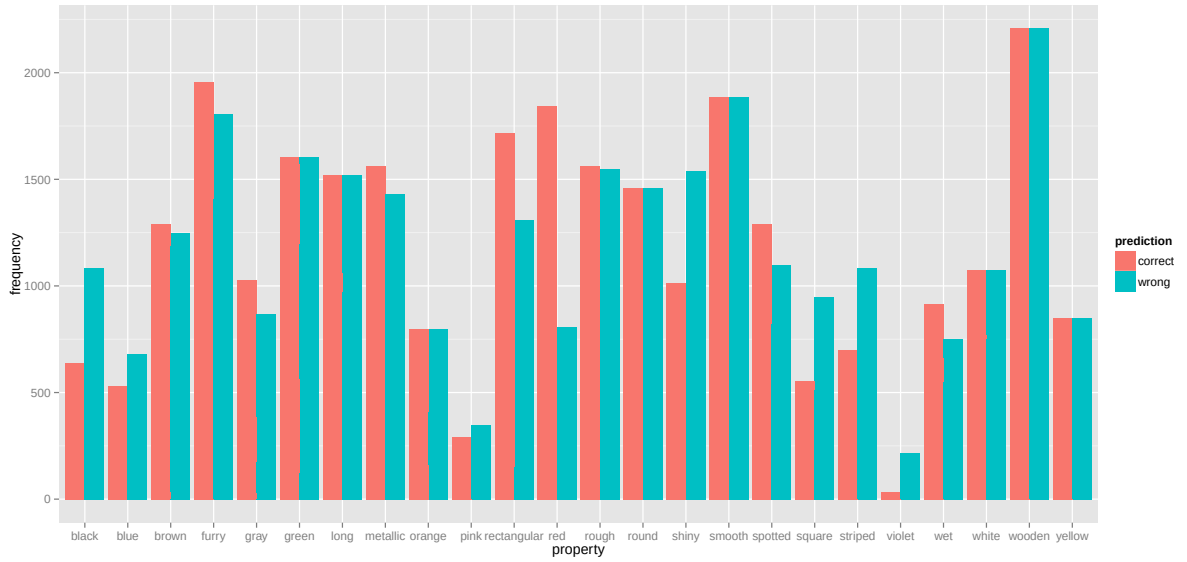


Figura 1: Risposte corrette e risposte sbagliate del modello in relazione alla frequenza della combinazione nome-proprietà.

tosto immune a possibili effetti collegati alla frequenza con cui ciascuna delle 24 proprietà occorre nel dataset. Si potrebbe pensare che combinazioni molto frequenti vengano imparate meglio di combinazioni poco frequenti, ma questo si rivela essere vero solo per pochi casi, su cui spiccano *red* e *rectangular*. Nella maggior parte dei casi, il modello si comporta in modo simile, in termini di riposte giuste/sbagliate, di fronte a combinazioni poco o molto frequenti. Nel caso di alcuni colori (*black*, *blue* e *violet*), notiamo un'accuratezza più alta per combinazioni poco frequenti rispetto a casi molto frequenti (vedi Figura 1). Possiamo dire che la performance del modello RNQ non dipende da effetti di frequenza delle combinazioni presenti nel dataset. Di nuovo, quindi, non si tratta di imparare correlazioni presenti nel dataset, ma per svolgere il compito di quantificazione è necessaria una comprensione più profonda dello scenario visivo e della domanda.

Ringraziamenti

Il secondo autore ringrazia l'Erasmus Mundus European Masters Program in Language and Communication Technologies (EM LCT). Gli altri autori sono stati finanziati dal progetto COMPOSES (ERC 2011 Starting Independent Research Grant n. 283554). Ringraziamo NVIDIA Corporation per la donazione delle GPU usati per questa ricerca.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *International Conference on Computer Vision (ICCV)*.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.

Sangeet Khemlani, Sarah-Jane Leslie, and Sam Glucksberg. 2009. Generics, prevalence, and default inferences. In *Proceedings of the 31st annual conference of the Cognitive Science Society*, pages 443–448. Cognitive Science Society Austin, TX.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Microsoft COCO: Common Objects in Context*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

M. Ren, R. Kiros, and R. Zemel. 2015. Image question answering: A visual semantic em- bedding model and a new dataset. In *In International Conference on Machine Learningt Deep Learning Workshop*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al.

2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Andrea Vedaldi and Karel Lenc. 2015. *MatConvNet – Convolutional Neural Networks for MATLAB*. Proceeding of the ACM Int. Conf. on Multimedia.

B. Zhou, Y. Tian, S. Suhkbaatar, A. Szlam, and R. Fergus. 2015. Simple baseline for visual question answering. Technical report, arXiv:1512.02167, 2015.

Appetitoso: A Search Engine for Restaurant Retrieval based on Dishes

Gianni Barlacchi[1,2], **Azad Abad**[1], **Emanuele Rossinelli**[3], **Alessandro Moschitti**[1,4]
[1]Department of Information Engineering and Computer Science, University of Trento
[2] TIM Semantics and Knowledge Innovation Lab, Trento
[3] Kloevolution S.r.l.
[4]Qatar Computing Research Institute, HBKU
{gianni.barlacchi,e.rossinelli,amoschitti}@gmail.com
azad.abad@unitn.it

Abstract

English. Recent years have seen an impressive development and diffusion of web applications to food domains, e.g., Yelp, TripAdvisors. These mainly exploit text for searching and retrieving food facilities, e.g., *restaurants*, *caffé*, *pizzerias*. The main features of such applications are: the location and quality of the facilities, where quality is extrapolated by the users' reviews. More recent options also enable search based on restaurant categorization, e.g., *Japanese*, *Italian*, *Mexican*. In this work, we introduce Appetitoso[1], an innovative approach for finding restaurants based on the dishes a user would like to taste rather than using the name of food facilities or their general categories.

Italiano. *Recentemente si è assistito ad un impressionante sviluppo e diffusione di applicazioni web per il dominio del cibo, e.g.,* Yelp, TripAdvisors. *Queste sfruttano principalmente il testo per la ricerca e il recupero di punti di ristoro, e.g.,* ristoranti, bar, pizzerie. *Le caratteristiche principali usate dalle applicazioni sono: la posizione e la qualitá delle strutture che servono il cibo, dove la qualitá é estrapolata dalle recensioni degli utenti. Opzioni piú recenti consentono anche la ricerca in base alla categoria del ristorante, e.g.,* Giapponese, Italiano, Messicano. *Questo articolo introduce Appetitoso, un nuovo modo di trovare punti di ristoro sulla base dei piatti che il cliente vuole gustare invece che sul nome del ristorante o su categories generali.*

1 Introduction

In late 2000's, we assisted to the explosion of TripAdvisor[2], the world's largest travel site, which offers advice about hotel and restaurants. In few years, it has revolutionized the restaurant industry, allowing its users to search restaurants by location, broad food categories (e.g., Mexican, Italian, French), reviews and ratings provided by other users.

However, the user expectation has evolved overtime: looking for restaurants is not enough anymore, people are now considering finer-grained properties of food, e.g., a particular way to cook a dish along with its specific ingredients. Thus, there is a clear gap between what the market proposes and the emerging trends.

In this work, we present Appetitoso, a search engine that seeks for restaurants based on dishes. This approach is designed to help users to find their restaurants having already a specific dish preference in mind, using fine-grained properties of the dish.

Appetitoso integrates state-of-the-art search engines, such as BM25, with a domain specific knowledge base describing properties and similarity relations between different Italian dishes. This knowledge is very useful, e.g., in our experiments, we show that it greatly boosts dish retrieval.

Appetitoso is available as a mobile phone application (e.g., Android and iOS) and website, released in 2014 for two languages, English and Italian. It is an end-to-end application for finding restaurants offering the desired dish. We evaluated it using a set of 547 popular queries typed by its users in the cities of Rome, Milan and Florence.

In the reminder of this paper, in Section 2, we report related work on systems for automatic food recommendation, In Section 3, we introduce Appetitoso, its knowledge base and the food search

[1]http://www.appetitoso.it

[2]http://www.tripadvisor.com

engine. Section 4, we describe our experiments on restaurant retrieval on Italian language and finally, in Section 5, we provide our conclusion.

2 Related Work

Nowdays, the importance of data analysis is becoming fundamental in many fields. From telecommunications to social media, the huge amount of available data allows scientists and researchers to address previously unsolved problems (Barlacchi et al., 2015). The food domain represents one of the field in which emerging big data techniques demonstrated to be very promising and able to impact the every daily life of people. In recipe recommendation, for instance, Teng et al. (2012) proposed an approach based on networks of ingredients, which has been built from a dataset of recipes. In order to capture both ingredient relations and users' knowledge for combining ingredients in new recipes, they created two separate networks used for recipe recommendation.

Moreover, Ahn et al. (2011) explored the impact of flavor compounds on ingredient combinations through a network-based approach. An interesting application was developed by IBM with Chef Watson[3], which is part of the cognitive computing applications developed by the company. The system models the chemical compounds of different ingredients together with textual information extracted from thousands recipes for suggesting new ones using innovative ingredient combinations.

Among the different kinds of data, text surely represents one of the richest sources of information from which we can extract a wide range of statements about food. The use of text in food domain has been widely explored showing promising results with different models, ranging from the measurement of sentiment in food reviews (Kang et al., 2012) and relation extraction (Wiegand and Klakow, 2013; Wiegand et al., 2012), to the prediction of attribute reviews in recipes (Druck, 2013).

3 Appetitoso

We introduce the idea of searching a dish and then finding the best restaurants that can offer it. Thus, the aim of our search engine, Appetitoso, is to find the best restaurants offering dishes relevant to the user's request. Starting from a query with food-related content, e.g., *bistecca alla fiorentina*

[3]https://www.ibmchefwatson.com

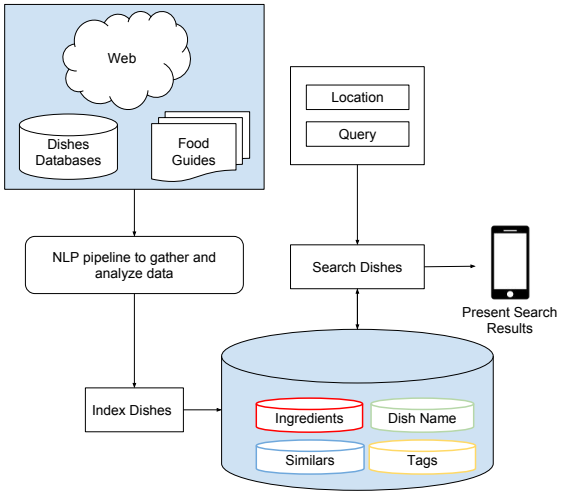


Figure 1: Architecture of Appetitoso.

(t-bone steak), the system retrieves places that satisfy the constraint on the location and, at the same time, prepare the desired dish or similar dishes.

Appetitoso retrieves restaurants from a semistructured database, Food Taste Knowledge Base (FKB), which contains text descriptions of dishes and restaurants: we in part manually inserted them or gathered them from various sources such as foodblogs, restaurants reviews and food guides. The search processes is divided in two phases: first, the user has to type the query and a location, e.g., the address of a target place or the current user position captured by GPS. These are both sent to the Appetitoso's search engine, which retrieves a list of related dishes from FKB. The results are grouped by dish name and shown to the user in different *course* categories, i.e., *antipasto*/entree, *primo*/first course, *secondo*/second course, dessert. The input location is used to restrict the search area of interest, relying on the restaurant position available in FKB.

The second phase of the searching process is devoted to select the best restaurant. Once the user chooses a dish from the list above, Appetitoso provides a list of restaurants that offer such food speciality. Indeed, all the restaurants offering that dish are stored in FBK. Additionally, Appetitoso provides a *DishScore*[4] for each restaurant, which is a measure of the goodness of the dish in that restaurant. Fig. 1 shows the high-level architecture

[4]We only inserted restaurant that have a good reputation in FBK. In order to generate the *DishScore*, we trained a logistic regression over 5 different review scores, e.g., 1 star, 2 star etc. We used various features, e.g., Tripadvisor and food guide scores. This description is however beyond the purpose of the current paper.

of the system. In the next section, we illustrate our FKB, which enables accurate retrieval of similar dishes.

3.1 The Food Taste Knowledge Base (FKB)

A quick analysis of Italian menus clearly show that, in many cases, the name of a dish is not enough to understand its content, which means that names do not support an accurate similarity measures between dishes. Thus, we created FKB, which also organizes dishes in a hierarchical structure, where each node is connected to others in case there is a similarity between them.

For instance, *Bucatini alla amatriciana* (bucatini with amatriciana sauce) can be extended from *Spaghetti alla amatriciana* (spaghetti with amatriciana sauce) since the only difference between the two dishes is the type of pasta (*spaghetti* vs. *bucatini*). In this case, we marked the first dish as a template for the second one. The relation is one-to-many: one dish can be a template for many others but it can be only assigned to one template. Since every restaurant can have its own way to prepare the dish, multiple instances of the same dish can be present in the FKB. We differentiate them by adding the restaurant ID.

Since there is no defined way to assess the similarity between two dishes: they may be similar as they are made by similar ingredients or because they are cooked in the same way, we built the FKB hierarchy with a semi-automatic approach. We used name similarity to select similar candidates, which are then manually annotated by food experts. We manually populated FKB with data collected from the web, food guides and foodblogs. Every dish belonging to a restaurant is represented by means of the following information:

- ID: unique identifier for the dish.
- Name of the dish: the name of the dish as reported in the restaurant menu.
- Ingredients: list of the principal ingredients. When the ingredients are not provided by the restaurant, we use a list of common ingredients for the dish (e.g., ingredients from online recipes).
- Tags: list of tags useful to characterize the dish. The tag list does not include ingredients but only categorical information that can help to characterize the dish (e.g., meat or fish).
- Similar dishes: list of similar dishes defined according to our hierarchy described above.
- Template: ID of the template dish, if it is present.

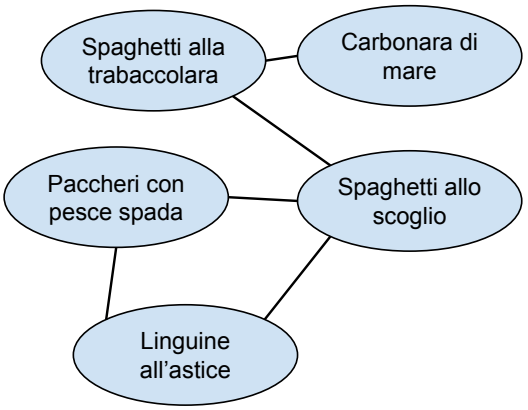


Figure 2: Connection between similar fish dishes.

- Restaurant: information about the restaurant that cook this dish (e.g. restaurant name and restaurant ID).
- DishScore: a value that indicates the goodness of the dish. It is calculated taking into account many factors such as the reputation of the restaurant in cooking that dish, the number of mentions in food guide and the sentiment extracted from foodblogger articles and restaurant reviews.

This hierarchical organization is very powerful and allows us to easily keep track of similarities that are not explicit. Fig. 2 shows an example of connections between similar dishes. It is worth to mention that Appetitoso aims to suggest only restaurants that own a good reputation in cooking target dishes, i.e., restaurants in Rome that are famous for *pasta alla carbonara*. Consequently, this limits the number of dishes contained in the FKB and thus on the territory coverage. On the other hand, it makes it possible to create a manually checked resource.

3.2 Dish Retrieval

Italy has long and variegated traditions on preparing food: it is possible to find different kinds of cuisine even in nearby cities. This makes the Italian food incredibly varied and fascinating, but, at the same time, difficult to interpret from a linguistic viewpoint. The same dish can be called in many different ways. In Florence people call *Carabaccia* the common dish *Zuppa di cipolle* with the consequence that the underlying retrieval problem cannot be addressed by just using a simple word matching approach. Indeed, even if a dish is conceptually the same of another, different restaurants (e.g., in different locations) have their own way to call it.

To tackle the problem above, we verified the hypothesis that a search engine can achieve a better result if we consider further information such as ingredients and tags. This approach significantly improves the accuracy of the retrieved list comparing to the simple word matching approach.

More specifically, we applied BM25 (Robertson et al., 1995) to FKB. Given a dish query, Q and a representation of a candidate dish, D, BM25 ranks the latter according to the following score:

$$s(Q,D) = \sum_{i=0}^{n} \frac{IDF\left(q_i\right) \cdot ((k+1) \cdot TF(q_i, D))}{\left(k \cdot (1 - b + b \cdot \frac{|D|}{avgD}) + TF(q_i, D)\right)},$$

where k and b are two free parameters that modify respectively the impact of term frequency (TF) and the document length through the term $\frac{|D|}{avgD}$, $|D|$ is the document length and $avgD$, i.e., the average of D over the whole dataset. Finally, $IDF(q_i)$ is the Inverse Document Frequency for the query term q_i, computed as:

$$log\left(1 + \frac{(N - DF(q_i) + 0.5)}{(DF(q_i) + 0.5)}\right),$$

where N is the total number of documents in the collection, and $DF(q_i)$ is the document frequency of the term q_i.

Additionally, we created four different indexes[5] with the information contained in FKB, i.e., the (i) dish name, (ii) ingredients, (iii) tags and (iv) similar dishes. Each list is built using the words describing the four items above. Thus, when we query a dish, we first retrive four different sets of results and then, since they have different importance, we combine them together assigning different weights, where the latter are set using cross-fold validation.

4 Experiments

Our experiments aim at demonstrating the effectiveness of our models on the task of dish retrieval. We used the well known metrics: Precision at rank 1 (P@1), Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP). P@1 indicates the percentage of queries with a correct answer (e.g., the desired dish) found in the first position. The MRR is computed as follows: MRR= $\frac{1}{|Q|}\sum_{q=1}^{|Q|}\frac{1}{\textit{rank(q)}}$, where *rank(q)* is the position of the first correct answer in the retrieved list. For a group of queries Q, MAP is the mean over the average precision scores for each query: $\frac{1}{Q}\sum_{q=1}^{Q} AveP(q)$.

[5]We use Lucene (McCandless et al., 2010)

Model	City	MRR	MAP	P@1
Baselines				
String Matching (on entire names)	Milan	53.28	53.28	53,28
	Rome	71.23	71.23	71.23
	Florence	44.87	44.87	44.87
	All	**56.46**	**56.46**	**56.46**
BM25 (on names only)	Milan	69.75	65.44	68.18
	Rome	63.86	60.32	58.90
	Florence	42.31	40.94	37.18
	All	**58.64**	**55.56**	**54.75**
Our Model				
Appetitoso (names, ingredients tags, similar names)	Milan	95.35	85.69	93.43
	Rome	87.40	76.23	84.93
	Florence	83.55	75.38	78.21
	All	**88.76**	**79.10**	**85.52**

Table 1: Ranking evaluation for different models

Due to the fact that FKB contains multiple instances of the same dish, we evaluated the collapsed list of results by considering the dish name. It is worth to mention that the names of the dishes are not standard, thus some dishes are the same still having slightly different names. To make them more similar, we normalized name forms by removing space, articles and punctuation. We considered a set of 547 popular queries typed by users in Milan (396 queries), Rome (73 queries) and Florence (78 queries). The number of retrieved dishes varies for the different queries with averages of 22.8, 22.3 and 37.4, for Florence, Milan and Rome, respectively. For each retrieved dish, we manually annotated the relevance respect to the input query. It should be noted that the same dish is associated (in FKB) with all of the restaurants that are offering it. Thus, restaurant retrieval is a side effect of dish retrieval.

We considered two baselines for evaluating our model, namely, String Matching and BM25. The first is based on simple string matching between the query and the dish names. The second is BM25, which can be applied to dish names only. We refer to our system (BM25 applied to the 4 indexes as described in Sec. 3.2) with the name Appetitoso.

Table 1 shows the results of the baselines and our model by cities and overall (All). Appetitoso largely outperforms String Matching and BM25 applied to names only, e.g., up to 32 and 24 absolute percent points in MRR and MAP, respectively.

5 Conclusion

In this paper we presented Appetitoso, a semantic search engine for food. The aims of the

search engine is to provide the users with a way of searching restaurants by dishes rather than just using the restaurants' address or cuisine type. We show that, given the complexity of dish naming, a semistructured database for dishes can largely improve BM25. Overall, Appetitoso shows good performance, e.g., achieving 88.76% in MAP. In the future, we would like to include more complex unstructured data such as the description of the dishes and also explore the possibility of word embeddings for the food domain. Moreover, it is also important increase the coverage of the system by adding more dishes to the FKB. Even if the manual annotation is important, and in some cases fundamental, it represents a bottleneck for the expansion process. For this reason, in the future it would be necessary consider approaches to automatically extract dish entities from text (e.g. NER for food).

Acknowledgments

We would like to thank the Appetitoso team for making available the system and for providing us with the data for this work. This work has been partially supported by the EC project CogNet, 671625 (H2020-ICT-2014-2, Research and Innovation action) and by an IBM Faculty Award. The first author was supported by a fellowship from TIM. Many thanks to the anonymous reviewers for their valuable suggestions.

References

Yong-Yeol Ahn, Sebastian E Ahnert, James P Bagrow, and Albert-László Barabási. 2011. Flavor network and the principles of food pairing. *Scientific reports*, 1.

Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. 2015. A multi-source dataset of urban life in the city of milan and the province of trentino. *Scientific data*, 2.

Gregory Druck. 2013. Recipe attribute prediction using review text as supervision. In *Cooking with Computers 2013, IJCAI workshop*.

Hanhoon Kang, Seong Joon Yoo, and Dongil Han. 2012. Senti-lexicon and improved naïve bayes algorithms for sentiment analysis of restaurant reviews. *Expert Systems with Applications*, 39(5):6000–6010.

Michael McCandless, Erik Hatcher, and Otis Gospodnetic. 2010. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co., Greenwich, CT, USA.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, 109:109.

Chun-Yuen Teng, Yu-Ru Lin, and Lada A Adamic. 2012. Recipe recommendation using ingredient networks. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 298–307. ACM.

Michael Wiegand and Dietrich Klakow. 2013. Towards the detection of reliable food-health relationships. *NAACL 2013*, page 69.

Michael Wiegand, Benjamin Roth, and Dietrich Klakow. 2012. Data-driven knowledge extraction for the food domain. In *KONVENS*, pages 21–29.

Gamification for IR: The Query Aspects Game

Giorgio Maria Di Nunzio
Dept. of Inf. Eng. (DEI)
University of Padua, Italy
Via Gradenigo 6/a 35131
dinunzio@dei.unipd.it

Maria Maistro
Dept. of Inf. Eng. (DEI)
University of Padua, Italy
Via Gradenigo 6/a 35131
maistro@dei.unipd.it

Daniel Zilio
Dept. of Inf. Eng. (DEI)
University of Padua, Italy
Via Gradenigo 6/a 35131
daniel.zilio.3@studenti.unipd.it

Abstract

English.

The creation of a labelled dataset for Information Retrieval (IR) purposes is a costly process. For this reason, a mix of crowdsourcing and active learning approaches have been proposed in the literature in order to assess the relevance of documents of a collection given a particular query at an affordable cost. In this paper, we present the design of the gamification of this interactive process that draws inspiration from recent works in the area of gamification for IR. In particular, we focus on three main points: i) we want to create a set of relevance judgements with the least effort by human assessors, ii) we use interactive search interfaces that use game mechanics, iii) we use Natural Language Processing (NLP) to collect different aspects of a query. [1]

Italiano

La creazione di una collezione sperimentale per l'Information Retrieval (IR) è un processo costoso sia dal punto di vista economico che in termini di sforzo umano. Per ridurre i costi legati all'attribuzione dei giudizi di rilevanza ai documenti di una collezione, sono stati proposti approcci che integrano tecniche di crowdsourcing e active learning. In questo paper viene presentata un'idea basata sull'utilizzo della gamification *('ludicizzazione') in IR per l'attribuzione di giudizi di rilevanza in maniera semi-automatica. In particolare, ci focalizzeremo su tre aspetti principali: i) si vuole creare una collezione in modo che l'assegnazione dei giudizi da parte dei valutatori richieda il minor sforzo possibile, ii) per mezzo di un'interfaccia che utilizza dinamiche di gioco iii) insieme a tecniche di NLP per la riformulazione della query.*

[1] This paper is partially an extended abstract of the paper "Gamification for Machine Learning: The Classification Game" presented at the GamifIR 2016 Workshop co-located with SIGIR 2016 (Di Nunzio et al., 2016)

1 Introduction

In Information Retrieval (IR), the performance of a system is evaluated using experimental test collections. These collections consist of a set of documents, a set of queries, and a set of relevance judgments, where each judgement explains whether a document is relevant or not to each query. The creation of relevance judgements is a costly, time-consuming, and non-trivial task; for these reasons, the interest in approaches that generate relevance judgements with the least amount of effort has increased in IR and related areas (i.e., supervised Machine Learning (ML) algorithms). In the last years, mixed approaches that use crowdsourcing (Ho et al., 2013) and active learning (Settles, 2011) have shown that it is possible to create annotated datasets at affordable costs. Specifically, crowdsourcing has been part of the IR toolbox as a cheap and fast mechanism to obtain labels for system evaluation. However, successful deployment of crowdsourcing at scale involves the adjustment of many variables, a very important one being the number of assessors needed per task, as explained in (Abraham et al., 2016).

1.1 Search Diversification and Query Reformulation

The effectiveness of a search and the satisfaction of users can be enhanced through providing various results of a search query in a certain order of relevance and concern. The technique used

to avoid presenting similar, though relevant, results to the user is known as a diversification of search results (Abid et al., 2016). While existing research in search diversification offers several solutions for introducing variety into the results, the majority of such work is based on the assumption that a single relevant document will fulfil a user's information need, making them inadequate for many informational queries. In (Welch et al., 2011), the authors propose a model to make a tradeoff between a user's desire for multiple relevant documents, probabilistic information about an average user's interest in the subtopics of a multifaceted query, and uncertainty in classifying documents into those subtopics.

Most information retrieval systems operate by performing a single retrieval in response to a query. Effective results sometimes require several manual reformulations by the user or semi-automatic reformulations assisted by the system. Diaz presents an approach to automatic query reformulation which combines the iterated nature of human query reformulation with the automatic behavior of pseudo relevance feedback (Diaz, 2016). In (Azzopardi, 2009), the author proposes a method for generating queries for ad-hoc topics to provide the necessary data for this comprehensive analysis of query performance. Bailey et al. explore the impact of widely differing queries that searchers construct for the same information need description. By executing those queries, we demonstrate that query formulation is critical to query effectiveness (Bailey et al., 2015).

1.2 Gamification in IR

Gamification is defined as "the use of game design elements in non-game contexts" (Deterding et al., 2011), i.e. tipical game elements are used for purposes different from their normal expected employment. Nowadays, gamification spreads through a wide range of disciplines and its applications are implemented in different and various aspects of scientific fields of study. For instances, gamification is applied to learning activities (Kotini and Tzelepi, 2015; Kapp, 2012), business and enterprise (Jurado et al., 2015; Stanculescu et al., 2016; Thom et al., 2012) and medicine (Eickhoff, 2014; Chen and Pu, 2014).

IR has recently dealt with gamification, as witnessed by the Workshop on Gamification for Information Retrieval (GamifIR) in 2014, 2015 and 2016. In (Galli et al., 2014), the authors describe the fundamental elements and mechanics of a game and provide an overview of possible applications of gamification to the IR process. In (Shovman, 2014), approaches to properly gamify Web search are presented, i.e. making the search of information and the scanning of results a more enjoyable activity. Actually, many proposals of game applied to different aspects of IR have been presented. For example in (Maltzahn et al., 2014), the authors describes a game that turns document tagging into the activity of taking care of a garden, with the aim of managing private archives. A method to obtain ranking of images by utilizing human computation through a gamified web application is proposed in (Lux et al., 2014). Fort et al. introduce a strategy to gamify the annotation of a French corpora (Fort et al., 2014).

In this paper, we want to apply game mechanics to the problem of relevance assessment with three goals in mind: i) we want to create a set of relevance judgements with the least effort by human assessors, ii) we use interactive search interfaces that use game mechanics, iii) we use NLP to collect different aspects of a query. In this context, we can define our web application as a Game with a Purpose (GWAP), that is a game which presents some purposes, usually boring and dull for people, within an entertaining setting, in order to make them enjoyable and to solve problem with the aid of human computation. The design and the implementation of this interactive interface will be used as a post-hoc analysis of two Text REtrieval Conference (TREC)[2] 2016 tracks, namely the Total Recall Track and the Dynamic Domain Track. These two tracks are interesting for our problem since they both re-create a situation where we need to find the best set (or the total amount) of relevant documents with the minimum effort by the assessor that has to judge the documents proposed by the system given an information need.

2 Design of the Experiment

In this first pilot study, we will implement a simple game based on a visual interpretation of probabilistic classifiers (Di Nunzio, 2014; Di Nunzio, 2009; Di Nunzio and Sordoni, 2012). The game consists in separating two sets of colored points on a two-dimensional plane by means of a straight line, as shown in Figure 1. Despite its simplicity,

[2] http://trec.nist.gov

this very abstract scenario received a good feedback by kids of primary schools who tested it during the European Researcher's Night at the University of Padua[3]. The next step will be to design and implement the game with real game development platforms like, for example, Unity[4] or Marmalade[5].

2.1 The Classification Game

The 'original game' (Di Nunzio et al., 2016) is based on the two-dimensional representation of probabilities (Di Nunzio, 2014; Singh and Raj, 2004), which is a very intuitive way of presenting the problem of classification on a two-dimensional space. Given two classes c_1 and c_2, an object o is assigned to category c_1 if the following inequality holds:

$$\underbrace{P(o|c_2)}_{y} < m\,\underbrace{P(o|c_1)}_{x} + q \tag{1}$$

where $P(o|c_1)$ and $P(o|c_2)$ are the likelihoods of the object o given the two categories, while m and q are two parameters that depend on the misclassification costs that can be assigned by the user to compensate for either the unbalanced classes situation or different class costs.

If we interpret the two likelihoods as two coordinates x and y of a two dimensional space, the problem of classification can be studied on a two-dimensional plot. The decision of the classification is represented by the 'line' $y = mx + q$ that splits the plane into two parts, therefore all the points that fall 'below' this line are classified as objects that belong to class c_1 (see Figure 1 for an example). Without entering into the mathematical details of this approach (Di Nunzio, 2014), the basic idea of the game is that the players can adapt the two parameters m and q in order to optimize the separation of points and, at the same time, can use their resources to improve the estimate of the two likelihoods by buying training data, and/or add more points to the plot, by buying validation data.

3 The Query Aspects Game

The classification game can be easily adjusted into a relevance assessment game if the two classes are 'relevant' and 'non-relevant' (we assume only binary relevance assessment for the moment). However, while in the classification game we already have a set of labelled documents and the goal is to find the optimal classifier, in this new game we need to find the relevant documents. To this purpose, we will follow the idea of the works described in the following subsections: i) building assessment by varying the description of the information need, ii) using an interactive interface that suggests the amount of relevant information that has to be judged, iii) using NLP approaches to generate variations of a query.

3.1 Building Relevance Assessments With Query Aspects

In (Efron, 2009), the author presents a method for creating relevance judgments without explicit relevance assessments. The idea is to create different "aspects" of a query: given a query q and a set of documents D, the same information need that generated q could also generate other queries that focus on another aspects of the same need. A query aspect is an articulation of a searcher's information need which might be a re-elaboration (for example, rephrasing, specification, or generalization) of the query. By generating several queries related to an information need and running each of these against our document collection, we can create a pool of results where each result set pertains to a particular aspect of the information need with a limited human effort.

In practice, in order to build a set of relevance assessments for q, we generate a number of query aspects using a single IR system. Then, the union of the top k documents retrieved for each aspect constitutes a list of pseudo-relevance assessments for the query q.

3.2 An Interactive Interface to Generate Query Aspects

Building different aspects of the same information need is not an easy task. As explained in (Umemoto et al., 2016), searchers often cannot easily come up with effective queries for collecting documents that cover diverse aspects. In general, experts that have to search for relevant documents usually have to issue more queries to complete the tasks if search engines return few documents relevant to unexplored aspects. Moreover, quitting this tasks too early without in-depth exploration prevents searchers from finding essential information.

[3] http://www.venetonight.it/padova/
[4] https://unity3d.com
[5] https://www.madewithmarmalade.com/

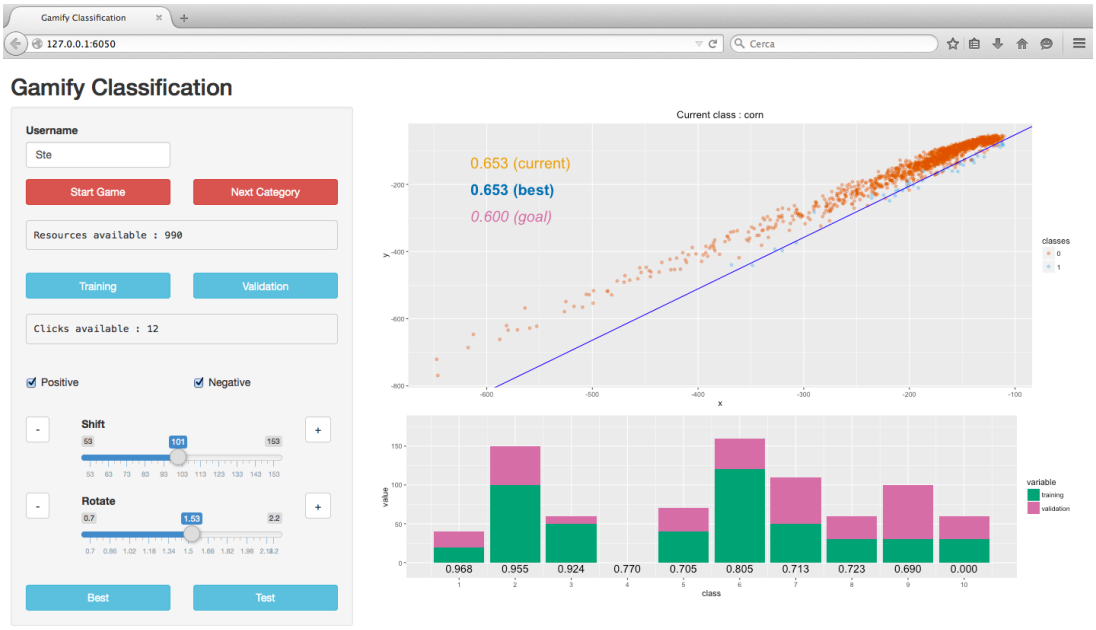


Figure 1: Layout of the original "classification game" that will be adapted to the "query aspects game".

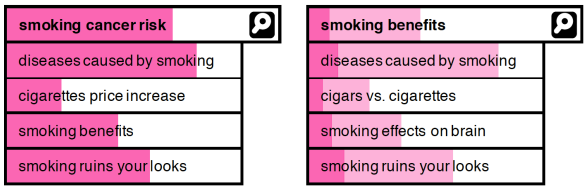


Figure 2: Scentbars and the visualization of missing information. Figure from (Umemoto et al., 2016)

Umemoto et al. propose an interactive interface, named ScentBar, that helps searchers to visualize the amount of missing information for both the search query and suggestion queries in the form of a stacked bar chart. The interface, a portion of which is shown in Figure 2, visualizes an estimate of missing information for each aspect of a query that could be obtained by the searcher. When the user collects new information during the browsing of the results, the bars of the different query aspects change color to indicate the amount of effort that the system estimates necessary to find most of the relevant information. The estimates of the required effort to complete a task are formalized as as a set-wise metric were the gain for each aspects is represented by a conditional probability.

3.3 Using NLP to Generate Query Aspects

The last part of the design of the query aspects game involves the use of natural language processing techniques to propose variations of a query to express the same information need. This problem has been studied for more than twenty years in IR. In (Strzalkowski et al., 1997), the authors discuss how the simplest word-based representations of content, while relatively better understood, are usually inadequate since single words are rarely specific enough for accurate discrimination. Consequently, a better method is to identify groups of words that create meaningful phrases, especially if these phrases denote important concepts in the domain.

Some examples of advanced techniques of phrase extraction, including extended N-grams and syntactic parsing, attempt to uncover concepts, which would capture underlying semantic uniformity across various surface forms of expression. Syntactic phrases, for example, appear reasonable indicators of content since they can adequately deal with word order changes and other structural variations. In the literature, there are examples of query reformulation using NLP approaches for example to the modification and/or expansion of both parts thematic and geospatial that are usually recognized in a geographical query (Perea-Ortega et al., 2013), or to support the refinement of a vague, non-technical initial query into a more precise problem description (Roulland et al., 2007), or to predict search satisfaction (Hassan et al., 2013).

4 Conclusions and Future Works

In this work, we presented the requirements of the design of an interactive interface that uses game mechanics together with NLP techniques to generate variation of an information need in order to label a collection of documents. Starting from the successful experience of the gamification of a machine learning problem (Di Nunzio et al., 2016), we are preparing a new pilot study of the 'query aspects game' that will be used to generate relevant documents for two TREC tracks: the Total Recall track and the Dynamic Domain track. The results of this study will be available at the end of November 2016, and can be presented and discussed at the workshop.

References

Adnan Abid, Naveed Hussain, Kamran Abid, Farooq Ahmad, Muhammad Shoaib Farooq, Uzma Farooq, Sher Afzal Khan, Yaser Daanial Khan, Muhammad Azhar Naeem, and Nabeel Sabir. 2016. A survey on search results diversification techniques. *Neural Computing and Applications*, 27(5):1207–1229.

Ittai Abraham, Omar Alonso, Vasilis Kandylas, Rajesh Patel, Steven Shelford, and Aleksandrs Slivkins. 2016. How many workers to ask?: Adaptive exploration for collecting high quality labels. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 473–482, New York, NY, USA. ACM.

Leif Azzopardi. 2009. Query side evaluation: An empirical analysis of effectiveness and effort. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 556–563, New York, NY, USA. ACM.

Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. User variability and ir system evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 625–634, New York, NY, USA. ACM.

Yu Chen and Pearl Pu. 2014. Healthytogether: Exploring social incentives for mobile fitness applications. In *Proceedings of the Second International Symposium of Chinese CHI*, Chinese CHI '14, pages 25–34, New York, NY, USA. ACM.

Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From Game Design Elements to Gamefulness: Defining "Gamification". In *Proc. of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, MindTrek '11, pages 9–15, New York, NY, USA. ACM.

Giorgio Maria Di Nunzio and Alessandro Sordoni. 2012. A Visual Tool for Bayesian Data Analysis: The Impact of Smoothing on Naive Bayes Text Classifiers. In *Proc. of the ACM SIGIR'12 conference on research and development in Information Retrieval, Portland, OR, USA, August 12-16, 2012*, page 1002.

Giorgio Maria Di Nunzio, Maria Maistro, and Daniel Zilio. 2016. Gamification for machine learning: The classification game. In *Proceedings of the Third International Workshop on Gamification for Information Retrieval co-located with 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016), Pisa, Italy, July 21, 2016.*, pages 45–52.

Giorgio Maria Di Nunzio. 2009. Using Scatterplots to Understand and Improve Probabilistic Models for Text Categorization and Retrieval. *Int. J. Approx. Reasoning*, 50(7):945–956.

Giorgio Maria Di Nunzio. 2014. A New Decision to Take for Cost-Sensitive Naïve Bayes Classifiers. *Information Processing & Management*, 50(5):653 – 674.

Fernando Diaz, 2016. *Pseudo-Query Reformulation*, pages 521–532. Springer International Publishing, Cham.

Miles Efron. 2009. Using multiple query aspects to build test collections without human relevance judgments. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 276–287, Berlin, Heidelberg. Springer-Verlag.

Carsten Eickhoff. 2014. Crowd-powered experts: Helping surgeons interpret breast cancer images. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, GamifIR '14, pages 53–56, New York, NY, USA. ACM.

Karën Fort, Bruno Guillaume, and Hadrien Chastant. 2014. Creating zombilingo, a game with a purpose for dependency syntax annotation. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, GamifIR '14, pages 2–6, New York, NY, USA. ACM.

Luca Galli, Piero Fraternali, and Alessandro Bozzon. 2014. On the Application of Game Mechanics in Information Retrieval. In *Proc. of the 1st Int. Workshop on Gamification for Information Retrieval*, GamifIR'14, pages 7–11, New York, NY, USA. ACM.

Ahmed Hassan, Xiaolin Shi, Nick Craswell, and Bill Ramsey. 2013. Beyond clicks: query reformulation as a predictor of search satisfaction. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, CIKM '13, pages 2019–2028, New York, NY, USA. ACM.

Chien-Ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. 2013. Adaptive task assignment for crowdsourced classification. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*, volume 28 of *JMLR Proceedings*, pages 534–542. JMLR.org.

Jose Luis Jurado, Alejandro Fernandez, and Cesar A. Collazos. 2015. Applying gamification in the context of knowledge management. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*, i-KNOW '15, pages 43:1–43:4, New York, NY, USA. ACM.

Karl M Kapp. 2012. *The gamification of learning and instruction: game-based methods and strategies for training and education*. John Wiley & Sons.

Isabella Kotini and Sofia Tzelepi. 2015. A Gamification-Based Framework for Developing Learning Activities of Computational Thinking. In Torsten Reiners and C. Lincoln Wood, editors, *Gamification in Education and Business*, pages 219–252. Springer Int. Publ., Cham.

Mathias Lux, Mario Guggenberger, and Michael Riegler. 2014. Picturesort: Gamification of image ranking. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, GamifIR '14, pages 57–60, New York, NY, USA. ACM.

Carlos Maltzahn, Arnav Jhala, Michael Mateas, and Jim Whitehead. 2014. Gamification of private digital data archive management. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, GamifIR '14, pages 33–37, New York, NY, USA. ACM.

José M. Perea-Ortega, Miguel A. García-Cumbreras, and L. Alfonso Ureña López. 2013. Applying nlp techniques for query reformulation to information retrieval with geographical references. In *Proceedings of the 2012 Pacific-Asia Conference on Emerging Trends in Knowledge Discovery and Data Mining*, PAKDD'12, pages 57–69, Berlin, Heidelberg. Springer-Verlag.

Frédéric Roulland, Aaron Kaplan, Stefania Castellani, Claude Roux, Antonietta Grasso, Karin Pettersson, and Jacki O'Neill, 2007. *Query Reformulation and Refinement Using NLP-Based Sentence Clustering*, pages 210–221. Springer Berlin Heidelberg, Berlin, Heidelberg.

Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1467–1478.

Mark Shovman. 2014. The Game of Search: What is the Fun in That? In *Proc. of the 1st Int. Workshop on Gamification for Information Retrieval*, GamifIR'14, pages 46–48, New York, NY, USA. ACM.

Rita Singh and Bhiksha Raj. 2004. Classification in likelihood spaces. *Technometrics*, 46(3):318–329.

Laurentiu Catalin Stanculescu, Alessandro Bozzon, Robert-Jan Sips, and Geert-Jan Houben. 2016. Work and play: An experiment in enterprise gamification. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 346–358, New York, NY, USA. ACM.

Tomek Strzalkowski, Fang Lin, Jose Perez-Carballo, and Jin Wang. 1997. Building effective queries in natural language information retrieval. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, pages 299–306, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jennifer Thom, David Millen, and Joan DiMicco. 2012. Removing gamification from an enterprise sns. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 1067–1070, New York, NY, USA. ACM.

Kazutoshi Umemoto, Takehiro Yamamoto, and Katsumi Tanaka. 2016. Scentbar: A query suggestion interface visualizing the amount of missed relevant information for intrinsically diverse search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 405–414, New York, NY, USA. ACM.

Michael J. Welch, Junghoo Cho, and Christopher Olston. 2011. Search result diversity for informational queries. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 237–246, New York, NY, USA. ACM.

Semantic Interpretation of Events in Live Soccer Commentaries

Anne-Lyse Minard[1,2], Manuela Speranza[1], Bernardo Magnini[1], Mohammed R. H. Qwaider[1]
[1] Fondazione Bruno Kessler, Trento, Italy
[2] Dept. of Information Engineering, University of Brescia, Italy
{minard,manspera,magnini,qwaider}@fbk.eu

Abstract

English. In the context of semantic interpretation of live soccer commentaries in Italian, we propose an annotation schema for relevant events and their argument structure, on whose basis we annotated a reference evaluation corpus. We investigated automatic event classification and used Active Learning to reduce the cost of acquiring domain-specific training data.

Italiano. *Nel contesto dell'interpretazione di commenti calcistici in diretta, proponiamo uno schema per l'annotazione di eventi (e relativa struttura argomentativa), sulla base del quale abbiamo creato un corpus di valutazione di riferimento. Ci siamo occupati della classificazione automatica di eventi utilizzando Active Learning per ridurre lo sforzo per l'acquisizione di dati annotati specifici del dominio.*

1 Introduction

This work focuses on understanding the content of live commentaries of sport games. This form of written reporting has become very popular in recent years, and almost every national Italian online newspaper has a section dedicated to live sport commentaries. Live commentaries have several interesting properties: (i) they are short descriptions of an event written by professionals while the event is happening; their form is much simpler than a full spoken running commentary; (ii) they have a clear and simple structure, typically based on the timing of the sport event; (iii) they are often associated with metadata (e.g. *La Roma passa in vantaggio* [Roma takes the lead] is associated with the metadata GOAL); (iv) finally, they describe visual scenes, which is relevant to automatic alignment of multimedia content (e.g. align a sequence of frames in a video with the corresponding commentary), a topic of emerging interest in Computational Linguistics (see, for instance, (Song et al., 2016)). Our work is part of a larger cross-disciplinary project, Understanding Multimedia Content, currently involving several research groups at FBK.

In this paper we first define an annotation framework for the semantic interpretation of online soccer commentaries in Italian (Section 3), which includes the detection and classification of relevant events, as well as the identification of their argument structure. Based on this annotation schema, which could also be used for the annotation of tweets or other short online comments, we manually annotated a collection of commentaries in Italian to be used as a gold standard (Section 4). As a first step towards a comprehensive system for automatic interpretation of soccer events we focused on event detection and classification (i.e. event extraction), and used Active Learning to build a training corpus (Section 5). We show that this procedure is very effective, allowing our system to reach an F1 of 77.25, with considerable savings of annotation time (Section 6).

2 Related Work

Most of the work on event detection and classification focuses either on the news (UzZaman et al., 2012) or medical domains (Sun et al., 2013). For Italian, two corpora annotated with events following the It-TimeML framework (Caselli et al., 2011a) are available: EVENTI (Caselli et al., 2014) and WItaC (Speranza and Minard, 2015).

Event detection and classification on news has been of interest for English, Italian and Spanish in the TempEval evaluation campaigns (Verhagen et al., 2010; UzZaman et al., 2012) and for Italian in the EVENTI task at Evalita 2014 (Caselli et al., 2014). As part of these evaluation campaigns, several event extraction systems, mainly

supervised, have been implemented (Caselli et al., 2011b; Jung and Stent, 2013; Bethard, 2013; Mirza and Minard, 2014). The development of supervised systems requires a significant amount of training data, whose creation is very time consuming. The effort needed to annotate these data can be reduced by using Active Learning methods, i.e. methods where instances to be annotated are selected according to their predicted impact on the model learned for a specific task. Active Learning has been used in various linguistic annotation tasks, such as Named Entity Recognition (Shen et al., 2004) and Part-of-Speech tagging (Ringger et al., 2007).

The surging interest of the NLP community for event detection and classification in the sport domain, on the other hand, is shown by the hackathon recently organized on extraction of soccer events from Tweets in French, English and Arabic (`http://hackatal.github.io/2016/`).

Fort and Claveau (2012) present a corpus of match commentaries and transcripts of video commentaries of soccer games in French, which has been annotated with entities (e.g. players, referees), events (e.g. corner, penalty) and some relations (e.g. pass, replace player) and van Oorschot et al. (2012) propose a method to extract relevant events of games in Dutch using the quantity of tweets posted per minute.

Event extraction in the sport domain is even more important as far as analysis of video (Xu et al., 2008; Han et al., 2008) and audio (Cabasson and Divakaran, 2003) data is concerned.

In the domain of automatic alignment of multimedia content, the analysis of both texts, videos and audio is necessary, and the research focuses on the alignment of the events detected in the three media (Malmaud et al., 2015; Regneri et al., 2013).

3 Task Definition and Annotation Framework

In our annotation framework, semantic interpretation of soccer events consists of the following steps: (i) soccer event recognition and classification, (ii) recognition and classification of the entities involved in the soccer event, and (iii) identification of the argument relations between the soccer event and the participant entities.

3.1 Event Recognition and Classification

Soccer event annotation is inspired by the It-TimeML definition of event and follows its minimal chunk rule, according to which only the head of the event phrase is included in the annotated text span (Caselli et al., 2011a). The main difference with the It-TimeML framework is that we restrict it to verbal and nominal events and to a semantically defined set of relevant events.

In particular, we identified six semantic categories of events relevant to the soccer domain (and a number of sub-categories):

`Referee decision` includes events that are characterized as such due to a referee's intervention; examples of subcategories are `Yellow card` and `Offside`;

`Kick` includes events in which the ball is kicked by a player; examples of subcategories are `Penalty`, `Corner`, `Pass` (e.g. *apre* in (1)), `Shot on goal`, and `Free kick`;

`Interruption` includes events in which a player interrupts the action of the opposing team examples of subcategories are `Clearance` and `Intercept` (e.g. *anticipato* in (1));

`Possession` includes events where the ball, although moving, does not go from one player to another; as subcategories we find, for example, `Dribbling` and `Holding possession`;

`Goal` includes events where a team scores (we did not devise subcategories for `Goal`);

`No ball` includes (i) events where a player doesn't have the ball (e.g. *inserimento* in (1)), and (ii) events not involving the ball, such as pushing or knocking to the ground (no subcategorization).

(1) 71: Griezmann passa a Pogba che *apre* per Matuidi, *inserimento* in area del centrocampista del Psg, che viene *anticipato*. [Griezman for Pogba who in turn passes to Matuidi, the Psg midfield player makes a forward run for the ball but gets beaten to it]

3.2 Entity Recognition and Classification

In order to annotate entities relevant to the soccer domain, we identified four categories, i.e. `Player`, `Team`, `Referee`, and `Coach`. Entities include both named entities (e.g. *Griezmann* and *Psg* in (1)) and nominal entities (e.g. *centrocampista* [middle field player] in (1)) and textual span is identified according to the minimal chunk rule (as was done for events).

3.3 Argument Structure Identification

The annotation of the argument structure of an event is performed through the creation of links called ARG_rel between each event and its arguments (which can be either entities or events). Inspired by PropBank (Bonial et al., 2010), we also defined four numbered arguments to be assigned to each ARG_rel in the form of an attribute: ARG_0 and ARG_1 correspond to the required arguments of a predicate, e.g. agent and patient respectively, while ARG_2 and ARG_3 correspond to arguments that occur with high-frequency for a certain predicate.

In (1), for instance, we have an ARG_rel between *passa* and *Griezmann* (ARG_0) and an ARG_rel between *passa* and *Pogba* (ARG_2).

4 Reference Annotated Corpus for Event Interpretation

Based on the annotation schema described in Section 3, we manually annotated a corpus of nine soccer games (five games from the Euro 2016 competition and four games from *Campionato di Serie A* 2015-2016) collected from La Repubblica,[1] Tuttosport,[2] and Eurosport.[3] Annotation was performed using the CAT tool (Bartalesi Lenzi et al., 2012). The result is a reference corpus for the evaluation of semantic interpretation of soccer events consisting of around 13,500 tokens, for a total of 1,372 annotated events and 1,600 argument relations (see Table 1).

We computed the inter-annotator agreement (IAA) over 46 commentaries annotated by two annotators (two halves from two different games). In terms of Dice's coefficient (Dice, 1945) we obtained an IAA of 0.70 and 0.96 (micro average) for event and entity classification respectively, and 0.69 for relation recognition (between events and entities marked by both annotators).

5 Event Extraction

In order to extract and classify soccer events in online commentaries, we used a supervised machine learning approach. We had a system for event detection (trained on news articles annotated following It-TimeML) available, which did not perform well on the soccer domain (it obtained an F1 of 40.8 and recall of 50.1 on our reference corpus).

[1] http://www.repubblica.it/

[2] http://www.tuttosport.com/

[3] http://it.eurosport.com/

	ref. corpus	training corpus
Games	9	101
Commentaries	652	1,377
Commentaries/game	72	14
Tokens	13,567	31,955
Tokens/com.	20.8	23.2
Goal	66	168
Kick	666	1,425
Interruption	274	390
Possession	71	181
Referee decision	254	807
No Ball Event	41	181
Player	1317	-
Referee	21	-
Coach	10	-
Team	291	-
ARG_rel	1,600	-

Table 1: Dataset statistics.

As a consequence, a training corpus specifically developed for this task was needed.

We therefore exploited the TEXTPRO-AL Active Learning platform (Magnini et al., 2016) which selects the most informative samples from an unlabeled set. More precisely, TEXTPRO-AL selects commentaries containing events that the system was not able to recognize correctly, pre-annotates them and asks the annotator to check them.

As illustrated in Figure 1, an AL cycle consists of the following steps:[4]

1. Train a model using the annotated commentaries[5] (step 3);
2. Repeat the following cycle until the batch[6] is full:
   (a) Select, from an unlabeled database of commentaries (see Section 5.1), a commentary that matches the first event string in the error queue[7] (i.e. the event with the lowest confidence) (step 4);
   (b) Pre-annotate the example (step 5);
   (c) Correct the annotation (done manually by an annotator) (step 1);
   (d) Add the annotated example to the batch (step 2a);

[4] The AL cycle is repeated until a stopping criteria is verified; for instance, until the system reaches a pre-defined performance.

[5] At the beginning the training corpus is empty, so the first commentary is randomly selected and added to the batch.

[6] The batch size was set to 2 for the first 24 examples and then to 10. These values were chosen to enable frequent retraining of the model and an update of the confidence scores and system errors.

[7] The error queue (or system global memory) contains the history of the system errors corrected by the annotator.

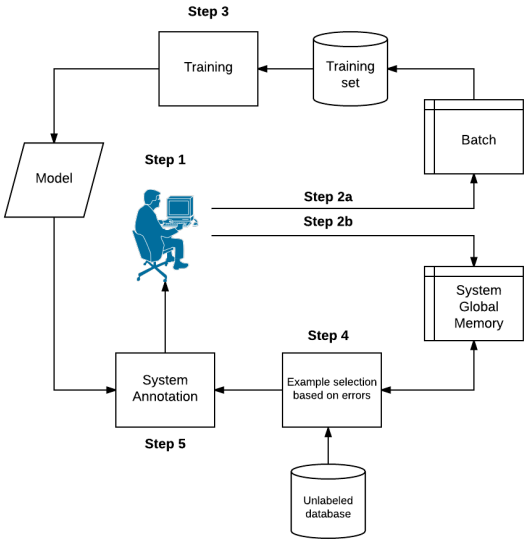


Figure 1: Active Learning schema adopted to build the training corpus.

(e) Save in the error queue the annotated events with their model confidence score (step 2b);

Our system is highly customizable: the event detection classification system can easily be substituted by a different system for different classification tasks, like NER and PoS tagging.

5.1 Unlabeled Database

The unlabeled database used in the AL procedure is composed of commentaries of 101 soccer games from DirettaGoal,[8] La Repubblica,[9] Tuttosport,[10] and Eurosport.[11] We extracted the online commentaries of all games of the Euro 2016 Cup and of the final 6 rounds of *Campionato di Serie A* 2015-2016. In total 6,573 commentaries were collected, with 155,005 tokens.

5.2 Error Selection

The error-based selection process exploits the idea that the corrections done by the annotator can be used to select new examples more efficiently. The system has a memory in which the events contained in the checked commentaries are stored, together with the system's confidence score and the indication of whether the system was right or wrong.

[8] http://www.direttagoal.it/
[9] http://www.repubblica.it/
[10] http://www.tuttosport.com/
[11] http://it.eurosport.com/

5.3 Event Detection and Classification

The system for event detection and classification is based on machine learning, using the SVM algorithm implemented in TinySVM and included in Yamcha (Kudo and Matsumoto, 2003). The task is treated as a multi-class classification task, where each token has to be classified in one of the 7 predefined classes.[12] The features used are those defined in the system of Mirza and Minard (2014), which took part in the EVENTI task at Evalita 2014 (Caselli et al., 2014), obtaining an F1 of 0.86 for the task of event detection and an F1 of 0.67 for event classification.

5.4 Annotation Editor

For the manual revision of linguistic annotations within the Active Learning method, we adapted an existing editor, MTEqual[13] (Girardi et al., 2014), originally developed for assessing the quality of machine translation.

6 Evaluation

The AL system described in the previous section was used by a non-expert annotator who annotated events in soccer commentaries for seven working days. This resulted in a training corpus of 1,377 commentaries, that is, around 200 commentaries per day (see Table 1).

The evaluation of our system was performed by comparing it to the reference annotated corpus described in Section 4. The learning curve in Figure 2 represents the results obtained by the system in terms of precision, recall and F1-measure as the training set was progressively extended. At the beginning the training set was empty, so the performance of the system was null. After the annotation of 200 commentaries, the system reached 53.27 F1, and after 800 commentaries it obtained 70.94 F1. At the end of our experiment, almost 1,400 commentaries had been annotated and the system's performance was 76.65 F1 (73.42 of recall and 80.16 of precision). The peak performance is 77.25 F1 and was reached with 1,347 commentaries (i.e. almost 32,000 tokens).

7 Conclusion and Future Work

We presented a new annotation framework for the interpretation of online soccer commentaries, as

[12] Referee decision, Kick, Interruption, Possession, Goal, No Ball Event and O for tokens that are not part of an event.
[13] https://github.com/hltfbk/MT-EQuAl

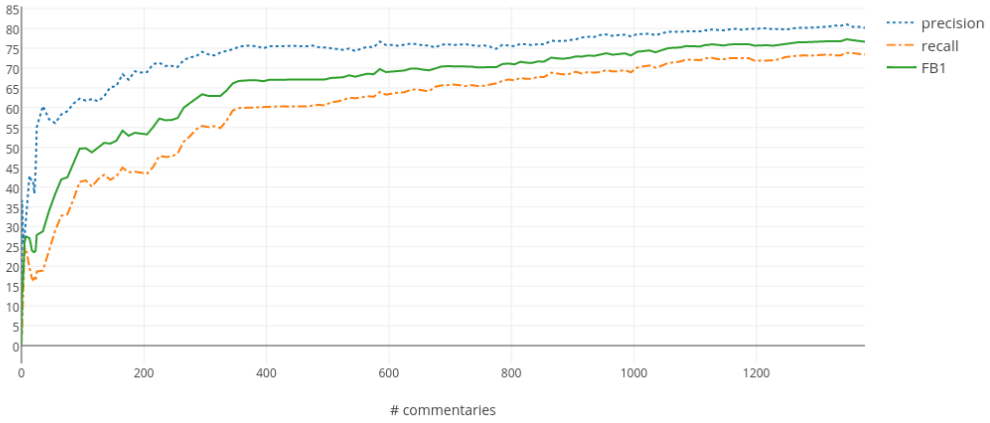


Figure 2: Event extraction performance as the training set was extended.

well as the reference annotated corpus we created.[14] We also described our system for event extraction from live soccer commentaries in Italian. It exploits the TEXTPRO-AL Active Learning platform, which allowed us to reach a significant F1 (77.25) in seven working days of a non-expert annotator. The annotation was performed for Italian but the method and the annotation schema we devised can be applied to other languages. The only language dependent component is the feature extractor used by the event detection module.

As for ongoing work, we are working at parameter optimization on the Active Learning framework (particularly, we are interested in the relations between the size of the unlabeled dataset, the frequency of the re-training, and the confidence score used by the selection procedure). We also plan to extend the current system by adding the detection of the argument structure of events.

Acknowledgments

We thank Valentino Frasnelli for his contribution, which consisted of manually annotating the data. This work has been partially funded by the Euclip project, in collaboration with Euregio.[15]

References

Valentina Bartalesi Lenzi, Giovanni Moretti, and Rachele Sprugnoli. 2012. CAT: the CELCT Annotation Tool. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 333–338, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Steven Bethard. 2013. Cleartk-timeml: A minimalist approach to tempeval 2013. In *Proceedings of the Seventh International Workshop on Semantic Evaluation*, SemEval '13, Atlanta, Georgia, USA.

Claire Bonial, Olga Babko-Malaya, Jinho D. Choi, Jena Hwang, and Martha Palmer. 2010. Propbank annotation guidelines, version 3.0. Technical report, Center for Computational Language and Education Research, Institute of Cognitive Science, University of Colorado at Boulder. `http://clear.colorado.edu/compsem/documents/propbank_guidelines.pdf`.

Romain Cabasson and Ajay Divakaran. 2003. Automatic extraction of soccer video highlights using a combination of motion and audio features. In *Storage and Retrieval for Media Databases 2003, Santa Clara, CA, USA, January 22, 2003*, pages 272–276.

Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. 2011a. Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank. In *Linguistic Annotation Workshop*, pages 143–151.

Tommaso Caselli, Hector Llorens, Borja Navarro-Colorado, and Estela Saquete. 2011b. Data-driven approach using semantics for recognizing and classifying timeml events in italian. In *Recent Advances in Natural Language Processing, RANLP 2011, 12-14 September, 2011, Hissar, Bulgaria*, pages 533–538.

Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. EVENTI

[14] Currently the annotated data are not be distributed due to copyright issues.

[15] `http://www.euregio.it`

EValuation of Events and Temporal INformation at Evalita 2014. In *Proceedings of the Fourth International Workshop EVALITA 2014*.

Lee Raymond Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, July.

Karën Fort and Vincent Claveau. 2012. Annotating football matches: Influence of the source medium on manual annotation. In *Proceedings of LREC 2012*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Christian Girardi, Luisa Bentivogli, Mohammad Amin Farajian, and Marcello Federico. 2014. Mt-equal: a toolkit for human assessment of machine translation output. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference System Demonstrations, August 23-29, 2014, Dublin, Ireland*, pages 120–123.

Yina Han, Guizhong Liu, and Gérard Chollet. 2008. Goal event detection in broadcast soccer videos by combining heuristic rules with unsupervised fuzzy c-means algorithm. In *Proceedings of ICARCV 2008, Hanoi, Vietnam, 17-20 December 2008, Proceedings*, pages 888–891.

Hyuckchul Jung and Amanda Stent. 2013. Att1: Temporal annotation using big windows and rich syntactic and semantic features. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of SemEval 2013*, pages 20–24, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Taku Kudo and Yuji Matsumoto. 2003. Fast Methods for Kernel-based Text Analysis. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 24–31, Stroudsburg, PA, USA.

Bernardo Magnini, Anne-Lyse Minard, Mohammed R. H. Qwaider, and Manuela Speranza. 2016. TEXTPRO-AL: An Active Learning Platform for Flexible and Efficient Production of Training Data for NLP Tasks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*.

Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. 2015. What's cookin'? interpreting cooking videos using text, speech and vision. *CoRR*, abs/1503.01558.

Paramita Mirza and Anne-Lyse Minard. 2014. FBK-HLT-time: a complete Italian Temporal Processing system for EVENTI-EVALITA 2014. In *Proceedings of the Fourth International Workshop EVALITA 2014*.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *TACL*, 1:25–36.

Eric Ringger, Peter McClanahan, Robbie Haertel, George Busby, Marc Carmen, James Carroll, Kevin Seppi, and Deryle Lonsdale. 2007. Active learning for part-of-speech tagging: Accelerating corpus annotation. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, pages 101–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Young Chol Song, Iftekhar Naim, Abdullah Al Mamun, Kaustubh Kulkarni, Parag Singla, Jiebo Luo, Daniel Gildea, and Henry A. Kautz. 2016. Unsupervised alignment of actions in video with text descriptions. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 2025–2031.

Manuela Speranza and Anne-Lyse Minard. 2015. Cross-language projection of multilayer semantic annotation in the NewsReader Wikinews Italian Corpus (WItaC). In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*.

Weiyi Sun, Anna Rumshisky, and Özlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *JAMIA*, 20(5):806–813.

Naushad UzZaman, Hector Llorens, James F. Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2012. Tempeval-3: Evaluating events, time expressions, and temporal relations. *CoRR*, abs/1206.5333.

Guido van Oorschot, Marieke van Erp, and Chris Dijkshoorn. 2012. Automatic extraction of soccer game events from twitter. In *Proceedings of the Workhop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2012)*, volume 902, pages 21–30, Boston, USA, 11.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 57–62, Stroudsburg, PA, USA. Association for Computational Linguistics.

Changsheng Xu, Yifan Zhang, Guangyu Zhu, Yong Rui, Hanqing Lu, and Qingming Huang. 2008. Using webcast text for semantic event detection in broadcast sports video. *IEEE Trans. Multimedia*, 10(7):1342–1355.

Linking IMAGACT ontology to BabelNet through action videos

Lorenzo Gregori
University of Florence
lorenzo.gregori@unifi.it

Alessandro Panunzi
University of Florence
alessandro.panunzi@unifi.it

Andrea Amelio Ravelli
University of Florence
aramelior@gmail.com

Abstract

English. Herein we present a study dealing with the linking of two multilingual and multimedia resources, BabelNet and IMAGACT, which seeks to connect videos contained in the IMAGACT Ontology of Actions with related action concepts in BabelNet. The linking is based on a machine learning algorithm that exploits the lexical information of the two resources. The algorithm has been firstly trained and tested on a manually annotated dataset and then it was run on all the data, allowing to connect 773 IMAGACT action videos with 517 BabelNet synsets. This linkage aims to enrich BabelNet verbal entries with a visual representations and to connect the IMAGACT ontology to the huge BabelNet semantic network.

Italiano. *In questo articolo si presenta uno studio sul linking tra due risorse linguistiche multilingui e multimediali, BabelNet e IMAGACT. L'esperimento ha l'obiettivo di collegare i video dell'ontologia dell'azione IMAGACT con i concetti azionali contenuti in BabelNet. Il collegamento è realizzato attraverso un algoritmo di Machine Learning che sfrutta l'informazione lessicale delle due risorse. L'algoritmo è stato addestrato e valutato su un dataset annotato manualmente e poi eseguito sull'insieme totale dei dati, permettendo di collegare 773 video di IMAGACT con 517 synset di BabelNet. Questo linking ha lo scopo di arricchire le entrate verbali di BabelNet con una rappresentazione visuale e di collegare IMAGACT alla rete semantica di BabelNet.*

1 Introduction[1]

Ontologies are widely used to represent language resources on the web, allowing them to be easily accessed and exploited by machines. For this reason, data interconnection between different semantic resources is a crucial task in order to enhance disambiguation and information retrieval capabilities in Artificial Intelligence, as evidenced by the increasing research into mapping and linking techniques among ontologies (Otero-Cerdeira et al., 2015). Nevertheless, ontology mapping has to face the problem of concept representation mismatch between resources, due to different building criteria and purposes (Siemoneit et al., 2015). Instance matching techniques play an important role in this context, allowing to connect entities from heterogeneous data resources which refer to the same real-world object (Castano et al., 2008; Nath et al., 2014).

Aside the general interest for knowledge bases interconnection in a web-based perspective, there is also a growing interest in multimodal resources, which combine textual and visual data. These resources can be exploited by intelligent algorithms integrating vision and natural language processing techinques[2]. This integrated approach was successfully applied for some challenging tasks involving verbs and their action reference as a video. Regneri et al. (2013) developed machine learning models for the automatic identification of similarity among actions, by using a corpus of natural language descriptions, derived from the videos of the *MPII Cooking Composite Activities* dataset, which represents actions involved in basic cooking tasks. Instead, the algorithm developed by Mathe

[1] Lorenzo Gregori developed the linking algorithm and wrote sections 3, 4, and 5; Andrea Amelio Ravelli performed the data annotation and wrote sections 1 and 2; Alessandro Panunzi supervised the research work and revised the paper.

[2] Several works in this field have been developed within The European Network on Integrating Vision and Language (iV&L Net), http://ivl-net.eu/

et al. (2008) extracts higher level semantic features in common among a sample set of verbs, using a fine-grained analysis of the represented action concepts, intended as a subsequent stable set of abstract features of the objects involved in the videos. Within this interdisciplinary perspective, a knowledge base which relates verbal lemmas in different languages with video prototypes can help in serveral applications, and be exploited by both humans and machines.

2 Resources

This paper presents a linking between BabelNet (Navigli and Ponzetto, 2012a) and IMAGACT (Moneglia et al., 2014a), two multilanguage and multimedia resources suitable for automatic translation and disambiguation tasks (Russo et al., 2013; Moneglia, 2014; Moro and Navigli, 2015).

2.1 BabelNet

BabelNet[3] is a multilingual semantic network created from the mapping together of the WordNet thesaurus and the Wikipedia enciclopedia. At present, BabelNet 3.7 contains 271 languages and it is the widest multilingual resources available for semantic disambiguation. Concepts and entities are represented by BabelSynsets (BS), extensions of WordNet synsets: a BS is a unitary concept identified by several kinds of informations (semantic features, glosses, usage examples, etc.) and related to lemmas (in any language) which have a sense matching with that concept. BSs are not isolated, but connected together by semantic relations. Moreover, BabelNet received a large contributions from its mapping with other resources such as ImageNet, GeoNames, OmegaWiki (along with many others), which increased its information beyond the lexicon and produced a wide-ranging, multimedia knowledge base.

2.2 IMAGACT

IMAGACT[4] is a visual ontology of action that provides a video-based translation and disambiguation framework for general verbs. The database evolves continously (Moneglia et al., 2014b) and at present contains 9 fully-mapped languages and 13 which are underway. The resource is built on an ontology containing a fine-grained categorization of action concepts, each represented by one or more video prototypes as recorded scenes and 3D animations. IMAGACT currently contains 1,010 scenes which encompass the action concepts most commonly referred to in everyday language usage[5]. The links between verbs and video scenes are based on the coreferentiality of different verbs with respect to the action expressed by a scene (i.e. different verbs can describe the same action, visualised in the scene). The visual representations convey the action information in a cross-linguistic environment and IMAGACT may thus be exploited for reference disambiguation in automatic and assisted translation tasks (Panunzi et al., 2014).

3 Related works

Other attempts have previously been made to link IMAGACT with other resources. Two experiments by De Felice et al. (2014) and by Bartolini et al. (2014) were conducted in an intralinguistic perspective: their aim was to evaluate the results of a mapping between the action concepts defined in ItalWordNet and the ones categorized by IMAGACT (in terms of perfect matches or hypernym/hyponym relations).

On the contrary, the objective behind our work is to obtain a light link between the resources by enriching the action concepts in BabelNet with a visual representation; in this way, we overpass the problem of finding a match between the generic semantic concepts in BabelNet and the specific pragmatic concepts in IMAGACT. This methodology is also enforced by the multilingual frame in which the experiment is conducted. As a matter of fact, the relation between words and concepts can deeply differ across languages, while the prototypical scenes ensure a language-independent modality which is able to keep together the different lexicalizations of the action space.

This work is a further step from a previous IMAGACT-BabelNet linking experiment (Gregori et al., 2015). Even if it was just a feasibility test to check the consistency of the linking, we reported good results in automatic assignment of IMAGACT prototypical scenes to BabelNet synsets. For this reason, we built a bigger dataset and we went from a metric-based to a Machine Learning algorithm to be run on the whole set of IMAGACT

[3] http://babelnet.org

[4] http://www.imagact.it

[5] The data is derived from the annotation of verb occurrences in spontaneous spoken corpora (Moneglia et al., 2012)

scenes.

4 Linking experiment

The aim of this experiment is to link the IMAGACT video scenes to the BabelNet interlinguistic concepts (BabelSynsets). In fact, the BabelNet objects are already enriched with visual objects, though this information contains static images which are inadequate for representing action concepts. In this way, adding video scenes to the verbs is very desirable and would suggest itself as a natural extension of BabelNet.

4.1 Training and test set

A manually annotated dataset of 50 scenes and 57 BabelSynsets (2,850 judgments) was created in order to test the algorithm and evaluate the results.

The sampling was carried on in two steps. First of all, a purely actional semantic area has been selected by taking BSs and scenes linked to 7 English action verbs, which are general and very frequent in the language use: *put*, *move*, *take*, *insert*, *press*, *give* and *strike*. The wide variation of these verbs allowed us to obtain a big set of concepts, with a high variation in terms of frequency and generality. On this set, a second sampling has been performed by preserving the variability in terms of number of connected verbs, that is a measurable parameter in both the resources.

Each ⟨BS,Scene⟩ pair has been evaluated to check if the scene is appropriate in representing the BS. Three annotators compiled the binary judgment table and we reported the values shared by at least 2 of 3. The measured Fleiss' kappa inter-rater agreement for this task was 0.74 [6].

Finally, the dataset has been split in a training set and a test set, with the proportions of 80% and 20% respectively (10 randomly chosen scenes for the test set and the remaining 40 scenes for the training set).

4.2 Algorithm

For this task, we developed a new algorithm which uses Machine Learning techniques, by exploiting the training set. As in the previous experiment, the features are extracted from the lexical items belonging to both the candidate BabelSynset and its neighbours[7]. Beside the algorithm, a baseline is determined by calculating the ratio $\frac{nsb}{nb+ns}$ for each pair and setting a threshold of 0.04, that maximizes the F-measure on our dataset.

Table 1 reports on the 17 languages common to both BabelNet and IMAGACT, detailing the relative number of verbs in each, and constitutes the quantitative data which the matching algorithms can exploit.

Language	BN Verbs	IM Verbs
English (EN)	29,738	1,299
Polish (PL)	9,660	1,193
Chinese (ZH)	9,507	1,171
Italian (IT)	7,184	1,100
Spanish (ES)	6,159	736
Russian (RU)	4,975	34
Portuguese (PT)	4,624	776
Arabic (AR)	3,738	804
German (DE)	3,754	992
Norwegian (NO)	1,729	115
Danish (DA)	1,685	646
Hebrew (HE)	1,647	160
Serbian (SR)	858	1,124
Hindi (HI)	831	466
Urdu (UR)	233	78
Sanskrit (SA)	33	276
Oriya (OR)	6	160
Total	86,361	11,130

Table 1: The 17 shared languages of BabelNet (BN) and IMAGACT (IM) with verbal lemma counts.

The basic features that we used for this experiment are:

- ns: the number of verbs connected to the Scene;
- nb: the number of verbs connected to the BS;
- nsb: the number of verbs that are shared between the Scene and the BS;

These 3 features have been calculated for each candidate BS and for the ones which are semantically related to it. We took the 8 BabelNet semantic relations available for verbs (see table 2) and for each BS we extracted 8 groups of related synsets, each one containing the set of BS connected to the main one by the same relation. Then, ns, nb and nsb are calculated for each group by summing the values of the BSs belonging to it.

[6] The manually annotated training set is published at `http://bit.ly/29J0ypx`

[7] This test is based on BabelNet 3.6; the data was extracted using the Java API (Navigli and Ponzetto, 2012b)

The feature set is comprised of 27 features: 3 features for the main BS and 3 features for each BabelNet relation. The set of candidates consists of all the possible BSs for each verb connected to the scene. A machine learning algorithm was trained on the annotated dataset: we used Support Vector Machine (SVM) classifier with a RBF kernel.

Table 2 shows the list of relations between the verbal BSs ranked by their relevance values for this task; this value is measured with Information Gain on the annotated dataset.

BabelNet relations	***IG* value**
Hyponym	0.057
Hypernym	0.026
Also See	0.019
Verb Group	0.019
Gloss Related	0.009
Entailment	0.003
Antonym	0.000
Cause	0.000

Table 2: Relations between verbal BSs.

4.3 Results

The algorithm was run on the training set and evaluated on the test set; the results are reported in Table 3.

	Baseline $th = 0.04$	**ML Algorithm** 27 features
Pr	0.580	0.833
Re	0.529	0.441
Fm	0.553	0.577

Table 3: Precision, Recall and F-measure of BSs to scenes linking task calculated on the test set for the algorithm and the baseline.

The results in terms of F-measure are not so satisfying and the value obtained with the algorithm is barely better than the baseline. Despite this, it's important to consider the difference with the baseline in terms of precision and recall, since precision is more important for this task: for this reason, the algorithm provides a much more reliable result compared to the baseline. We also have to point out that a low recall is mainly caused by multiple possiblities in the interpretation of a scene from different points of view: for example, the scene linked to the English verb *to throw* described by the sentence *John throws the ball to Mark* can represent not only a sense of *throw*, but also senses of other verbs, like *to play* or *to catch*, that refer to different semantic concepts; in these cases, the scene in IMAGACT is not linked to the alternative verbs, but it can be described with them (i.e. *John and Mark play with the ball*, *Mark catches the ball*). For this reason, the manual annotation provides more BS-to-scene relations than an algorithm can foresee on the basis of a lexical match, causing a low recall value.

Table 4 reports some statistics about the linking process; the entire results are browsable at the page `http://bit.ly/2a4FefT`.

IM Scenes linked to BS	773
BS linked to Scenes	517
IM English Verbs related to Scenes	544
BabelNet English Verbs related to BS	1,100

Table 4: IMAGACT-BabelNet linking numbers

Switching to Machine Learning had a strong impact on this linking task. The main advantage from the previous linking experiment (Gregori et al., 2015) is that now the number of BSs that can be assigned to each scene is variable, depending on the different reference possibilities that the BSs have. This is coherent with the BabelNet structure where we find very general concepts, that can be represented by several action prototypes, and specific ones, for which one prototype is enough to provide a clear representation.

For example the BS "bn:00090224v" (*Put into a certain place or abstract location*) expresses a general concept and is linked to 72 scenes, comprising the actions involving one or more objects or a body part, relating to different ways of *putting* (like *inserting*, *throwing*, *attaching*,...) or to different states of the Theme (e.g. solid or liquid). Conversely, the BS "bn:00084326v" (*Fasten with buttons*) is much more specific and is linked to only one scene (*c17d7346*) that represents a man that fastens his jacket.

5 Conclusions

The experiment described in this paper shows that is possible to obtain an extensive linking between IMAGACT and BabelNet through visual entities (see Figure 1 for a visual representation of a linking example); this can be advantageous for both the resources. BabelNet can add a clear video representation of the verbal synsets that refer to

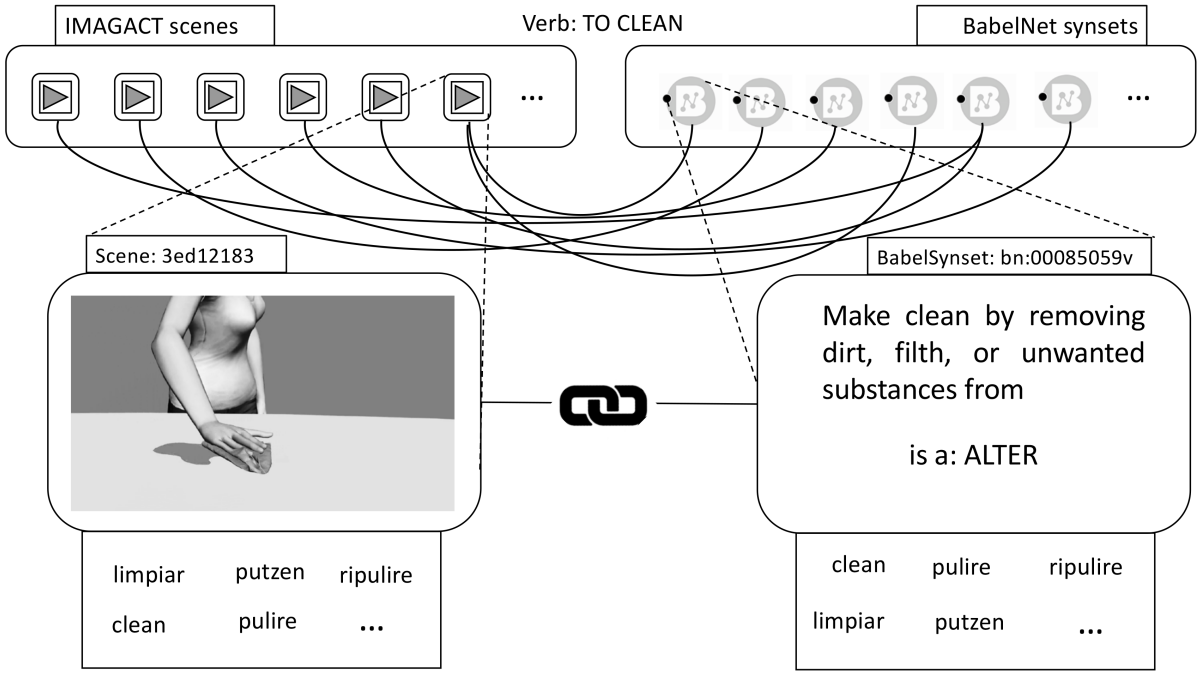


Figure 1: IMAGACT scene to BabelNet synset linking example

actions; IMAGACT can import verb translation candidates from many languages by exploiting the BabelNet semantic network; their integration can be exploited as a unified multimedial resource to accomplish complex tasks that combine natural language processing and computer vision.

Finally, we feel important to note that this procedure is scalable and the statistical model can be retrained at resource changes. This is a fundamental feature, especially considering the continuous update of IMAGACT languages and lemmas inventory.

Acknowledgments

This research has been supported by the MODELACT Project, funded by the Futuro in Ricerca 2012 programme (Project Code RBFR12C608); `http://modelact.lablita.it`.

References

[Bartolini et al.2014] Roberto Bartolini, Valeria Quochi, Irene De Felice, Irene Russo, and Monica Monachini. 2014. From synsets to videos: Enriching italwordnet multimodally. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

[Castano et al.2008] S. Castano, A. Ferrara, D. Lorusso, and S. Montanelli. 2008. On the ontology instance matching problem. In *Database and Expert Systems Application, 2008. DEXA '08. 19th International Workshop on*, pages 180–184, Sept.

[De Felice et al.2014] Irene De Felice, Roberto Bartolini, Irene Russo, Valeria Quochi, and Monica Monachini. 2014. Evaluating ImagAct-WordNet mapping for English and Italian through videos. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors, *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & the Fourth International Workshop EVALITA 2014*, volume I, pages 128–131. Pisa University Press.

[Gregori et al.2015] Lorenzo Gregori, Andrea Amelio Ravelli, and Alessandro Panunzi. 2015. Linking dei contenuti multimediali tra ontologie multilingui: i verbi di azione tra imagact e babelnet. In C. Bosco, F.M. Zanzotto, and S. Tonelli, editors, *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 150–154. Accademia University Press.

[Mathe et al.2008] S. Mathe, A. Fazly, S. Dickinson, and S. Stevenson. 2008. Learning the abstract motion semantics of verbs from captioned videos. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, pages 1–8, June.

[Moneglia et al.2012] Massimo Moneglia, Francesca Frontini, Gloria Gagliardi, Irene Russo, Alessandro

Panunzi, and Monica Monachini. 2012. Imagact: deriving an action ontology from spoken corpora. *Proceedings of the Eighth Joint ACL-ISO Workshop on Interoperable Semantic Annotation (isa-8)*, pages 42–47.

[Moneglia et al.2014a] Massimo Moneglia, Susan Brown, Francesca Frontini, Gloria Gagliardi, Fahad Khan, Monica Monachini, and Alessandro Panunzi. 2014a. The IMAGACT Visual Ontology. An Extendable Multilingual Infrastructure for the Representation of Lexical Encoding of Action. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

[Moneglia et al.2014b] Massimo Moneglia, Susan Brown, Aniruddha Kar, Anand Kumar, Atul Kumar Ojha, Heliana Mello, Niharika, Girish Nath Jha, Bhaskar Ray, and Annu Sharma. 2014b. Mapping Indian Languages onto the IMAGACT Visual Ontology of Action. In Girish Nath Jha, Kalika Bali, Sobha L, and Esha Banerjee, editors, *Proceedings of WILDRE2 - 2nd Workshop on Indian Language Data: Resources and Evaluation at LREC'14*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

[Moneglia2014] Massimo Moneglia. 2014. Natural Language Ontology of Action: A Gap with Huge Consequences for Natural Language Understanding and Machine Translation. In Zygmunt Vetulani and Joseph Mariani, editors, *Human Language Technology Challenges for Computer Science and Linguistics*, volume 8387 of *Lecture Notes in Computer Science*, pages 379–395. Springer International Publishing.

[Moro and Navigli2015] Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado, June. Association for Computational Linguistics.

[Nath et al.2014] Rudra Nath, Hanif Seddiqui, and Masaki Aono. 2014. An efficient and scalable approach for ontology instance matching. *Journal of Computers*, 9(8).

[Navigli and Ponzetto2012a] Roberto Navigli and Simone Paolo Ponzetto. 2012a. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

[Navigli and Ponzetto2012b] Roberto Navigli and Simone Paolo Ponzetto. 2012b. Multilingual WSD with just a few lines of code: the BabelNet API. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, Korea.

[Otero-Cerdeira et al.2015] Lorena Otero-Cerdeira, Francisco J. Rodrguez-Martnez, and Alma Gmez-Rodrguez. 2015. Ontology matching: A literature review. *Expert Systems with Applications*, 42(2):949 – 971.

[Panunzi et al.2014] Alessandro Panunzi, Irene De Felice, Lorenzo Gregori, Stefano Jacoviello, Monica Monachini, Massimo Moneglia, Valeria Quochi, and Irene Russo. 2014. Translating Action Verbs using a Dictionary of Images: the IMAGACT Ontology. In *XVI EURALEX International Congress: The User in Focus*, pages 1163–1170, Bolzano / Bozen, 7/2014. EURALEX 2014, EURALEX 2014.

[Regneri et al.2013] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

[Russo et al.2013] Irene Russo, Francesca Frontini, Irene De Felice, Fahad Khan, and Monica Monachini. 2013. Disambiguation of Basic Action Types through Nouns Telic Qualia. In Roser Saur, Nicoletta Calzolari, Chu-Ren Huang, Alessandro Lenci, Monica Monachini, and James Pustejovsky, editors, *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon. Generative Lexicon and Distributional Semantics*, pages 70–75.

[Siemoneit et al.2015] Benjamin Siemoneit, John Philip McCrae, and Philipp Cimiano. 2015. Linking four heterogeneous language resources as linked data. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 59–63, Beijing, China, July. Association for Computational Linguistics.

Relation mining from clinical records

Anita Alicante, Anna Corazza, Francesco Isgrò
Department of Electrical Engineering and Information Technologies (DIETI)
Università di Napoli Federico II
via Claudio 21, 80125 Napoli, Italy
{anita.alicante|anna.corazza|francesco.isgro}@unina.it

Stefano Silvestri
Institute for High Performance Computing and Networking, ICAR-CNR
via P. Castellino, 111, 80131 Napoli, Italy
stefano.silvestri@icar.cnr.it

Abstract

English. We propose a system to extract entities and relations from a set of clinical records in Italian based on two preceding works (Alicante et al., 2016b) and (Alicante et al., 2016a). This approach does not require annotated data and is based on existing domain lexical resources and unsupervised machine learning techniques.

Italiano. *Proponiamo un sistema per estrarre entità e relazioni da un insieme di cartelle cliniche in Italiano basato su due precedenti lavori (Alicante et al., 2016b) e (Alicante et al., 2016a). Questo approccio non richiede dati annotati e si basa su risorse lessicali di dominio già esistenti e tecniche di apprendimento automatico senza supervisione.*

1 Introduction

The digitization of medical documents in hospitals has produced plenty of information which should be adequately organized. While part of the material, mainly including international scientific publications, is in English, increasingly more material is being created in the language of the country of the medical institution. The main part of the local language material is represented by patient records. They contain important information not only for preparing care plans or solve problems for the particular patient, but also to extract statistics useful for research and also for logistics administration.

Automatic processing of such repositories still can not be straightforwardly applied. One of the principal issues to be solved is the automatic extraction of relevant information, usually consisting in entities and relations connecting them (Alicante et al., 2016b). In the cited work, we extensively discuss a domain entity and relation recognition system for Italian. Such step is at the basis of more sophisticated analyses, including semantics-based indexing of documents for improved retrieval, advanced query based information extraction, and the application of ontology-based strategies for privacy protection.

General tools, such as TextPro (Pianta et al., 2008), are not adapted for technical domains such as the medical one, as they are trained on generic documents, rather than domain-specific ones. Furthermore, a lot of tools are available for English and only a few of them have been ported to Italian. Another problem to take into account is the occurrence, in clinical records, of typos and nonstandard abbreviations, in addition to the most usual acronyms. Last but not least, passing from text to knowledge processing raises tricky privacy problems. In fact, especially but not only in small hospitals, obscuring the patient names is not sufficient to hide their identity as the medical information reported in records are often sufficient to reconstruct a precise profiling of the patients.

Therefore, *ad hoc* solutions represent the only way to build effective applications to solve this kind of problems. For example, not only domain entities and relations can help identifying potentially dangerous information, but also ontological information can be exploited to better protect patient privacy (Bonatti and Sauro, 2013). Again, ontologies construction and population are based on entity and relation extraction.

Efforts to port systems to languages different from English require, first of all, the development of lexical resources for the considered language. However, they are not sufficient, because of the intrinsic differences between languages. A widely

adopted way to tackle such difficulties is represented by machine learning approaches.

Although supervised approaches are usually more effective, they require large corpora of annotated data, which are quite expensive to obtain, as they require that domain experts invest time in a long and tedious annotation activity. In the medical domain, staff should invest part of their precious time to annotate data with information about the presence and the type of domain relevant entities and relations in records to be used for the training phase. Things would be much easier if domain experts are only required to check an automatically produced annotation. We therefore propose to integrate a knowledge-based and a text mining approaches to develop an application which requires the expert intervention only to check on medical and pharmaceutical labels associated to groups of relations.

More in detail, we propose here to integrate the systems discussed in (Alicante et al., 2016b) and in (Alicante et al., 2016a): the former adopts domain dependent lexical resources to extract entities and unsupervised machine learning approaches to decide where relations occur in the text. The latter clusters and labels the extracted relations with an approach based on lexical semantics.

The paper is organized with Section 2 detailing the approach implementation and Section 3 for conclusions and future works.

2 Proposed approach

The framework proposed is composed by three modules, and its logical structure is depicted in Figure 1. The first one is devoted to domain entity (i.e., medical and pharmaceutical entities) identification and classification, and exploits domain related lexical resources and standard natural language tools. The second one is based on an unsupervised machine learning approach, namely *clustering*, to avoid the necessity of annotating data, for the relation extraction. A potential relation is hypothesized among all pairs of the entities identified in the preceding phase. Clustering is then applied to group similar entity pairs. Small clusters indicate the lack of repetitive patterns and will therefore be considered as entity pairs which are not in relation to each other, while larger clusters are likely to correspond to different relation types.

Relations are clustered and labeled using the approach proposed in (Alicante et al., 2016a). The decision about how a relation can be labeled is only based on the terms involved in the corresponding entity pair, without considering the context in which it occurs. In fact, this is complementary with respect to the task of deciding whether two entities are related, which should be decided on the basis of the context where the two entities occur, as in (Alicante et al., 2016b). On the other hand, by considering only the two involved entities, we can only decide the *type* of a relation. Then, to decide whether the relation is stated or negated, also the context should be considered in the analysis.

The third module of the framework is based on Word Embeddings (WEs) (Mikolov et al., 2013) to represent the words involved in each entity with a real valued array. WEs most interesting characteristic consists in the fact that the mutual position of words in a metric space strongly depends on their meanings, so that words having similar semantics have large similarity, when this is computed, for example, by cosine similarity. Embeddings can be automatically built from a large collection of unannotated text with a very efficient algorithm. Therefore, they can be easily applied to any language, in our case to Italian, provided that enough texts are available. We used documents extracted from Wikipedia for training. In particular, we considered pages flagged as Medicine, Biology and Pharmacy in Italian. For the extraction, we used CatScan v3.0[1], Wikipedia Export tool[2] and Wikiextractor[3].

For each entity, we then consider the embeddings corresponding to each token. As shown in (Paperno and Baroni, 2016), a good representation for a string of words is given by the sum of the corresponding WEs. However, as we do not want that such representation depends on the string length, we normalize the sum by the number of words involved in the entity, obtaining the average or centroid of the corresponding WEs. Each pair of entities occurring in the same sentence represents a possible candidate for a relation. We therefore build the feature vector for each entity pair by juxtaposing the average vectors for each

[1] https://tools.wmflabs.org/catscan2/catscan2.php

[2] https://en.wikipedia.org/wiki/Special:Export

[3] medialab.di.unipi.it/Project/SemaWiki/Tools/WikiExtractor.py

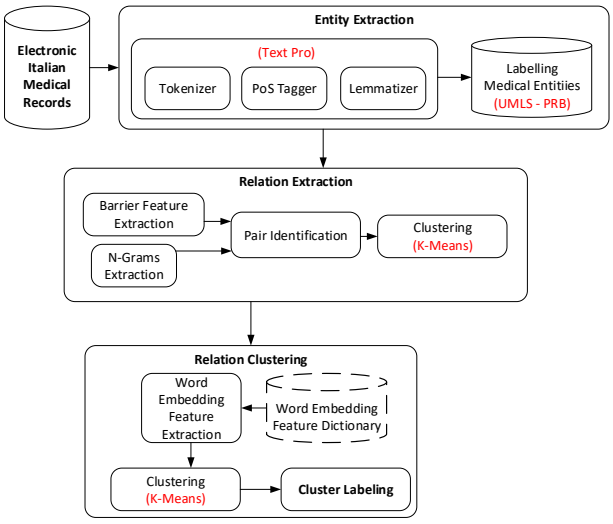


Figure 1: Architecture for Relation mining from clinical records.

entity and input this representation into a k-means clustering (Manning et al., 2008; Shalev-Shwartz and Ben-David, 2014).

2.1 Input Preprocessing

The text, processed by our system, is extracted from anonymized medical records, in the form of plain text encoded in UTF-8. The text includes a small set of special characters, used as delimiters and/or formatters. The largest part of these medical records has been produced by an HL7-compatible information system. At the end of each medical record, there is often an ICD9M (International Standard for Encoding and Classifying Diseases) disease code, which we disregard together with the rest of the structured part of the records.

The text is initially preprocessed for extracting textual parts from the medical records, and to get rid of non-textual characters. The plain text, produced by this preprocessing step, is passed to the natural language processing suite TextPro to perform tokenization, sentence splitting, PoS tagging and lemmatization.

2.2 Entity Extraction

Entity extraction is crucial for our analysis, and a specific module has been implemented with the goal of extracting entities which are relevant for the application domain: biomedical and pharmaceutical entities in our case. The module follows a pattern matching approach by identifying each occurrence of a number of PoS patterns in the input text as a candidate to be further analysed.

Afterwards, for each token occurring in the identified pattern, we search for matches of the corresponding lemma in the dictionaries. In case of multi-word expressions, when several patterns apply to overlapping strings of tokens, we apply a greedy approach by choosing the longest one matching the input.

The output is produced following the TextPro format, that is a line for each token, and a column for each analysis level. In our system these files are enriched by the information about Medical and Pharmaceutical entities obtained from the dictionaries provided by UMLS[4] and PRB[5]. These information are labeled as MED for the medical entities, and FAR for the pharmaceutical ones (the whole entity tag list is shown in the Table 1).

Table 1: List of medical sub-categories

Description	Label
Medical	MED
Pharmaceutical	FAR
Anatomy	ANA
Organisms	ORG
Diseases	MAL
Chemicals and Drugs	CHE
Technical medical equipment	TEC
Psychology and Psychiatric	PSI
Biology	BIO
Natural Sciences	NAT
Anthropology and Social Science	SOC
Technology, Industry and Agriculture	IND
Humanities	UMA
Computer Science	INF
Groups of People	GRU
Health care	ASS
Characteristics of Publication	PUB
Locations	LOC

In addition to a label indicating whether the entity is medical (MED) or pharmaceutical (FAR), we also add to each medical entity annotation the sub-categories included in the UMLS database in correspondence to the dictionary entry. The list of sub-categories labels are summarized in Table 1. A side-effect of such sub-categorization is that the number of potential relations increases while it becomes possible to find more specific relations.

[4]Unified Medical Language System, `http://www.nlm.nih.gov/research/umls`

[5]Pharmaceutical Reference Book, officially mantained by Agenzia Italiana del Farmaco

2.3 Relation Clustering

We apply the *k-means* approach that identifies groups of relations of the same type appearing in the data set. Each pair of entities occurring in the same sentence identifies a potential relation, therefore all possible entity pairs must be considered. We then apply a clustering algorithm to the set of all the potential relations identified. We will disregard all entity pairs belonging to clusters having a size smaller than a given threshold.

We then concentrate on the remaining entity pairs, which are likely to represent actual relations and semantically cluster them. The approach proposed for this is structured in three main modules: *Feature Construction*, *Clustering*, and *Cluster Labeling*. The first module builds a feature vector based on WEs for each relation candidate; for doing this, first it constructs a WE dictionary by using a large collection of unannotated texts, in our case extracted from Wikipedia. This module is based on *word2vec*[6] (Mikolov et al., 2013). For the feature vectors length we chose 500, which is the default choice, and set the minimum word count to 3, to exclude the less frequent words from the dictionary, obtaining a set of $260,680$ vectors.

After that, the k-means clustering is applied to the set of feature vectors obtained by the first module. For every entity pair we then construct a Feature Vector (FV) starting from the WE of each word involved. Each entity can be composed by one or more words, as for example *conati di vomito*: in this case, for each entity, we take the average among the WEs of the words composing the entity associated to the entity pair. Finally, we concatenate the FVs of the two entities, obtaining a FV of $1,000$ entries.

The clustering algorithm is then applied to the FV data set by means of the *C Clustering library* (de Hoon et al., 2004), a fast C implementation of the k-means algorithm. As the k-means is characterized by a random initial choice of the seeds, we repeated each run 10 times, always choosing the best solution. We considered the cosine similarity, choosing a number of clusters equal to 40, which seemed a reasonable choice given the results from the experiments in (Alicante et al., 2016b) and in (Alicante et al., 2016a).

Eventually, to label each cluster we ordered the pairs in each cluster according to its cosine similarity from the cluster centroid: the first four pairs are then chosen to characterize the cluster.

As discussed above, each FV can be partitioned in two parts: the first half corresponds to the first entity in the pair, the second one to the other. Such partition is consistently maintained during the whole processing. Also in the computation of centroids in the k-means clustering algorithm, the former half of each centroid derives from the average of the former half of the involved FVs and then corresponds to the first entity. Correspondingly, the latter half of each centroid vector only depends on the second entity of each involved pair.

The choice of the cluster to which a given item is assigned is based on the cosine similarity. Its computation can be divided in three parts: the dot product of the part of the two FVs corresponding to the first entity, the same for the second entity and eventually the normalization with respect to the whole FV. Therefore, the evaluation of the cosine similarity is based on a trade-off between how similar are the first and the second entities in each pair. In other words, they represent actual entities pairs which are similar to the (abstract) cluster representative, corresponding to the centroid.

3 Conclusions and future work

In this paper we presented a system for the extraction of information from clinical records in Italian. A first part of the system aims to extract domain relevant entities from medical reports by a pattern matching approach. A second part takes the output of the former step and applies a clustering approach to explore possible relations between such entities. A third part is based on WE and aims to give cues about the type of the relations.

Interestingly, the approach does not require annotated data, but only easily available data such as Wikipedia and off-the-shelf tools in addition to the documents to process. Naturally, available tools have been trained on annotated data, but without any adaptation to the specific domain. It would therefore be interesting to port it to a new language, possibly different from English, which represents the most widely studied among all languages.

Acknowledgments

The research presented in this paper was partially supported by the national projects CHIS - Cultural Heritage Information System (PON), and BIG4H

[6]The software is freely available at `https://code.google.com/p/word2vec/`

- Big Data Analytics for E-Health Applications (POR).

References

Anita Alicante, Anna Corazza, Francesco Isgrò, and Stefano Silvestri. 2016a. Semantic cluster labeling for medical relations. In *Proceeding of Innovation in Medicine and Healthcare 2016*, pages 183–193, Puerto de la Cruz, Tenerife, Spain. Springer.

Anita Alicante, Anna Corazza, Francesco Isgrò, and Stefano Silvestri. 2016b. Unsupervised entity and relation extraction from clinical records in Italian. *Computers in Biology and Medicine*, 72:263–275.

Piero A. Bonatti and Luigi Sauro. 2013. A confidentiality model for ontologies. In Harith Alani, Lalana Kagal, Achille Fokoue, Paul T. Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha F. Noy, Chris Welty, and Krzysztof Janowicz, editors, *International Semantic Web Conference (1)*, volume 8218 of *Lecture Notes in Computer Science*, pages 17–32. Springer.

Michiel J.L. de Hoon, Seiya Imoto, John Nolan, and Satoru Miyano. 2004. Open source clustering software. *Bioinformatics*, 20(9):1453–1454.

C.D. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proc. of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12.

Denis Paperno and Marco Baroni. 2016. When the Whole is Less than the Sum of its Parts: How Composition Affects PMI Values in Distributional Semantic Vectors. *Computational Linguistics*, 42(2):345–350.

Emanuele Pianta, Christian Girardi, and Roberto Zanoli. 2008. The TextPro Tool Suite. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 28–30, Marrakech, Morocco. European Language Resources Association (ELRA).

Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA.

Subjective Well-Being and Social Media: A Semantically Annotated Twitter Corpus on Fertility and Parenthood

Emilio Sulis, Cristina Bosco, Viviana Patti
Dipartimento di Informatica
University of Turin
Italy
{bosco,patti,sulis@di.unito.it}

Mirko Lai
Delia Irazú Hernández Farías
University of Turin, Italy
Univ. Politècnica de València, Spain
{hernande,lai}@unito.it

Letizia Mencarini
Dondena Centre for Research on
Social Dynamics and Public Policy
Bocconi University, Italy
letizia.mencarini@unibocconi.it

Michele Mozzachiodi
Daniele Vignoli
University of Florence, Italy
{mozzachiodi,vignoli}@disia.unifi.it

Abstract

English. This article describes a Twitter corpus of social media contents in the Subjective Well-Being domain. A multi-layered manual annotation for exploring attitudes on fertility and parenthood has been applied. The corpus was further analysed by using sentiment and emotion lexicons in order to highlight relationships between the use of affective language and specific sub-topics in the domain. This analysis is useful to identify features for the development of an automatic tool for sentiment-related classification tasks in this domain. The gold standard is available to the community.

Italiano. *L'articolo descrive la creazione di un corpus tratto da Twitter sui temi del Subjective Well-Being, fertilità e genitorialità. Un'analisi lessicale ha mostrato il legame tra l'uso di linguaggio affettivo e specifiche categorie di messaggi. Questo esame è utile per se e per l'addestramento di sistemi di classificazione automatica sul dominio. Il gold standard è disponibile su richiesta.*

1 Introduction

The key research questions we address in this paper concern how subjective well-being drives fertility trends (and vice versa). We developed a Twitter Italian corpus annotated with a novel semantic annotation scheme for marking information not only about sentiment polarity, but also about the specific semantic areas/sub-topics which are the target of sentiment in the fertility-SWB domain. The relationship between big data and official statistics is increasingly a subject of attention (Mitchell et al., 2013; Reimsbach-Kounatze, 2015; Sulis et al., 2015; Zagheni and Weber, 2005). In this work we focus on Twitter data for two main reasons. First, Twitter individuals' opinions are posted spontaneously (not responding to a question) and often as a reaction to some emotional driven observation. Moreover, using Twitter we can incorporate additional measures of attitudes towards children and parenthood, with a wider geographical coverage than what is the case for traditional survey. Sentiment analysis in Twitter has been also used to monitor political sentiment (Tumasjan et al., 2010), to extract critical information during times of mass emergency (Verma et al., 2011; Buscaldi and Hernández Farías, 2015), or to analyse user stance in political debates on controversial topics (Stranisci et al., 2016; Bosco et al., 2016; Mohammad et al., 2015). A comprehensive overview of sentiment analysis with annotated corpora is offered in (Nissim and Patti, 2016). Focusing on Italian, among the existing resources we mention the Senti-TUT corpus (Bosco et al., 2013) and the TWITA corpus (Basile and Nissim, 2013) that were recently exploited in the SENTIment POLarity Classification (SENTIPOLC) shared task (Basile et al., 2014). The corpus described in this paper enriches the scenario of datasets available for Italian, enabling also a finer grained analysis of sentiment related phenomena in a novel domain related to parenthood and fertility.

2 Dataset and Methodology

As a reference dataset, we adopted all the tweets posted in Italian language in 2014 (TWITA14

henceforth), which were retrieved through the Twitter Streaming API and applying the Italian filter proposed within the TWITA project (Basile and Nissim, 2013). The TWITA14 dataset included 259,893,081 tweets (4,766,342 geotagged). We applied a multi-step methodology in order to filter and select those relevant tweets concerning fertility and parenthood.

2.1 Filtering steps on the dataset

A number of filtering steps have been applied for selecting from TWITA14 a corpus of tweets where users talk about fertility and parenthood (TW-SWELLFER corpus, henceforth). We could not rely on the exploitation of one or few hashtags or other elements that allow identifying posts on fertility and parenthood. In fact, these topics are somehow spread in the dataset and messages may contain relevant information on such subjects even if the main topic of the post is different. Therefore, we are facing a situation where, on the one hand, the set of the data that are potentially relevant for our specific analysis is wider than usual; on the other hand, it is more difficult to identify the presence of information related to the topics we are interested in. This leaded us to adopt a multi-step thematic filtering approach. In a first step (*Keyword-based filtering step*), eleven hashtags[1] and other 19 keywords have been chosen for selecting tweets of interest, including 8 roots to consider diminutives, singulars and plurals. This list is the result of a combination of a manual content analysis on 2,500 tweets sampled at completely random (taken as a starting point) and a linguistic analysis on synonyms. We obtain a total amount of 3.9 million tweets. A second filtering step consisted in removing noisy tweets from corpus (*User-based filtering step*), as the off-topic ones (messages not concerning individual expression on fertility and parenthood topics). Tweets posted by company/institutions/newspapers accounts have been deleted. Finally, duplicated tweets not marked as RT were deleted (*Duplicate-based filtering step*). The resulting TW-SWELLFER corpus consists of 2,760,416 tweets.

[1] #papà, #mamma, #babbo, #incinta, #primofiglio, #secondofiglio, #futuremamme, #maternità, #paternità, #allattamento, #gravidanza

2.2 Annotation scheme

Given the TW-SWELLFER dataset, we developed and applied an annotation model aimed at studying not only the sentiment expressed in the tweets, but also specific parenthood-related topics discussed in Twitter that are the target of the sentiment. To build our annotation model, we relied on a standard annotation scheme on sentiment polarity (POLARITY), by exploiting the same labels POS, NEG, NONE and MIXED provided the organizers of the shared task for sentiment analysis in Twitter for Italian (Basile et al., 2014). Also the presence/absence of irony has been marked in order to be able to reason on sentiment polarity also in case of use of figurative devices. Annotating the presence of ironic devices is a challenging task because the inferring process of this figure of speech does not always lie on semantic and syntactic elements of texts (Ghosh et al., 2015; Reyes et al., 2013; Hernández Farías et al., 2016), but often requires contextual knowledge (Wilson, 2006). In order to mark irony, we introduced two polarized ironic labels: HUMNEG, for ironic tweets with negative polarity, and HUMPOS for ironic tweets with positive polarity. Finally, a set of labels marks the specific semantic areas (or SUBTOPICS) of the tweets related to the parenthood domain. This part of the annotation scheme is very important since somehow provides us with a semantic grid in order to analyse which are the aspects of parenthood that are discussed on Twitter. For the annotation of sub-topics we considered 7 labels, suggested by a group three experts on the SWELLFER (subjective well-being and fertility) domain, after a manual analysis of a subset of the tweets:

- TOBEPA - *To be parents*. This tag is introduced to mark when the user generically comments about his status of parent.
- TOBESO - *To be sons*. This tag marks the sons point of view, i.e. on when the user is a son that comments on the parent-son relationship.
- DAILYLIFE - *Daily life*. This tag marks messages commenting on recurring situation in everyday life for what concerns the relationship between parents and children.
- JUDGOTHERPA - *Judgment over other parents behaviour*. The tag allows to mark

comments on educations of children, for instance comments of behaviours which does not seems to be appropri - ated for the parent role.

- FUTURE - *Children' future*. This tag is used for tweets where parents do express sentiments about the future of children.
- BECOMPA - *To become parents*. This tag is introduced to mark tweets where users speak about the prospect or fear of being parents.
- POL - *Political side*. This tag is introduced to mark tweets talking about laws having impact on being parents.

Finally, two additional tags (IN-TOPIC/OFF-TOPIC) have been added to allow annotators to mark if the tweet is relevant. The addition of this tag was necessary because of the noise still present in the dataset. Moreover, in this way, the manual annotation will produce also data to be used in order to create a supervised topic classifier from the whole TW-SWELLFER corpus. This opens the way to the exploitation of the corpus for a fine-grained sentiment analysis, by identifying different aspects and topics of the Twitter debate on parenthood and the sentiment expressed towards each aspect/topic.

2.3 Manual annotation

A random sample of 5,566 tweets from TW-SWELLFER has been collected. On this sample we applied crowdsourcing for manual annotation via the Crowdflower platform already used in literature (Nakov et al., 2016). We relied on CrowdFlower controls to exclude unreliable annotators and spammers based on hidden tests, which we created by developing a set of gold-standard test questions equipped with gold reasons[2]. The annotator's task was, first, to mark if the post is IN- or OFF-TOPIC (or unintelligible), and then to mark for IN-TOPIC posts, on the one hand, the polarity and presence of irony, on the other hand, the subtopics. Precise guidelines were provided to the annotators. Overall, for each tweet at least three independent annotations were provided[3]. In order to select the true label we used majority voting.

In-topic vs off-topic: manual annotation on this aspect resulted in 2,355 in-topic tweets (42.3%) and 3,136 off-topic (56.3%); the remaining 75 tweets were unknown or null (cases of disagreement). Thanks to the preliminary filtering steps, the proportion of in-topic tweets is pretty high compared to common results from different Twitter based content and opinion analysis (Ceron et al., 2014).

Polarity, irony, sub-topics (in-topic tweets): at the end of the manual annotation process we collected 1,545 labeled with the same tags for all the layers.

Notice that in the analysis in the next section will report results also on tweets labeled as IN-TOPIC after the manual annotation (2,355), but where annotators did not agree on the polarity, irony and subtopics labels. We refer to those tweets as NULL messages.

Summarizing, the TWSWELLFER-GOLD corpus includes 1,545 IN-TOPIC tweets labeled with the same tags for all the layers (POLARITY, IRONY and SUBTOPICS).

3 Analysis of the gold standard

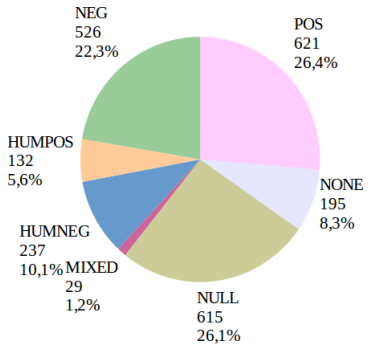


Figure 1: TW-SWELLFER: Distribution of polarity tags in IN-TOPIC messages

Regarding IN-TOPIC tweets, the 26.4% has been labeled as positive and 22.3% as negative (See Fig.1), giving us a guidance on what might be the general feeling in Twitter about the research topics on happiness and parenthood. The irony issue is limited to a 15.7% of all the messages and negative irony prevails (10.1% of negative ironic tweets and 5.6% of positive ironic tweets), while neutral tweets are just the 8.3%.

[2]Test questions resulted from the agreement of three expert annotators.

[3]We selected the CrowdFlower's *dynamic judgment* option: having the goal of collecting at least 3 reliable annotations for each tweet, the system was collecting up to a maximum of 5 annotations (to deal with cases when row's confidence score is low). In our jobs we set 0.7 as minimum accuracy threshold.

The amount of mixed tweets is limited to 1.2% (remaining 26% are labelled as NULL because of annotators disagreement). Regarding these results, it appears that positive and negative feelings towards family, parenthood and fertility appear more or less equally spread through Twitter Italy. Even if the positive posts are a little bit more than the negative ones, ironic tweets must be considered: most of them are negative ironic posts (i.e., insulting/damaging the target) balancing the slight difference between pure positive and negative tweets. Furthermore, this particular topic, combined with the Twitter nature which provides short direct message, discourages people to stand in the grey (neutral) area, as could happens in other cases: about the 90% of the tweets shows an explicit polarity, meaning people take a side and express their opinions.

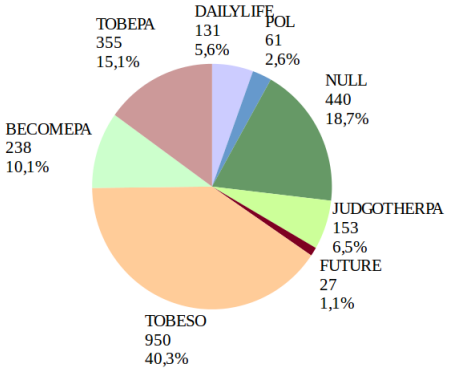


Figure 2: TW-SWELLFER: Distribution of subtopic tags in IN-TOPIC messages

Which are these opinions and about what? Going further with the analysis and looking also at the contents, so taking into consideration the "topic specification attribute and its values (Fig. 2), the largest category refers to sons tweets (TOBESO) (40.3%), in which children are discussing and posting about being children and/or about relating themselves with parents. Parents tag (TOBEPA) settles on 15% and becoming tag (BECOMEPA) on 10%. Remaining categories have minor impact, all being in between 1% and 6%.

3.1 Sentiment and emotion analysis

The exam of the corpus includes a lexical analysis on different aspects of affect: sentiment and emotions. The distribution of terms in each group of messages reveals interesting patterns. Adopting sentiment lexical resources[4] the whole polarity of messages is computed summing positive and negative terms. A normalization is finally performed, i.e. dividing the polarity value by the number of terms in each group. In particular, the four lexica considered count more positive terms in positive messages. Similarly, negative terms are more frequent in negative messages. Ironic messages reveal a similar pattern, even if smoothed. Table 1 presents some of these results.

In addition, the emotion lexicon indicates a larger frequency of terms related to anger, sadness, fear and disgust in negative messages than in positive ones (See Fig. 3). On the contrary, positive

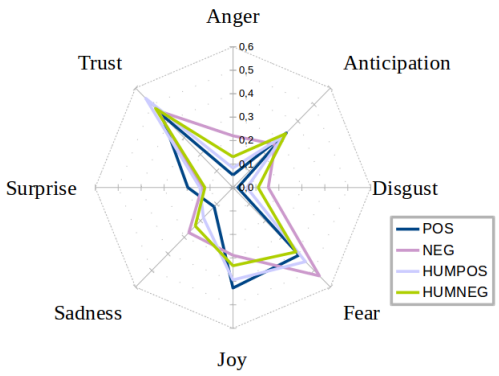


Figure 3: Distribution of emotions by polarity tags

messages contain more terms related to joy, anticipation and surprise. Some suggestions can be derived in the comparison of polarity categories and the corresponding ironic ones. For instance, terms related to joy are more frequent in ironic negative messages than in negative ones. It is an insight of the polarity reversal phenomena, where a shift is produced by the adoption of a seemingly positive statement, to reflect a negative one (Sulis et al., 2016).

The analysis of topic specification messages reveals a positive polarity for messages concerning TOBEPA (to be parents), while BECOMEPA (to become parents) has a more negative polarity (See Table 1). Focusing on emotion lexicon, TOBEPA has an higher incidence of Joy words (Fig. 4).

Messages concerning educations of children (JUDGOTHERPA) contain a high frequency of anger and disgust term. The category TOBESO (to

[4]EmoLex (Mohammad and Turney, 2013) as well as an own-house Italian version of LIWC (Pennebaker et al., 2001), Hu&Liu (Hu and Liu, 2004), AFINN (Nielsen, 2011). Lexicons were translated from English in (Buscaldi and Hernández Farías, 2015).

tag	pLIWC	pHuLiu	pEmolex	Afinn	pAVG
POS	1.06	0.22	0.62	3.51	1.35
NEG	-1.61	0.04	0.12	0.39	-0.27
HUMPOS	0.19	0.12	0.23	2.29	-0.71
HUMNEG	-0.34	0.08	0.64	0.61	0.25
TOBESO	1.97	0.88	0.02	1.56	1.11
TOBECOMEPA	1.5	0.73	0.18	-1.64	0.19
TOBEPA	1.94	1.38	0.18	5.04	2.13
DAILYLIFE	1.13	1.56	0.32	6.04	2.26

Table 1: Polarity values according to different lexicons in tweets tagged with the following labels: POS, NEG, HUMPOS, HUMNEG (polarity tags) and TOBESO, TOBECOMEPA,TOBEPA, DAILYLIFE (sub-topic tags).

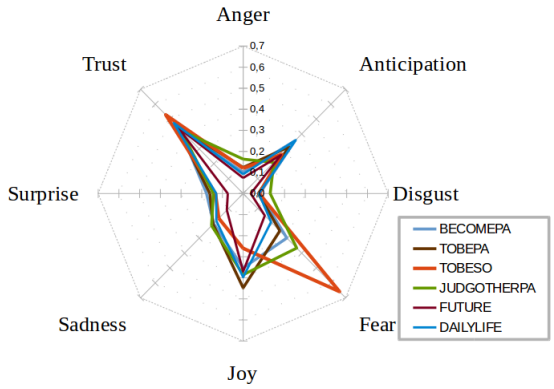


Figure 4: Distribution of emotions by sub-topic tags

be sons) is more controversial, having the higher frequency of negative terms as fear, but also trust, as well as having the lower frequency of Joy terms. Coherently, anticipation is more frequent in the BECOMEPA group of messages. Summarizing, it seems that children are more critics toward parents. On the contrary, parents seem express an attitude more positive towards children.

4 Conclusions and Future Work

The contribution of this paper is the exploration of opinions and semantic orientation about fertility and parenthood by scrutinizing about 3 million Italian tweets. This analysis is useful to identify features for the development of an automatic system to address automatic classification tasks in this domain. The corpus is available to the community. Its development constitutes a first step and a precondition to a further analysis that can be applied on such contents in order to extract, from semantically enriched data, measures of SWB constructed in an indirect way. This will hopefully improve our understanding of attitudes on fertility and parenthood.

We are currently extending the corpus by exploring the very interesting debate around the "Fertility Day's initiative" from the Italy's Minister of Health Beatrice Lorenzin, which had a remarkable echo on social media such as Twitter, with a substantial number of (also sarcastic) messages with hashtag #fertilityday posted.

Acknowledgments

The authors gratefully acknowledge financial support from the European Research Council under the European ERC Grant Agreement n. StG-313617 (SWELL-FER: Subjective Well-being and Fertility, P.I. Letizia Mencarini).

References

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta, Georgia. Association for Computational Linguistics.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proc. of EVALITA 2014*, pages 50–57, Pisa, Italy. Pisa University Press.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 28(2):55–63, March.

Cristina Bosco, Mirko Lai, Viviana Patti, and Daniela Virone. 2016. Tweeting and being ironic in the debate about a political reform: the french annotated corpus twitter-mariagepourtous. In *Proceedings of*

the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 1619–1626, Portoroz, Slovenia. ELRA.

Davide Buscaldi and Delia Irazú Hernández Farías. 2015. Sentiment analysis on microblogs for natural disasters management: A study on the 2014 genoa floodings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 1185–1188. ACM.

A. Ceron, L. Curini, and S.M. Iacus. 2014. *Social Media e Sentiment Analysis: L'evoluzione dei fenomeni sociali attraverso la Rete*. SxI - Springer for Innovation / SxI - Springer per l'Innovazione. Springer Milan.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, Antonio Reyes, and Jhon Barnden. 2015. Semeval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–475, Denver, Colorado, USA. Association for Computational Linguistics.

Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso. 2016. Irony detection in Twitter: The role of affective content. *ACM Transaction of Internet Technology*, 16(3):19:1–19:24.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, Seattle, WA, USA. ACM.

L. Mitchell, M. R. Frank, K. D. Harris, P. S. Dodds, and C. M. Danforth. 2013. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5), 05.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing and Management*, 51:480–499.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California, June. Association for Computational Linguistics.

Finn Årup Nielsen. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98, Heraklion, Crete, Greece. CEUR-WS.org.

Malvina Nissim and Viviana Patti. 2016. Semantic aspects in sentiment analysis. In Fersini Elisabetta, Bing Liu, Enza Messina, and Federico Pozzi, editors, *Sentiment Analysis in Social Networks*, chapter 3, pages 31–48. Elsevier.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. Linguistic Inquiry and Word Count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71.

Christian Reimsbach-Kounatze. 2015. The proliferation of big data and implications for official statistics and statistical agencies.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in twitter. *Lang. Resour. Eval.*, 47(1):239–268, March.

Marco Stranisci, Cristina Bosco, Delia Irazú Hernández Farías, and Viviana Patti. 2016. Annotating sentiment and irony in the online italian political debate on #labuonascuola. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Emilio Sulis, Mirko Lai, Manuela Vinai, and Manuela Sanguinetti. 2015. Exploring sentiment in social media and official statistics: a general framework. In *Proceedings of the 2nd International Workshop on Emotion and Sentiment in Social and Expressive Media, co-located with AAMAS 2015, Istanbul, Turkey, May 5, 2015.*, volume 1351 of *CEUR Workshop Proceedings*, pages 96–105. CEUR-WS.org.

Emilio Sulis, Irazú Hernández Farías, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo. 2016. Figurative messages and affect in twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*, 108:132 – 143. New Avenues in Knowledge Bases for Natural Language Processing.

Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *International AAAI Conference on Web and Social Media*.

Sudha Verma, Sarah Vieweg, William Corvey, Leysia Palen, James Martin, Martha Palmer, Aaron Schram, and Kenneth Anderson. 2011. Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency. In *International AAAI Conference on Web and Social Media*.

Deirdre Wilson. 2006. The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10):1722 – 1743.

Emilio Zagheni and Ingmar Weber. 2005. Demographic research with non-representative internet data. *International Journal of Manpower*, 36(1):13–25.

An extended version of the KoKo German L1 Learner corpus

Andrea Abel, Aivars Glaznieks, Lionel Nicolas, Egon Stemle
Institute for Specialised Communication and Multilingualism
EURAC Research
Bolzano/Bozen, Italy
andrea.abel@eurac.edu, aivars.glaznieks@eurac.edu
lionel.nicolas@eurac.edu, egon.stemle@eurac.edu

Abstract

English. This paper describes an extended version of the KoKo corpus (version KoKo4, Dec 2015), a corpus of written German L1 learner texts from three different German-speaking regions in three different countries. The KoKo corpus is richly annotated with learner language features on different linguistic levels such as errors or other linguistic characteristics that are not deficit-oriented, and is enriched with a wide range of metadata. This paper complements a previous publication (Abel et al., 2014a) and reports on new textual metadata and lexical annotations and on the methods adopted for their manual annotation and linguistic analyses. It also briefly introduces some linguistic findings that have been derived from the corpus.

Italiano. *Il contributo descrive una versione estesa del corpus KoKo (versione KoKo4, Dic 2015), corpus che raccoglie produzioni scritte di apprendenti di tedesco L1, provenienti da tre distinte regioni germanofone, a loro volta situate in tre diversi paesi. Il corpus KoKo è annotato dettagliatamente su differenti livelli linguistici rilevanti, quali gli errori o altre caratteristiche linguistiche non direttamente ricollegabili a deficit individuali, ed arricchito da un'ampia gamma di metadati. Questo contributo integra una precedente pubblicazione (Abel et al., 2014a) è informa sui nuovi metadati testuali e sulle nuove annotazioni lessicali cosi come sui metodi adottati per la loro annotazione manuale e per le loro analisi linguistiche. Inoltre presenta brevemente alcuni risultati ricavati dal corpus.*

1 Introduction

The study of linguistically annotated learner corpora has received a growing interest over the past 20 years (Granger et al., 2013). In learner corpus linguistics, such corpora are usually defined as "systematic computerized collections of texts produced by language learners" (Nesselhauf, 2005). Unlike most learner corpora focusing on L2/FL learners (i.e. learners learning a foreign language), the KoKo corpus focuses on advanced L1 speakers that are still learning their mother tongue, which typically happens in educational contexts.

This paper describes an extended version of the KoKo corpus (Abel et al., 2014a), a corpus created for the purposes of the KoKo project which aims at investigating the writing skills of German-speaking secondary school pupils. The creation of the corpus was guided by two goals: on the one hand to describe writing skills at the end of secondary school, on the other hand to consider external socio-linguistic factors (e.g. gender, socio-economic background etc.).

The previous description focused on the data collection, the data processing, the annotation of orthographic and grammatical features as well as on aspects regarding annotation quality (Abel et al., 2014a). This paper, however, introduces the new textual metadata and lexical annotations.

The paper is structured as follows. In section 2, key facts are briefly reported, including references to related work. The new textual metadata and lexical annotations are then described in section 3, alongside with the methods adopted for their manual annotation and linguistic analyses and some examples of linguistic findings. In section 4, future works are discussed right before concluding in section 5.

2 Key Information about the Corpus

The KoKo corpus is a collection of 1,503 authentic argumentative essays, and the corresponding survey information about their authors, produced in classrooms under standardized conditions by learners of 85 classes of 66 schools from three different German-speaking areas: South Tyrol in Italy, North Tyrol in Austria and Thuringia in Germany.[1] Such areas are particularly suitable for comparative studies because of differences regarding the German standard varieties, the use of dialectal vs. standard varieties and the monolingual vs. plurilingual environments (Abel et al., 2014a).

The corpus is roughly equally distributed over the three regions and amounts to 824,757 tokens (punctuation excluded). All writers were attending secondary schools one year before their school-leaving examinations. 83% of the pupils were native speakers of German. The corresponding L1 part of the corpus amounts to 726,247 tokens. Metadata annotations amount to 52,605 annotations whereas manual annotations amount to 117,422 annotations. Furthermore, 366 features to measure linguistic complexity[2](Hancke et al., 2012; Hancke and Meurers, 2013) were automatically calculated per text (550,098 in total) and added as metadata.

Previous evaluation showed high accuracy of manual transcriptions ($> 99\%$), and automatic tokenization ($> 99\%$), sentence splitting ($> 96\%$) and POS-tagging ($> 96\%$) (Glaznieks et al., 2014).

As it is among the first accessible richly linguistically annotated German L1 learner corpora, the KoKo corpus is particularly relevant to L1 learner language researchers, and for the field of didactics of German as L1. Other comparable language resources are either not accessible (Berg et al., 2010; DESI-Konsortium, 2006; Nussbaumer and Sieber, 1994), or although accessible, have not been enriched with linguistic information (Augst et al., 2007; Fix and Melenk, 2002) or are only partly annotated (Thelen, 2010). Some other corpora include L1 data, but as reference for L2/FL learner corpus research (Reznicek et al., 2010; Zinsmeister and Breckle, 2012).

[1] We followed the privacy policy for such surveys and requested a signed consent from all adult participants and parents of minors. In addition, all students participated anonymously, no names of the students were collected, names of schools were codified and made anonymous.

[2] e.g. syntactic features such as the average length of NPs, VPs and PPs as well as their number per sentence, morphological features such as the number of modal verbs per total number of verbs or the average compound depth of nouns, and lexical features such as lexical diversity described by means of different measures

3 New Metadata and Annotations

This section describes the main features of the latest corpus version KoKo4 (Dec. 2015) that have been added to the version KoKo3 (Dec. 2014). It thus focuses on a new set of textual metadata and a new layer of lexical annotations which is, due to the selected features and the degree of granularity, a novelty in (corpus-based) modeling of L1-writing competences for German .

3.1 Textual Metadata

In the KoKo corpus, two kinds of **Metadata information** are available: (1) non-linguistic, i.e. person-related information provided by each participant via a questionnaire survey in class that is available for the whole sample and (2) linguistic, i.e. text-related information provided for a subsample of the corpus (569 texts, equally distributed over the three regions involved) through an online evaluation form by three different specially trained raters originating from the different participating regions.

While type (1) metadata allow for sociolinguistic analyses in order to detect relations between linguistic features (e.g. text length, sentence length, orthographic errors, grammatical errors, etc.) and non-linguistic person-related information, type (2) metadata constitute a further expansion of our analysis by including textual features as well. Text analysis was done holistically using an evaluation form and detailed guidelines that were elaborated on the basis of recent findings in writing research and text analyses (Brinker, 2010; Feilke, 2010; Augst et al., 2007; Böttcher and Becker-Mrotzek, 2006; Jechle, 1992; Augst and Faigel, 1986) and the curricula in the participating regions. The text evaluation form distinguishes four categories : **(A) formal completeness, (B) content, (C) formal and linguistic means of text arrangement and (D) overall impression**.

For **category A**, 10 questions of the online evaluation form focused on the presence of obligatory text parts (introduction, main part, closing part) and explicitly requested constituents of argumentative essays (opinion of the author, conclusion). The 25 questions of **category B** belong to

two subcategories: (B1) the topics of the essay (9 questions), (B2) patterns of topic development (16 questions). B1 comprises evaluations on e.g. the topics of each text part, gaps, and the overall coherence of the text. B2 refers to the main pattern of topic development (argumentative, etc.), the argumentation strategies (point of view, concessive or not), and the motivation of arguments (objective vs. subjective stance, quality of arguments). Formal and linguistic means of text arrangement (**category C**, 7 questions) focus on the use of paragraphs, the explicit announcement of and commitment to the function of the essay, and the use of linguistic means to structure the text with regards to content. Finally, **category D** (20 questions) aims for an overall impression and therefore focuses on the completion of the task (successful or not), the overall quality of the text and the overall consistency of both the quality and coherence. Of all 62 questions of the entire online evaluation form, we used 57 for each document of the subcorpus (alltogether 33,972 annotations).

The analyses revealed, among other things, that the text quality is classified as quite satisfactory on a 5 point Likert-scale[3]. More specifically, there are significant correlations between text quality assessment and other linguistic variables: thus, a lower number of e.g. lexical errors is connected to a higher text quality score[4], and, finally, a variety of group differences could be detected (e.g. concerning school type: lower text quality scores within vocational schools compared to general high schools[5]).

3.2 Lexical Annotations

As for the **manual annotations** of orthographic and grammatical features added to previous corpus versions (Abel et al., 2014a), a specifically crafted tag set and annotation manual were used for the annotation of lexical features. 61,728 lexical annotations were manually performed by trained annotators on a subcorpus of 980 texts, almost equally distributed over the three regions.

The analyses of lexical features focuses on lexical knowledge as a central part of lexical competence which includes the dimensions of lexical breadth (quantitative aspect) and lexical depth (qualitative aspect) (Steinhoff, 2009; Böttcher and Becker-Mrotzek, 2006; Mukherjee, 2005; Read and Nation, 2004; Read, 2000; Nation, 2001). Whereas the analyses of quantitative aspects of lexical knowledge were performed automatically by using different measures (e.g. lexical diversity measures such as MTDL and Yule's K, or lexical frequency scores based on dlexDB (Hancke and Meurers, 2013)), the analyses of qualitative aspects were done by means of manual annotations. We focus hereafter exclusively on the manual annotations allowing us to model qualitative aspects of lexical knowledge.

Category	Sub-category	Total
Single words	Neol. & occas.	4,670
	Arg. adv. & conj.	14,345
Phrasemes	Referential	18,708
	Communicative	4,824
	Structural	2,704
Particularities	Semantic	8,397
	Stylistic	236
	Form	1,923
	Metalinguistic	1,412
Target hyp.		4,509

Table 1: Quantitative figures for the 980 documents annotated with the new lexical annotations.

For annotating lexical features, we developed a new hierarchically-structured linguistic classification scheme inspired by previous work that focused on L2 learner languages (Abel et al., 2014b; Konecny et al., 2016). The classification scheme takes both into account occurrences of selected lexical phenomena and defective as well as non-defective particularities of learner languages considering two dimensions: (1) the linguistic subcategory, e.g. collocations and idioms, and (2) a target modification classification, e.g. omission, addition (Díaz-Negrillo and Domínguez, 2006; Abel et al., 2014b). Furthermore, we formulated target hypotheses for those categories that we annotated as defective in order to make the error interpretation transparent (Lüdeling et al., 2005). The corresponding annotation scheme contains 77 different tags including a set of further attributes.

In a multi-stage annotation procedure, all occurrences of phenomena on both single words and formulaic sequences (FS) were annotated (Wray, 2005). Annotations for particularities were subsequently added in order to distinguish between er-

[3] percentages: 1 (scarse): 6.2 - 2: 22.9 - 3: 39.0 - 4: 26.3 - 5 (excellent): 5.7)

[4] Kruskal Wallis H Test: FS errors $X^2(1) = 10.417$, $p = .036$, single word errors: ANOVA $F(4, 338) = 2.805$, $p = .026$

[5] Kruskal Wallis H Test: $X^2(1) = 49.147$, $p = .000$

rors concerning correctness, errors concerning appropriateness of usage (Eisenberg, 2007; Schneider, 2013), non-defective modifications (to capture, for example, creative use of language), and diasystematic markedness. At the single word level, we considered all out-of-vocabulary tokens of the part-of-speech tagger (Schmid, 1994) as candidates of neologisms or occasionalisms. In addition, we captured a variety of tokens relevant for the text genre of an argumentative essay (i.e. argumentative adverbs and conjunctions). At the level of FS, we applied a function-based approach distinguishing between three main categories of phrasemes (Burger, 2007), each of them with further subcategories (Abel et al., 2014b; Konecny et al., 2016; Granger and Paquot, 2008; Burger, 2007; Stein, 2007; Steinhoff, 2007), as well as a "mixed classification" (Burger, 2007):

Referential phrasemes include collocations[6] and idioms[7], distinguished among other things with respect to their degree of idiomaticity. **Communicative phrasemes** are subdivided into those bound to specific situations[8], and those not bound to specific situations[9]. Finally, **structural phrasemes** comprise complex conjunctions and prepositions[10] and concessive constructions[11].

For particularities, we considered four main categories, each with further subcategories:

On a **semantic dimension** a distinction is made between denotative errors concerning correctness or appropriatness of use[12], and connotative markedness or appropriatness of use[13]. The **stytlistic dimension** considers repetition, and redundancy. The **form dimension** focusses on word formation errors (concerning single word units only[14]), and on omission, choice, position and addition errors as well as creative modifications (concerning FS). Concerning **metalinguistic markers** the appropriateness of the use of quotation marks for highlighting units is considered.

An overview of the number of annotations is provided in Table 1.

Results of the analyses showed, among others, that pupils use different types of FS quite frequently, on average 5.12 constructions per 100 words: with 62%, non idiomatic referential phrasemes constitute the major part, followed by idiomatic referential phrasemes (19%), and, finally, structural (10%) and communicative phrasemes (9%). However, lexical errors in general affect more often FS than single word units (10% of the FS vs. 1.04% of the single words). The latter are most frequently form errors (5.50% of FS affected, especially choice errors: 4.17%).

4 Future Work

The KoKo project was completed and presented to the public in December 2015. We will start releasing the data via the corpus exploration interface ANNIS3 (Krause and Zeldes, 2016) and for download on request, after signing a license agreement.[15] Aside from the aforementioned data, future versions will also include additional metadata information about the authors integrated for the purposes of future socio-linguistic analyses.

Consensus in the annotations among annotators, and as such an indication of its reliability, will be evaluated on sub-sets of texts that were annotated for this purpose by more than one annotator. Three annotators independently annotated the text level metadata annotations on 27 texts, and six annotators independently annotated the lexical level annotations on the same 27 texts. Inter-annotator agreement will be calculated for annotations and segmentation, i.e. the agreement on the decision which word sequence needs to be tagged vs. what annotation needs to be assigned to it, and will be evaluated and reported in the form of Fleiss Multi-k and boundary similarity (Artstein and Poesio, 2008; Fournier, 2013).

Finally, thanks to its relatively large size and its richly annotated nature, potential additional uses

[6] further divided into restricted and loose collocations, light verb constructions (called "Funktionsverbgefüge" in German) as well as special classes such as irreversible bi- and trinominals, similes etc.

[7] further divided into nominative idioms and fixed phrases, and special classes such as irreversible bi- and trinominals etc.

[8] further divided into general routine or speech act formulas, special classes such as commonplaces, slogans, proverbs etc., and empty formulas

[9] further divided into text organising formulas, and interaction organising formulas

[10] further divided into phraseological connectors and syntactically complex connectors, and secondary prepositions

[11] further divided into constructions with "although" and a correlate of the "but"-class, and constructions with a modal word and a correlate of the "but"-class

[12] further divided into reference/function, contextual fitness, semantic compatibility, and precision

[13] further divided into speaker's attitude, and diasystematic markedness concerning language usage, i.e. diaphasic markedness, diachronic markedness, diatopic markedness

[14] distinguishing between errors with respect to derivation and to composition

[15] We have been trying to make the data available for direct download – but have to take more legal hurdles.

of the KoKo corpus in Natural Language Processing and Corpus Linguistics are being considered. Regarding Natural Language Processing, the error annotations paired with target hypothesis annotations allow for creating an aligned corpus. Such corpora can be used to improve machine translation for automatically correcting learner texts (Ng et al., 2014). Regarding Corpus Linguistics, machine learning methods can be used (e.g. as being done in WebAnno (Yimam et al., 2014)) to drive linguistic intuitions when performing annotations or analyses. Because of the richness of its annotation schemes, the KoKo corpus constitutes a challenging but at the same time promising dataset to test if the developed methods are able to uncover relevant correlations that have already been investigated, or to uncover even new ones that are worth considering for future linguistic analyses.

5 Conclusion

This paper described the most recent version of the KoKo corpus, a collection of richly annotated German L1 learner texts, and focused on the new textual metadata and lexical annotations.

Because other comparable language resources are either not accessible, or have not been enriched with linguistic information or are only partly annotated, the corpus is a valuable resource for research on L1 learner language, in particular for the research on writing skills, and for teachers of German as L1, in particular for the teaching of L1 German writing skills.

References

Andrea Abel, Aivars Glaznieks, Lionel Nicolas, and Egon Stemle. 2014a. Koko: An L1 learner corpus for german. In *Proceedings of LREC 2014*, pages 2414–2421.

Andrea Abel, Katrin Wisniewski, Lionel Nicolas, and Detmar Meurers. 2014b. A trilingual learner corpus illustrating european reference levels. *RICOGNIZIONI— Rivista di Lingue, Letterature e Culture Moderne*, 1(2):111–126.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Gerhard Augst and Peter Faigel. 1986. *Von der Reihung zur Gestaltung: Untersuchungen zur Ontogenese der schriftsprachlichen Fähigkeiten von 13-23 Jahren*, volume 5 of *Theorie und Vermittlung der Sprache*. Peter Lang, Frankfurt.

Gerhard Augst, Katrin Disselhoff, Alexandra Henrich, Thorsten Pohl, and Paul Völzing. 2007. *Text-Sorten-Kompetenz. Eine echte Longitudinalstudie zur Entwicklung der Textkompetenz im Grundschulalter*. Peter Lang, Frankfurt.

Margit Berg, Anne Berkemeier, Reinold Funke, Christian Glück, Christiane Hofbauer, and Jordana Schneider, editors. 2010. *Sprachliche Heterogenität in der Sprachheil- und der Regelschule. Abschlussbericht im Programm „Bildungsforschung" der Landesstiftung Baden-Wrttemberg, Germany*.

Ingrid Böttcher and Michael Becker-Mrotzek. 2006. *Schreibkompetenz entwickeln und beurteilen*. Cornelsen, Berlin.

Klaus Brinker. 2010. *Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden. Bearbeitet von Sandra Ausborn-Brinker, 7., durchgesehene Auflage*, volume 29 of *Grundlagen der Germanistik*. Erich Schmidt Verlag, Berlin.

Harald Burger. 2007. *Phraseologie: Eine Einführung am Beispiel des Deutschen*, volume 36 of *Grundlagen der Germanistik*. Erich Schmidt Verlag, Berlin.

DESI-Konsortium, editor. 2006. *Unterricht und Kompetenzerwerb in Deutsch und Englisch*. Beltz Verlag, Weinheim – Bern.

Ana Díaz-Negrillo and Jesús Fernández Domínguez. 2006. Error tagging systems for learner corpora. *Revista española de lingüística aplicada*, (19):83–102.

Peter Eisenberg. 2007. Sprachliches Wissen im Wörterbuch der Zweifelsfälle. Über die Rekonstruktion einer Gebrauchsnorm. *Aptum. Zeitschrift für Sprachkritik und Sprachkultur*, 3(2007):209–228.

Helmuth Feilke. 2010. Schriftliches Argumentieren zwischen Nähe und Distanz am Beispiel wissenschaftlichen Schreibens. *Nähe und Distanz im Kontext variationslinguistischer Forschung*, pages 209–231.

Martin Fix and Hartmut Melenk. 2002. *Schreiben zu Texten-Schreiben zu Bildimpulsen: das Ludwigsburger Aufsatzkorpus; mit 2300 Schülertexten, Befragungsdaten und Bewertungen auf CD-ROM*. Schneider-Verlag, Hohengehren.

Chris Fournier. 2013. Evaluating Text Segmentation using Boundary Edit Distance. In *Proceedings of 51st Annual Meeting of the ACL*, pages 1702–1712. ACL.

Aivars Glaznieks, Lionel Nicolas, Egon Stemle, Andrea Abel, and Verena Lyding. 2014. Establishing a standardised procedure for building learner corpora. *Apples - Journal of Applied Language Studies*, 8(3):5–20.

Sylviane Granger and Magali Paquot. 2008. Disentangling the phraseological web. *Phraseology. An interdisciplinary perspective*, pages 27–50.

Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier. 2013. *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead: Proceedings of the First Learner Corpus Research Conference (LCR 2011)*.

Julia Hancke and Detmar Meurers. 2013. Exploring CEFR classification for German based on rich linguistic modeling. In *Proceedings of the Learner Corpus Research Conference (LCR 2013)*, pages 54–56.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for german using lexical, syntactic, and morphological features. In Martin Kay and Christian Boitet, editors, *Proceedings of COLING 2012*, pages 1063–1080, Mumbai.

Thomas Jechle. 1992. *Kommunikatives Schreiben: Prozess und Entwicklung aus der Sicht kognitiver Schreibforschung*, volume 41 of *ScriptOralia*. Gunter Narr Verlag, Tübingen.

Christine Konecny, Andrea Abel, Erica Autelli, and Lorenzo Zanasi. 2016. Identification and Classification of Phrasemes in an L2 Learner Corpus of Italian. In Gloria Corpas Pastor, editor, *Computerised and Corpus-based Approaches to Phraseology*, pages 533–542. Editions Tradulex, Geneva.

Thomas Krause and Amir Zeldes. 2016. ANNIS3: A New Architecture for Generic Corpus Query and Visualization. *Digital Scholarship in the Humanities*, 31(1):118–139.

Anke Lüdeling, Maik Walter, Emil Kroymann, and Peter Adolphs. 2005. Multi-level error annotation in learner corpora. In *Proceedings of Corpus Linguistics 2005*, pages 15–17.

Joybrato Mukherjee. 2005. The native speaker is alive and kicking: Linguistic and language-pedagogical perspectives. *Anglistik*, 16(2):7–23.

I.S.P. Nation. 2001. *Learning Vocabulary in Another Language*. Foreign Language Study. Cambridge University Press.

Nadja Nesselhauf. 2005. *Collocations in a Learner Corpus*, volume 14 of *Studies in Corpus Linguistics*. John Benjamins Publishing, Amsterdam.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. ACL.

Markus Nussbaumer and Peter Sieber. 1994. Texte analysieren mit dem Zürcher Textanalyseraster. In Peter Sieber, editor, *Sprachfähigkeiten–Besser als ihr Ruf und nötiger den je!*, pages 141–186. Verlag Sauerländer, Aarau.

John Read and Paul Nation. 2004. Measurement of formulaic sequences. In Norbert Schmitt, editor, *Formulaic sequences: Acquisition, processing and use*, Language Learning & Language Teaching, pages 23–35. John Benjamins Publishing, Amsterdam.

John Read. 2000. *Assessing vocabulary*. Cambridge University Press.

Marc Reznicek, Maik Walter, Karin Schmidt, Anke Lüdeling, Hagen Hirschmann, Cedric Krummes, and Torsten Andreas. 2010. Das Falko-Handbuch. Korpusaufbau und Annotationen. Technical report, Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Jan Georg Schneider. 2013. Sprachliche ‚Fehler' aus sprachwissenschaftlicher Sicht. In *Sprachreport*, volume 1-2/2013, pages 30–37. Institut fr Deutsche Sprache, Mannheim.

Stephan Stein. 2007. Mündlichkeit und Schriftlichkeit aus phraseologischer Perspektive. In Harald Burger, Dmitrij Dobrovolskij, Peter Kühn, and Neal R. Norrick, editors, *Phraseologie. Ein internationales Handbuch zeitgenössischer Forschung*, volume 1, pages 220–236. de Gruyter, Berlin – New York.

Torsten Steinhoff. 2007. *Wissenschaftliche Textkompetenz: Sprachgebrauch und Schreibentwicklung in wissenschaftlichen Texten von Studenten und Experten*, volume 280 of *Reihe Germanistische Linguistik*. de Gruyter, Berlin – New York.

Torsten Steinhoff. 2009. *Wortschatz–eine Schaltstelle für den schulischen Spracherwerb?*, volume 17/2009 of *SPASS*. Universität Siegen, FB 3.

Tobias Thelen. 2010. *Automatische Analyse orthographischer Leistungen von Schreibanfängern*. Ph.D. thesis, University of Osnabrück.

Alison Wray. 2005. *Formulaic language and the lexicon*. Cambridge University Press.

Seid Muhie Yimam, Richard Eckart de Castilho, Iryna Gurevych, and Chris Biemann. 2014. Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. In Kalina Bontcheva and Zhu Jingbo, editors, *Proceedings of the 52nd Annual Meeting of the ACL. System Demonstrations*, pages 91–96. ACL, jun.

Heike Zinsmeister and Margit Breckle. 2012. The alesko learner corpus: design–annotation–quantitative analyses. *Multilingual Corpora and Multilingual Corpus Analysis. Amsterdam: John Benjamins*, pages 71–96.

Annotating Content Zones in News Articles

Daniela Baiamonte
University of Pavia
daniela.baiamonte01@universitadipavia.it

Tommaso Caselli
VU University Amsterdam
t.caselli@vu.nl

Irina Prodanof
University of Pavia
irina.prodanof@gmail.com

Abstract

English. This paper presents a methodology for the annotation of the semantic and functional components of news articles (Content Zones, henceforth CZs). We distinguish between narrative and descriptive zones and, within them, among finer-grained units contributing to the overall communicative purpose of the text. Furthermore, we show that the segmentation in CZs could provide valuable cues for the recognition of time relations between events.

Italiano. *In questo lavoro viene presentata una metodologia per l'annotazione delle componenti semantiche e funzionali del testo giornalistico (Zone di Contenuto). Distinguiamo tra zone narrative e descrittive e, al loro interno, tra ulteriori unità che contribuiscono al dispiegamento dello scopo comunicativo del testo. Inoltre, mostriamo che la segmentazione in Zone di Contenuto offre preziosi indizi per il riconoscimento delle relazioni temporali tra eventi.*

1 Introduction

The logical structure of a document, i.e. its hierarchical arrangement in sections, paragraphs, sentences and the like, reflects a functional organization of the information flow and creates expectations on where the desired information may be located. As it is often the case, however, breakups in sections and paragraphs are motivated by style or even arbitrary choices.

The segmentation of the text in Content Zones (CZs, henceforth), i.e. functional categories contributing to the overall message or purpose, as induced by the genre of the text[1], provides more reliable and fine-grained cues to access the structure of its *types of functional content*. Previous attempts to annotate CZs have mainly focused on highly standardized texts like scientific articles (Teufel et al., 2009; Liakata et al., 2012; Liakata and Teufel et al., 2010) and scheduling dialogues (Taboada and Lavid, 2003), or on semi-structured texts like film reviews (Bieler et al., 2007; Taboada et al., 2009). Other work (Palmer and Friedrich, 2014; Mavridou et al., 2015) adopts the theory of discourse modes (Smith, 2003) to distinguish between the different types of text passages in a text document.

To the best of our knowledge, no efforts have been undertaken to devise an annotation scheme targeting the functional structure of news articles in terms of their content: the inverted pyramid structure, i.e. the gathering of key details at the beginning, followed by supporting information in order of diminishing importance, is too coarse-grained to be effectively used for information extraction purposes. Our hypothesis is that modeling the document's content via CZs could yield benefits for high-level NLP tasks such as Temporal Processing, Summarization, Question-Answering, among others. In addition to this, CZs qualify as a higher-level analysis of a text/discourse which captures different information with respect to Discourse Relations. The remainder of the paper is structured as follows: in Section 2 the motivations of this work are presented, together with related studies. Section 3 reports on our inventory of CZs, used to annotate a corpus of English news articles. Details on the corpus are provided in Section 4. In Section 5, we describe a case-study on the correlation between CZs and temporal re-

[1] We adopt Systemic Functional Linguistics' view of genre as "a staged, goal oriented, purposeful activity in which speakers engage as members of our culture" (Martin, 1984:25).

lations to show that the segmentation in CZs can provide cues in recognizing temporal relations between events. Finally, Section 6 draws on conclusion and suggests directions for future work.

2 Motivations and related work

The bulk of the work on discourse structures has focused on low-level structures corresponding to Discourse Relations holding between textual segments pairs. CZs take a different view on texts, as they perform a function towards the text as a whole. As an instance of a particular genre, every text is meant to accomplish a culturally-established communicative purpose, e.g. a news article reports on events happening in the world. This goal is not accomplished all at once: separate functional *stages* (i.e. CZs) convey fragments of its overall meaning (Eggins and Martin, 1997). Therefore, the knowledge about the typical functional structure of genres can be exploited to predict the internal organization of a text. This kind of information can be of help to produce balanced summaries or to select the passages most likely to contain the answer to a question.

Teufel et al. (2009) and Liakata et al. (2010)'s works present two complementary perspectives on scientific papers: the former models their argumentative/rhetorical structure (following the knowledge claims made by the authors); the latter treats them as the humanly readable representations of scientific investigations. In the works of Bieler et al. (2007) and Taboada et al. (2009), two different kinds of zones are recognized in film reviews: formal zones (required by the genre, e.g. credits and cast) and functional zones (reflecting the abstract functions of describing and commenting).

In the elaboration of news articles' CZs, we were mostly inspired by Labov (2013)'s study of oral narratives of personal experiences and by Bell (1991)'s analysis of the structure of news stories.

3 Annotation Schema

The opposition between dynamicity and staticity, mainly realized by grammatical and lexical aspect, is adopted as the basic parameter for differentiating between two macro CZs: NARRATION and DESCRIPTION. The former is aimed at reporting temporally interrelated (dynamic) events, the latter is used to comment by focusing on selected entities, properties, and states of affairs. Each of these two macro CZs is further divided into more fine-grained categories.

The class NARRATION (NARR) includes the following zones:

- `Foreground` (FGR): text span containing the most salient events, i.e. those in the focus of attention (as intended by Boguraev and Kennedy, 1999). The information it conveys is both referentially and relationally new (Gundel and Fretheim, 2005), as it is usually mentioned at the beginning of the article.
- `Background` (BGR): ancillary, referentially and relationally old information performing an explanatory function (through causal and temporal precedence relations) towards FGRs.
- `Follow-up` (FUP): reactions and consequences to FGR events (to whom they're related through cause-effect and temporal succession relations), i.e. relationally new information moving the discourse forward.
- `Expectation` (EXP): assumptions and probable or possible outcomes, i.e. non factual information (e.g. conditionals, modality).

The class DESCRIPTION (DSCR) includes the following zones:

- `Description` (DES): characteristics of a person or an object, customary circumstances, or states of affairs.
- `Evaluation` (EVL): subjective descriptions, explicit judgements showing the author or some other agent's attitude towards a target.

In addition, a third macro-class is posited, OTHER (OTHR), containing categories performing auxiliary functions towards the other CZs:

- `Attribution` (ATT): text span containing the source and, if present, the cue of an attribution (as intended by Pareti and Prodanof, 2010) - while the content is assigned the relevant CZ(s).
- `Metatext` (MTX): text span guiding the reader's attention towards metatextual elements like figures or tables.
- `Interrogative` (INT): questions directly addressed to the reader, e.g. to introduce a new topic or to prompt a reaction.

Major approaches to functional discourse structuring adopt the sentence or the paragraph as unit of annotation. On the other hand, we have opted for a clause level annotation as this allows us to better deal with news articles' high level of information density. Although CZs are conceptually non-overlapping, empirical analysis indicates that an annotation unit may fit into more than one category, that is a clause may represent complex contents. Cases as such suggest that the more informative content should be preferred. In the example below, the tag ATT is assigned, even though a descriptive content may as well be recognized.

1. [On an office wall of the Senate intelligence committee hangs a quote from Chairman David Boren,]$_{ATT}$ {PDTB[2], wsj_0771}

The annotation of CZs is further complicated by the fact the distribution of the zones does not follow the linear order of the text. In most cases, CZs are discontinuous, that is either their contiguity may be "broken" by the presence of other CZs or the same CZ may appear in different sentences along the entire document (see example **??** for the FGR zone).

2. [South Korea registered a trade deficit of $101 million in October,]$_{FGR}$ [reflecting the country's economic sluggishness,]$_{EVL}$ [according to government figures released Wednesday.]$_{ATT}$ [Preliminary tallies by the Trade and Industry Ministry showed another trade deficit in October, the fifth monthly setback this year,]$_{FGR}$ [casting a cloud on South Korea's export-oriented economy.]$_{EVL}$ {PDTB, wsj_0011}

In other cases, due to the use of the clause as minimal annotation span of a CZ, nested CZs may occurr (see example **??**).

3. [South Korea's economic boom, [which began in 1986,]$_{BGR}$ stopped this year because of prolonged labor disputes, trade conflicts and sluggish exports.]$_{BGR}$ {PDTB, wsj_0011}

4 Description of the corpus

We used the CZs annotation schema and the annotation tool CAT (Bartalesi Lenzi et al., 2012) to construct a small corpus of 57 news articles (20 from the test section of TempEval-3 (UzZaman et al., 2013), 20 shared between the PDTB (Prasad et al., 2008) and the training section of the TimeBank (Pustejovsky et al., 2003), 17 from the PDTB). The corpus contains 2059 annotation units and it is dominated by narrative sections (57%). Within them, the most frequent CZ is the BGR (26.5%), followed by FGR (12.4%), EXP (9.6%) and FUP (8.4%). These figures show that news articles mostly consist of redundant information, only mentioned in order to help the reader to anchor the new data to the prior knowledge. Descriptive sections constitute the 25.5% of the corpus: EVLs are slightly more frequent than DESs (14.8% vs. 8.9%) — contradicting the alleged objectivity expected in news reports (note, however, that EVLs tend to occur in association with ATTs). As to the OTHER macro CZ, it makes up the 17.4% of the corpus: this percentage almost entirely refers to ATTs, since MTXs and INTs are only marginal zones (0.19% and 0.33%, respectively).

Tense	**FGR**	**BGR**	**FUP**	**EXP**	**DES**	**EVL**	**ATT**	**MTX**	**INT**
PRESENT	46	46	26	21	35	58	34	0	0
PAST	66	204	29	3	2	10	172	0	0
FUTURE	21	1	25	22	0	2	0	0	0
INFINITIVE	32	51	26	18	2	12	1	0	0
PRESPART	6	19	10	5	2	3	9	0	0
PASTPART	2	11	1	1	0	2	1	0	0
NONE	6	8	6	21	0	2	2	0	0

Table 1: Distribution of tenses among CZs.

To test our hypotheses about some formal properties of CZs, we carried out a corpus study. The results are reported below.

Position in the text. 71.7% of FGRs are located in the opening sections of the articles and their occurrence decreases towards the central (18.4%) and closing sections (9.8%). BGRs show a fairly complementary distribution to FGRs, as they mostly occupy the central (31.6%) and closing sections (27.3%) of the articles. As expected, ATTs are quite evenly distributed among the three sections. The remaining CZs do not show any clear-cut tendencies.

Verbal tenses. Table **??** shows the distribution

[2]Penn Discourse TreeBank (Prasad et al., 2008).

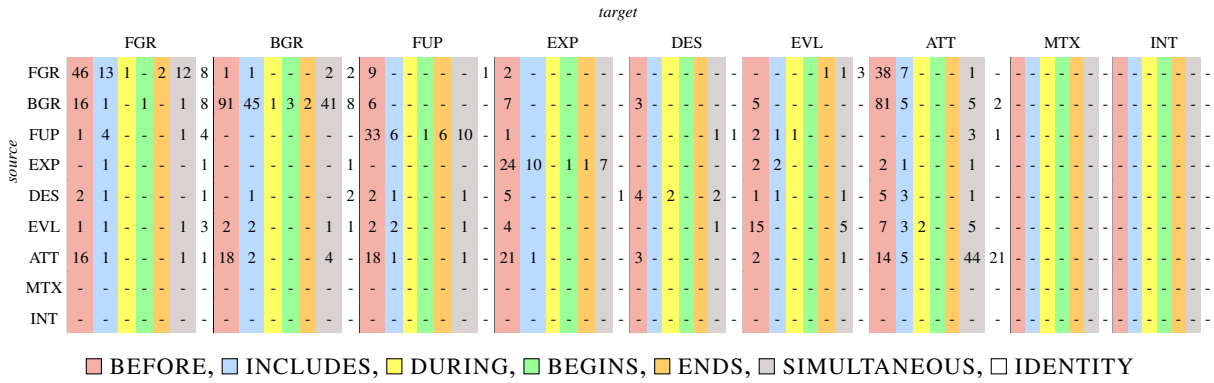


Table 2: Distribution of time relations among CZs.

of verbal tenses, as annotated in the TimeBank corpus, among CZs. BGRs and ATTs are dominated by the past tense, this is in accordance with our expectations as the former is characterized by temporal precedence relations to FGR events and the second mostly contains events of saying. CZs belonging to the DSCR class are significantly dominated by the present tense, usually associated with imperfective aspect and staticity. The high frequency of present tenses in FGRs and BGRs doesn't necessarily defy our expectations, since FGRs contain both dynamic and static events and the tag PRESENT is also used to refer to instances of present perfect.

Modality markers. The majority of modality markers is located in EXPs and, more broadly, in the narrative CZs, as shown in Figure **??**. In the TimeBank corpus, the MODALITY tag is mostly assigned to modal auxiliaries, we believe that the annotation of modal adverbs would further raise the percentages observed in EXPs and in the NARR class.

Pronouns. Looking at Figure **??** we can see that almost 50% of all pronouns is located in BGRs. The percentages are consistent with our expectations as BGRs convey referentially old information and, although FUPs and EXPs elaborate on FGR events, they often introduce new referents. Note that the distribution of pronouns is not, alone, a sufficient indicator of referential oldness since also lexical and zero anaphoras should be taken into account.

5 Interactions between CZs and time relations

In news articles events are not iconically presented in the linear order of their real succession, this poses a challenge to systems aimed at uncover-

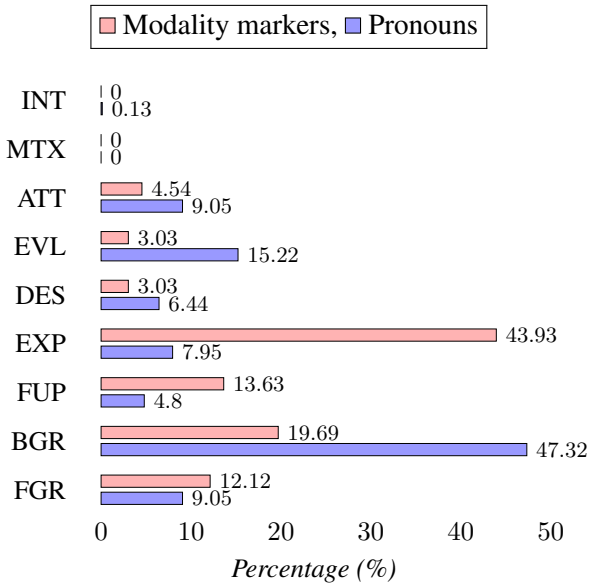


Figure 1: Distribution of modality markers and pronouns among CZs.

ing their temporal event structure. Therefore, we used the annotations available for the TimeBank section of the corpus to check whether some connections between CZs and temporal relations between event pairs exist. The full set of temporal relations specified in TimeML contains 14 types of relations: BEFORE, AFTER, IBEFORE, IAFTER, BEGINS, BEGUN_BY, ENDS, ENDED_BY, DURING, DURING_INV, INCLUDES, IS_INCLUDED, SIMULTANEOUS and IDENTITY. We simplified the set as follows: the relation types that invert each other were collapsed into a single type; given the low frequency of the relation type IBEFORE, it was mapped to the corresponding more coarse-grained type BEFORE.

Given the narrative shape of news articles, the corpus is considerably dominated by precedence

source - target	BFR	INCL	DUR	BEG	ENDS	SMLT	IDNT
NARR - NARR	218	77	1	6	11	71	26
NARR - DSCR	17	1	1	0	1	2	7
NARR - OTHR	119	9	0	0	0	10	3
DSCR - DSCR	30	5	3	0	0	12	6
DSCR - NARR	20	8	0	0	0	4	6
DSCR - OTHR	16	10	2	0	0	6	0
OTHR - OTHR	14	5	0	0	0	44	21
OTHR - NARR	72	5	0	0	0	6	0
OTHR - DSCR	6	0	0	0	0	1	1

Table 3: Distribution of time relations among the macro-classes.

(BEFORE) and succession (AFTER) relations. Table **??** shows that the majority of temporal relations holds between events belonging to the same CZ types: events tend to precede, include, occur during, begin, end, be simultaneous to and anaphorically evoke (through TimeML IDENTITY temporal relations) other events belonging to the same zone.

FGR events precede rather than follow ATT, FUP and EXP events. BGR events, the most involved in BEFORE relations, tend to precede other events, especially if located in ATTs and FGRs. Unexpected outcomes mostly occur in cases like the following, where the FGR event precedes the BGR one. This is because conflicting contents may be expressed in the same unit (in this case a reaction to the FGR event and the list of its premises):

4. [Delta Air Lines earnings soared to 33% to a record in the fiscal first quarter,]$_{FGR}$ [bucking the industry trend toward declining profits.]$_{FUP}$ [The Atlanta-based airline, the third largest in the U.S., attributed the increase to higher passenger traffic, new international routes and reduced service by Rival Eastern Airlines...]$_{BGR}$ {PDTB, wsj_1011}

As highlighted in Table **??**, NARR events begin or end other NARR or DSCR events (more specifically, these relations hold between events belonging to instances of the same CZ) and DSCR events include (rather than being included in) other events.

IDENTITY relations mostly involve FGRs: as a result of their textual salience, FGR events can be mentioned in other FGRs or further clarified in narrative or descriptive sections.

6 Conclusions and future work

We have developed an inventory of zone labels for the genre *news article* and shown that the so-generated *content structure* could help narrowing down the range of time relations connecting events.

Future work would involve testing the stability and reproducibility of the annotation scheme through the measurement of inter-annotator agreement and elaborating a separate annotation scheme for editorials, whose argumentative style reflects different structuring principles than those acting in news reports. Finally, we would like to automatize the process of annotation and test the effectiveness of the approach in texts belonging to different genres, e.g. novels (Ouyang and McKeown, 2014) and historical essays. Even the basic distinction between narrative and descriptive zones could facilitate the performance of more complex NLP tasks by targeting the relevant informational zones. The corpus and the annotation guidelines are publicly available[3].

Acknowledgment

This has been partially supported by the Erasmus + Traineeship Program 2015/2016 from University of Pavia and the NWO Spinoza Prize project Understanding Language by Machines (sub-track 3).

References

Baiamonte, D. 2016. *Annotazione di Zone di Contenuto: una strutturazione funzionale del testo giornalistico.* Thesis of the Master in Theoretical and Applied Linguistics. University of Pavia, Pavia.

Bärenfänger, M., Hilbert, M., Lobin, H., Lüngen, H., Puskás, C. 2006. Cues and constraints for the relational discourse analysis of complex text types - the role of logical and generic document structure. Sidner C.L., Harpur J. Benz A., Kühnlein P. (eds.), *Proceedings of the Workshop on Constraints in Discourse*. Maynooth, Ireland. 27-34.

Bartalesi Lenzi, V., Moretti, G., Sprugnoli, R. 2012. CAT: the CELCT Annotation Tool. *Proceedings of LREC 2012.* Istanbul.

[3] https://github.com/cltl/ContentZones.git

Bell, A. 1991. *The Language of News Media.* Blackwell Publishers, Oxford.

Bieler, H., Dipper, S., Stede, M. 2007. Identifying Formal and Functional Zones in Film Reviews. Keizer S., Bunt H., Paek T. (eds.), *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*. Antwerp, Belgium. 75-78.

Boguraev, B. and Kennedy, C. 1999. Salience-based content characterisation of text documents. Inderjeet M. and Maybury M. T. (eds.), *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA.

Eggins, S. and Martin, J. R. 1997. Genres and registers of discourse. van Dijk T. (ed.), *Discourse Studies. Discourse as structure and process*, volume 1. Sage, London (UK) and Thousand Oaks (CA). 230-257.

Gundel, J. K. and Fretheim, T. 2005. Topic and Focus. Horn L. and Ward G. (eds.), *The Handbook of Pragmatics*. Blackwell Publishing, 175-196.

Labov, W. 2013. *The Language of Life and Death*. Cambridge University Press, Cambridge, UK.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., Rebholz-Schuhmann, D. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics 2012*, volume 28. 991-1000.

Liakata, M., Teufel, S., Siddharthan, A., Batchelor, C. 2010. Corpora for conceptualisation and zoning of scientific papers. *Proceedings of the 7th conference on Internation Language Resource and Evaluation (LREC10)*.

Martin, J. R. 1984. Language, register and genre. Christie F. (ed.), *Language studies: Children writing. Reader.* Deakin University Press, Geelong, Australia. 21-30.

Mavridou, K., Friedrich, A., Peate Sørensen, M., Palmer, A., and Pinkal, M. 2015. Linking discourse modes and situation entity types in a cross-linguistic corpus study. September 2015. In Proceedings of Linking Models of Lexical. *Sentential and Discourse-level Semantics (LSDSem).* , Lisbon, Portugal.

Ouyang, J. and McKeown, K. 2014. Towards automatic detection of narrative structure. *Proceedings of LREC14*, Reykjavik, Iceland.

Palmer, A. and Friedrich, A. 2014. Genre distinctions and discourse modes: Text types differ in their situation type distributions. *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing. Forlí-Cesena, Italy*.

Pareti, S. and Prodanof, I. 2010. Annotating Attribution Relations: Towards an Italian Discourse Treebank. *Proceedings of LREC10*.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., Webber, B. 2008. The Penn Discourse TreeBank 2.0. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC).* Marrakech, Morocco.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Day, D., Ferro, L., Gaizauskas, R., Lazo, M., Setzerr, A., Sundheim, B. 2003. The TimeBank Corpus. *Corpus Linguistics.* 647-56.

Smith, S. Carlota 2003. *Modes of discourse: The local structure of texts.* Cambridge University Press.

Stede, M. 2011. *Discourse Processing.* Morgan & Claypool Publishers. 7-38.

Taboada, M., Brooke, J., Stede, M. 2009. Genre based paragraph classification for sentiment analysis. *Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue.* Queen Mary University of London. 62-70.

Taboada, M. and Lavid, J. 2003. Rhetorical and Thematic Patterns in Scheduling Dialogues: A Generic Characterization. *Functions of Language, 10(2).* 147-179.

Teufel, S., Siddharthan, A., Batchelor, C. 2009. Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009.* Singapore.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J. Verhagen, M., Pustejovsky, J. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*

LICO: A Lexicon of Italian Connectives

Anna Feltracco
Fondazione Bruno Kessler
University of Pavia, Italy
University of Bergamo, Italy
feltracco@fbk.eu

Elisabetta Jezek
University of Pavia
Pavia, Italy
jezek@unipv.it

Bernardo Magnini
Fondazione Bruno Kessler
Povo-Trento, Italy
magnini@fbk.eu

Manfred Stede
University of Potsdam
Potsdam, Germany
stede@uni-potsdam.de

Abstract

English. This paper presents the first release of LICO, a Lexicon for Italian COnnectives. LICO includes about 170 discourse connectives used in Italian, together with their orthographical variants, part of speech(es), semantic relation(s) (according to the Penn Discourse Treebank relation catalogue), and a number of usage examples.

Italiano. *Questo contributo presenta la prima versione di LICO, un lessico di connettivi per l'italiano. LICO comprende circa 170 connettivi del discorso usati in italiano, di cui abbiamo raccolto varianti ortografiche, le parti del discorso, le relazioni semantiche (ricavate dal catalogo del Penn Discourse Treebank) espresse dal connettivo, e alcuni esempi d'uso.*

1 Introduction

Discourse connectives are explicit lexical markers that are used to express functional relations between parts of the discourse. As an example, the italian word "quando" in the sentence "Quando si preme sul bottone, la porta si apre da sola" (When you press the button, the door opens by itself) expresses a conditional relation between two parts of the sentence (from now on, *arguments*).

Work on discourse connectives in Computational Linguistics was initially part of Rhetorical Structure Theory (Mann and Thompson, 1988), where the focus is on discourse relations, which are at the basis of the notion of textual coherence. In Computational Linguistics, being able to identify connectives is a central task in "shallow discourse parsing", which has become very popular in recent years (e.g., (Lin et al., 2014)) and constituted the shared task of the CONLL conference in 2015 and 2016[1]. Downstream applications that can benefit from shallow discourse structure are, inter alia, sentiment analysis (e.g., (Bhatia et al., 2015) and argumentation mining (e.g., (Peldszus and Stede, 2013)).

Our work on connectives is mainly motivated by the fact that, to the best of our knowledge, still there is no high coverage resource of discourse connectives available for Italian. LICO, the Lexicon for Italian COnnectives, aims at filling this gap, providing a repository of Italian connectives aligned with recent developments in discourse relations (i.e. the last version (3.0) of the Penn Discourse Treebank (PDTB)).

In addition, the LICO lexicon takes advantage from DimLex, a similar repository for German (Scheffler and Stede, 2016; Stede and Umbach, 1998); in fact DimLex served as the main inspiration for creating LICO (see section 4). DimLex is an XML-encoded resource that can be used for NLP; the public version provides information on orthographical variants, syntactic behavior, semantic relations (in terms of PDTB), and usage examples. It is used for automatic discourse parsing, and also for semi-automatic text annotation using the ConAno tool (Stede and Heintze, 2004). Another relevant resource for connectives is LEXCONN, for French, (Roze et al., 2012), which contains about 300 connectives with their syntactic category and coherence relations from Segmented Discourse Representation Theory (Asher and Lascarides, 2003)(and to some extent Rhetorical Structure Theory (Mann and Thompson, 1988)).

LICO is freely distributed under a CC-BY licence.

[1] http://www.cs.brandeis.edu/ clp/conll16st/

2 Discourse Connectives

The definition of discourse connective is controversial both in traditional grammar and in the linguistic literature. Our definition is based on the encyclopedia entry on connectives by Ferrari (2010), included in the reference work for the Italian language recently published by Treccani. In this entry, connectives are defined as "each of the invariable forms [...], that introduce relations that structure "logically" the meanings of the sentence and of the text"[2]. The definition provided in Ferrari (2010) is restrictive, as it does not include variable forms, i.e. those forms which are subject to morphological modifications, such as *ne consegue/conseguiva che* 'it follows/followed/ that', nor does it include pragmatic uses of connectives (also known as *discourse markers*) such as causal *perché* 'why' in "Che ore sono? *Perché* ho dimenticato l'orologio" ('what time is it? Because I forgot my watch'). On the other end, it assumes that logical relations marked by connectives hold between events or assertions, and therefore includes as arguments for the relation nominal expressions such as "dopo il pressante invito ..." 'after the pressing invitation ...', i.e. expressions that contain an event nominal, - although the event is, in this case, referred to instead of predicated.

In our work, we partly drop the invariability criteria; we do not include forms which exhibit morphological inflection or conjugation, but we do include connectives which show a certain degree of lexical variability that is, multi-word expressions which are not totally rigid from a lexical point of view (*ad esempio/per esempio* 'for example'; see section 3).

3 The Structure of the Lexicon

Each entry in the LICO lexicon corresponds to a connective (including its variants). Currently, for each entry LICO specifies:

- whether the connective (or its variants) is composed by a single token ("part = single", e.g. *perché*) or by more than one token ("part = phrasal" e.g. *di conseguenza*);
- whether the connective is composed by correlating part ("orth = discont") or not ("orth = cont") and the specification of the two correlating parts, e.g. "orth = discont": *da una parte* ("part = phrasal"), *dall'altra* ("part = phrasal"); "orth = cont": *perché* ("part = single");
- possible orthographic variants: e.g. *ciò nonostante* ("part = phrasal") and *ciononostante* ("part = single");
- possible lexical variants: e.g *dopo di ché* and *dopo di ciò*. Notice that in some cases this lexical variants determine a different syntactic environment, such as *in modo da* and *in modo che*, the first being followed by infinitive form, the following by a subjunctive form;
- pos category: adverbs, preposition subordinating or coordinating conjunctions;
- the semantic relation(s) that the connective indicates, according to the PDTB 3.0 schema (see section 3.1);
- examples of the connectives for each semantic relation;
- possible alignments with lexicon of connectives in other languages.

Table 1 shows the entry for *quando*, which presents more than one semantic relation, and the entry for *ciononostante, ciò nonostante, nonostante ciò*, as example of a connective with orthografic variants in LICO.

3.1 Semantic relations

For the annotation of the semantic relation we used the PDTB 3.0 schema of relations (Webber et al., 2016; Rehbein et al., 2016) as proposed in the DimLex resource (Scheffler and Stede, 2016), which is our main reference resource.

The schema is a most recent version of PDTB 2.0 (Prasad et al., 2008; Prasad et al., 2007) and includes semantic relations structured in a hierarchy composed by three levels. In the first level, the *class level*, the relations are grouped in four major classes: *TEMPORAL*, *CONTINGENCY*, *COMPARISON* and *EXPANSION*. The second level, the *type level*, specifies further the semantics of the class level. For example, the *TEMPORAL: Synchronous* tag is used for connectives that indicate that the two arguments are simultaneous, while the

[2] "Il termine *connettivo* indica in linguistica ciascuna delle forme invariabili [...], che indicano relazioni che strutturano 'logicamente' i significati della frase e del testo".

▷ entry-id	146
▷ orth	cont
▷ part	single
	quando
▷ POS	subordinating
▷ sem relation	TEMPORAL: Synchronous ex.: Quando lasciò l'appartamento, arrivò la chiamata rel. to German id: 5
▷ sem relation	CONTINGENCY:Condition ex.: Quando si preme sul bottone, la porta si apre da sola. ex.: Quando me lo chiedi, lo lascerò stare. rel. to German id: 116
▷ entry-id	30
▷ orth	cont
▷ part	single
▷ variant	orthographic
	ciononostante
▷ orth	cont
▷ part	phrasal
▷ variant	orthographic
	ciò nonostante
▷ orth	cont
▷ part	phrasal
▷ variant	orthographic
	nonostante ciò
▷ POS	coordinating
▷ sem relation	COMPARISON:Concession:Arg2-as-denier ex.: La procura ha ordinato la restituzione dell'esemplare confiscato. Ciononostante l'istruttoria prosegue. rel. to German id: 74

Table 1: The connectives *quando* and *ciononostante, ciò nonostante, nonostante ciò* in LICO.

TEMPORAL: Asynchronous tag is used for connectives that indicate a before-after relation between the arguments. The third level (*subtype level*)[3] varies according to the role of the two arguments involved in the relation. For example, *CONTINGENCY:Cause:Reason* is used if the argument introduced by the connective -Arg2- is the reason for the situation in the other argument -Arg1- (e.g. I stayed at home, because it was raining), while *CONTINGENCY:Cause:Results* is used if Arg2 represents the result/effect of Arg1 (e.g. It was raining, therefore I stayed at home). Not every *type* has a further *subtype*.
In the LICO structure, each connective is assigned with one or more three-level tags.

4 The Current Resource

In this Section, we present the current resource and its construction. In particular, we focus on describing how the list of entries has been identified so far and how we proceeded to acquire the semantic information for each entry.

List of connectives. Currently, LICO is composed by 173 entries, each one corresponding to a connective and its orthographical or lexical variants. In order to compile this list we used a number of grammatical and lexical resources for Italian and for other languages.

First, we retrieved the list of connectives mentioned by Ferrari (2010) in the Enciclopedia Treccani for the entry *connettivi*[4] for a total of 33 connectives. Then, we retrieved the list of connectives tagged as *congiunzione testuale* in Sabatini Coletti 2006 (Sabatini-Coletti, 2005) discarding the ones of literary use, for a total of 70 entries. Finally, we benefited from the DimLex resource for German, as we enriched our list by identifying the equivalent Italian terms of the German connectives[5]. This process was facilitated by the presence of examples in the German resource in which the connective is displayed in context: only the Italian candidates that maintain the sense of the German connectives were added to LICO. We keep trace of this "German-Italian" links and we will use this information to enrich also the characteristic of the entry in LICO (e.g. *aber* → *ma*). A total of 127 entries were collected with this method. Figure 1 shows the overlap between the three resources and Table 2 shows a sample of the connectives in LICO and the respective sources.

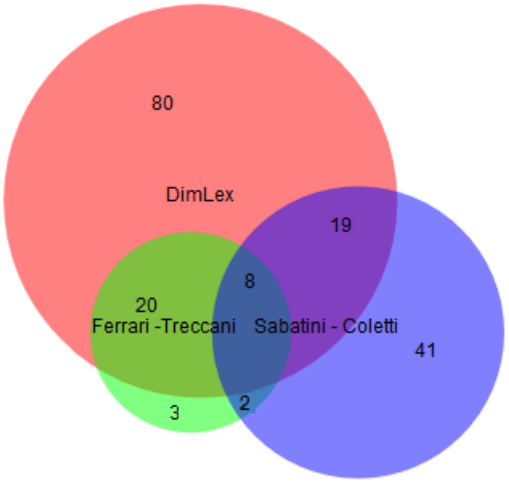


Figure 1: Overlap between the resources.

Semantic relations in LICO.
In LICO connectives are tagged with the semantic relations that the connective can indicate in a text, selecting the most appropriate ones in the PDTB 3.0 schema. In this process we took advantage from the information which was already

[3]The names of the levels are taken form Prasad et al. (2007).

[4]http://www.treccani.it/enciclopedia/connettivi_(Enciclopedia-dell'Italiano)/, last access July 21st 2016.

[5]https://github.com/discourse-lab/dimlex

LICO Entries	Resources		
	Ferrari Treccani	**Sabatini Coletti**	**DimLex (equivalent)**
dopo	dopo	dopo	dopo
dopo di che dopodiché		dopodiché	dopo di che
dopotutto		dopotutto	
dunque	dunque	dunque	dunque
e	e		e
ebbene		ebbene	
eccetto			eccetto
eppure		eppure	eppure

Table 2: Sample of connectives in different resources.

present in the resources we used for building the list. In fact, the DimLex resource provides this information for the German connectives, and both the Italian resources previously mentioned provide useful information about the semantic relation triggered by the connective.[6] A total of 23 different PTDB relations have been used to describe LICO entries. In order to validate the tagging of semantic relations, we conducted a research by observing examples of the use of the connectives in corpora, i.e. we wanted to verify whether the relation that a connective introduces in a portion of text is one of the relations already tagged for that same connective in the first step. In particular, we searched for 20 connectives in the ItWac corpus (Baroni et al., 2009) and we retrieved occurrences with 400 characters on both sides of the connective. We limited our observation to 5 retrieved segments of text in which the connective is actually playing such a role. We finally tagged each connective in each portion of text with the semantic relation it indicates.

To further confirm the corpus-driven evidences for the semantic relations, we asked two annotators (one being an expert annotator, the other not) to perform the same tagging task. We then calculated the interannotator agreement between the two annotators adopting the Dice's coefficient (Rijsbergen, 1997)[7] for three configurations, one for each level of the relation schema: *class agreement*, *type agreement*, *subtype agreement*. We considered that there was agreement if both annotators identify exactly the same *class*, *type*, *subtype* respectively. The Dice values result in 0,78 for *class agreement* and 0,71 for both *type agreement* and *subtype agreement*.

Observing cases of disagreement, we can make the following preliminary considerations. The main cases of disagreement regard the COMPARISON:Contrast relation (on one hand) and the COMPARISON:Concession and EXPANSION:Substitution relations (on the other hand). These relations in fact appear to be the ones that connect arguments that are in contrast. As an example, the connective *anziché* 'rather than' in Example (1) has been annotated as COMPARISON:Contrast by annotator1 and as EXPANSION:Substitution:Arg1-as-subst by annotator2: the first enlightens the contrast between "emissione attraverso il Tesoro" and "usare il tradizionale sistema", the second emphasises that Arg2 represents the alternative to the Arg1.

(1) [..] chiedeva l' emissione di dollari in banconote statunitensi attraverso il Tesoro *anziché* usando il tradizionale sistema della Federal Reserve.

Another interesting case concerns the disagreement between the relations TEMPORAL:Asynchronous:precedence (in which Arg2 follows Arg1) and CONTINGENCY:Cause:Result (in which Arg2 is the results of Arg1), being the two strictly connected (i.e. in a cause-effect relation, the effect follows the cause). As an example, in (2) one annotator marks the connective as indicator of the temporal sequence of Arg1 and Arg2, while the other prefers to mark it as an indicator of the cause-effect relation.

(2) [..] Il bello è che i tipi hanno pure accennato a prendersela con me, *al che* io gli ho abbaiato contro una sequela di insulti [..]

In general, the relations that were initially as-

[6] In particular, in the online version of Sabatini Coletti (http://dizionari.corriere.it/dizionario italiano/D/dizionario .shtml, last access July 21st 2016) the semantic relations the connectives can trigger are described in the definition of the connective itself, e.g. "*quindi*, cong. testuale: Con valore deduttivo-conclusivo, perciò, di conseguenza, per questo motivo, dunque". Ferrari (2010) in the Enciclopedia Treccani proposes a non hierarchical classification which includes the following relations: "temporal relation" "causal relation", "consequence relation", "condition relation", "opposition relations".

[7] Dice's coefficient measures how similar two sets are by dividing the number of shared elements of the two sets by the total number of elements they are composed by. This produces a value from 1, if both sets share all elements, to 0, if they have no element in common.

signed to these connectives were confirmed by the corpus-based exercise (i.e. at least one annotator assigns the tag in at least one portions of text); viceversa, in some cases one of the two annotators assigned a relation that was not initially identified.[8]

5 Conclusion and Further work

In this paper we have presented LICO, a new resource for the Italian language describing lexical properties of discourse connectives. While LICO fills a gap with respect to similar resources existing for other languages, it is still under construction under several aspects. Our short term plans include the completion of the lexical entries with corpus derived examples and the observation of the connectives in Italian corpora, in order to acquire more information about the semantic relations that each connective can indicate and thus extend the annotation of the semantic relations in LICO.

Acknowledgment

We acknowledge Denise Pangrazzi for her contribution to identify the Italian equivalents of the German connectives.

References

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from rst discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, September. Association for Computational Linguistics.

Angela Ferrari. 2010. Connettivi. In *Enciclopedia dell'Italiano*. diretta da Raffaele Simone, con la collaborazione di Gaetano Berruto e Paolo D'Achille, Roma, Istituto della Enciclopedia Italiana.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184.

[8]For this moment, the "new" relations are not included in LICO.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L Webber. 2007. The Penn Discourse Treebank 2.0 Annotation Manual.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May.

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating Discourse Relations in Spoken Language: A Comparison of the PDTB and CCR Frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May.

CJ van Rijsbergen. 1997. Information retrieval. 1979.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. LEXCONN: a French lexicon of discourse connectives. *Discours. Revue de linguistique, psycholinguistique et informatique.*, (10).

Il Sabatini-Coletti. 2005. Dizionario della lingua italiana 2006, con CD-ROM. *Milano, Rizzoli Larousse*.

Tatjana Scheffler and Manfred Stede. 2016. Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May.

Manfred Stede and Silvan Heintze. 2004. Machine-assisted rhetorical structure annotation. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 425–431, Geneva.

Manfred Stede and Carla Umbach. 1998. Dimlex: A lexicon of discourse markers for text generation and understanding. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 1238–1242. Association for Computational Linguistics.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. A Discourse-Annotated Corpus of Conjoined VPs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 22–31. Association for Computational Linguistics.

Formatio formosa est.

Building a Word Formation Lexicon for Latin

Eleonora Litta, Marco Passarotti, Chris Culy
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo Gemelli, 1 – 20123 Milan, Italy
{eleonoramaria.litta, marco.passarotti}@unicatt.it,
chrisculy@mac.com

Abstract

English. This paper presents the steps undertaken for building a word formation lexicon for Latin. The types of word formation rules are discussed and the semi-automatic procedure to pair their input and output lexical items is evaluated. An on-line graphical query system to access the lexicon is described as well.

Italiano. *Questo articolo presenta le procedure di realizzazione di un lessico morfologico derivazionale per il latino. Sono descritti i tipi di regole di formazione di parola e viene valutata la qualità del sistema semi-automatico di individuazione delle parole in input e in output ad esse. Il sistema grafico d'interrogazione on-line dei dati è altresì presentato.*

1 Introduction

In the area of Natural Language Processing (NLP), derivational morphology has always been neglected if compared to inflectional morphology, which plays a central role in fundamental annotation tasks like PoS tagging. Yet enhancing textual data with derivational morphology tagging promises to provide strong outcomes. First, it organises the lexicon at higher level than words, by building word formation based sets of lexical items sharing a common derivational ancestor. Secondly, derivational morphology acts like a kind of interface between morphology and semantics, since core semantic properties are shared at different extent by words built by a common word formation process.

Lately, some lexical resources for derivational morphology have been made available. Among them are the lexical network for Czech DeriNet (Ševčíková and Žabokrtský, 2014), the derivational lexicon for German DErivBASE (Zeller et al., 2013) and that for Italian derIvaTario (Talamo et al., 2016). Furthermore, stemming is a technique largely used for detecting word formation processes (Goldsmith, 2001), and language independent NLP tools were trained to extract derivation information from inflectional lexica (Baranes and Sagot, 2014).

On the Classical languages front, although the number of resources and NLP tools for Ancient Greek and Latin is now manifold and varied (ranging from digital libraries, treebanks and computational lexica to PoS taggers and parsers), no lexical resource for derivational morphology is available yet, where words are connected by word formation processes. The first steps towards building such a word formation lexicon for Latin were made by Passarotti and Mambrini (2012), who described a model for the semi-automatic extraction of word formation rules from the list of lemmas of *Lexicon Totius Latinitatis* by Forcellini (fifth edition; 1940) and the subsequent pairing of lexical entries and their derivational ancestor(s).

The *Word Formation Latin* project has received funding from the EU Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Individual Fellowship to expand on these efforts and create a word formation lexicon (working as an NLP tool as well) for Latin. In this paper, we describe the steps undertaken to build such a lexicon.

The paper is organised as follows. Section 2 presents the lexical basis supporting the lexicon; section 3 details the way the lexicon is built; section 4 describes how to access the data; section 5 concludes the paper and sketches the future work.

2 Lemlat

The lexical basis used for building the word formation lexicon is the one provided by the morphological analyser for Latin Lemlat (Passarotti, 2004). Resulting from the collation of three Latin dictionaries (Georges and Georges, 1913-1918; Glare, 1982; Gradenwitz, 1904), it counts 40,014 lexical entries and 43,432 lemmas (as more than one lemma can be included into the same lexical entry). Recently, the lexical basis of Lemlat was further enlarged by adding most of the Onomasticon (26,250 lemmas out of 28,178) provided by Forcellini (1940).

The basic component of the lexical look-up table used by Lemlat to morphologically analyse (and lemmatise) the input wordforms is the so-called les ("LExical Segment"), which roughly corresponds to the invariable part of the inflected forms. In other words, the les is the sequence (or one of the sequences) of characters that remains the same in the inflectional paradigm of a lemma (hence, the les does not necessarily correspond to the word stem). For instance, puell is the les for the lemma puell–a ("girl").

Lemlat includes a LES archive, in which each LES is assigned a number of inflectional features among which are a tag for the gender of the lemma (for nouns only) and a code (CODLES) for its inflectional category. For instance, the CODLES for the LES *puell* is N1 (first declension regular nouns) and its gender is F (feminine).

3 Building the Lexicon

The word formation lexicon is built in two steps. First, word formation rules are detected. Then, they are applied to lexical data.

3.1 Detecting Word Formation Rules

Word formation rules (WFRs) are conceived according to the so-called *Item-and-Arrangement* model, outlined by Hockett (1954), which considers word forms either as simple morphemes (not derived word forms) or as a concatenation of morphemes (derived word forms). The following conditions on bases and affixes do hold: (1) Baudoin's assumption that both bases and affixes are lexical elements (i.e. they are both morphemes); (2) as a consequence, they exist in the lexicon (Bloomfield's "lexical morpheme" theory); (3) they are dualistic, i.e. they have both form and meaning (Bloomfield's "sign-base" morpheme theory). The first two conditions motivate the fact that in our word formation lexicon affixes are recorded with the same status of lexical bases; the third condition concerns the semantic properties of WFRs mentioned in Section 1.

WFRs fall into two main types: (1) derivation and (2) compounding. Derivation rules are further organised into two subcategories: (a) affixal, in its turn split into prefixal and suffixal, and (b) conversion, a derivation process that changes the PoS of the input word without affixation.

Compounding and conversion WFRs are automatically detected, by considering all the possible combinations of main PoS (verbs, nouns, adjectives), regardless of their actual instantiations in the lexical basis. For instance, there are four possible types of conversion WFRs involving verbs: V-To-N (*claudo* → *clausa*; "to close" → "cell"), V-To-A (*eligo* → *elegans*; "to pick out" → "accustomed to select, tasteful"), N-To-V (*magister* → *magistro*; "master" → "to rule"), A-To-V (*celer* → *celero*; "quick" → "to quicken"). Each compounding and conversion WFR type is further specified by the inflectional category of both input and output. For instance, A1-To-V1 is the conversion WFR from first class adjectives to first conjugation verbs.

Affixal WFRs are found both according to previous literature on Latin derivational morphology (Jenks, 1911; Fruyt, 2011; Oniga, 1988) and in semi-automatic fashion. The latter is performed by extracting from the list of lemmas of Lemlat the most frequent sequences of characters occurring on the left (prefixes) and on the right (suffixes) side of lemmas. The PoS for WFR input and output lemmas as well as their inflectional category are manually assigned. Further affixal WFRs are found by confrontation with data. So far, we have detected 167 affixal WFRs: 71 prefixal and 96 suffixal.

We recorded the rules in a table of a MySQL relational database where each WFR is classified by type and it is assigned the required PoS, inflectional category and gender for its input and output.

3.2 Applying Word Formation Rules

Each morphologically derived lemma is assigned a WFR. All those lemmas that share a common (not derived) ancestor belong to the same "morphological family". For instance, lemmas *formatio* ("formation"), *formo* ("to form") and *formosus* ("beautiful", lit. "finely formed") all belong to the morphological family whose ancestor is the lemma *forma* ("form").

Lemmas and WFRs are paired by using a MySQL relational database whose main tables are the LES archive of Lemlat, the list of its lemmas (each assigned its PoS, inflectional category and, for nouns only, gender) and the list of WFRs.

A number of MySQL queries provide the candidate lemmas for each WFR. Some of these queries run on the list of lemmas, while others on the LES archive. In particular, most candidate lemmas of prefixal WFRs are found by running queries on the list of lemmas, as such rules tend to just add the characters of the prefix to the input lemma, like in the case of *accuso* → *sub+accuso* ("to blame" → "to blame somewhat"). Instead, suffixal WFRs are mostly assigned to their candidate input and output lemmas by running queries on the LES archive, because suffixes attach to LES instead of modifying full lemmas, like in *amo* → *amabilis* ("to love" → "lovable") where suffix *–bil–* attaches to LES *am* (plus the thematic vowel *–a–*, used for first conjugation verbs) instead of full lemma *amo*. Also, there are suffixal WFRs whose input is the basis of the irregular perfect participle of the input verb, like in *duco* → *ductilis* ("to lead" → "that may be led") where suffix *–il–* attaches to the basis of the irregular perfect participle of the verb *duco* (*duct*). Such irregular bases are recorded explicitly in the LES archive with a specific CODLES.

3.3 State of Affair and Evaluation

The procedure described above is not sufficient neither for detecting nor for applying the WFRs and, ultimately, for building the morphological families. Manual checking is largely needed for identifying false results and disambiguating duplication, as well as for filling lacunas resulting from the automatic process.

For example, while looking for the candidates of the WFR that forms adjectives from nouns with the addition of the suffix *–ax/–acis*, two candidate input nouns are found for the adjective *fugax* ("swift, transitory"): *fuga* ("flight") and *fugium* (rare, scarcely used in place of *fuga*). Such duplicate results need to be checked and disambiguated manually, as there must be only one input lemma for each output lemma resulting from a WFR of the derivation type, just like there must be only one WFR associated with each derived lemma.

Morphotactically obscure word formation processes, like most compounding WFRs, are examples of lacunas of the automatic process of assigning WFRs, which are thus fully manually hard-coded. For instance, the compound lemma *matricida* ("matricide") is derived by compounding the input lemmas *mater* ("mother") and *caedo* ("to cut"), thus showing quite an obscure morphotactic configuration.

So far, we have applied to data 134 WFRs (45 prefixal, 80 suffixal, 6 conversion and 3 compounding), which corresponds to having assigned a WFR to 18,774 lemmas. Evaluation is performed by calculating the precision rate (Van Rijsbergen, 1979) of MySQL queries, i.e. the percentage of the correct candidate input-output pairs that are automatically assigned to a WFR by a query.

As expected, precision is higher when morphotactic mutations are lower. Indeed, while precision rates for prefixal rules range between 0.95 and 0.8, as they imply quite a few graphical mutations, precision for suffixal rules can vary heavily, ranging from 0.75 to as little as 0.3. Instead, the recall of queries has to be calculated later in the project, as currently we are unable to verify how many derived lemmas are not automatically picked up by queries.

4 Accessing the Data

The word formation lexicon can be accessed on-line through a visualisation query system (http://wfl.marginalia.it). The lexicon can be browsed either by WFR, affix, or input and output PoS or lemma. Drop down menus provide the available options for each selection, like for instance the list of affixes and lemmas.

Results are visualised as tree graphs, whose nodes are lemmas and edges are WFRs. Trees are interactive. Clicking on a node shows the full derivation tree ("word formation cluster", which is calculated dynamically) for the lemma reported in that node. For example, figure 1 shows the currently available word formation cluster for the lemma *amo*. One can see that *amabilis* derives from *amo* and it is in turn the input for two other derived lemmas: *amabilitas* ("loveliness") and *inamabilis* ("unlovely"). Clicking on an edge shows the lemmas built by the WFR concerned in that edge. Lemmas are provided both as a derivation graph and as an alphabetical list. For instance, clicking on the edge going from *amo* to *amabilis* in figure 1 shows the lemmas built by the derivation WFR that builds second class adjectives (A2) from first conjugation verbs (V1) with suffix *–bil–*.

Figure 2 presents a portion of the derivation graph for this rule.

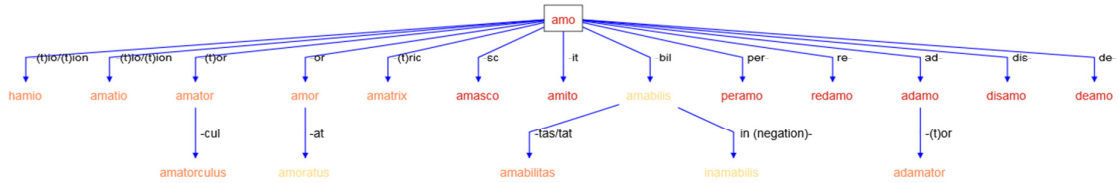


Figure 1. Word formation cluster for *amo*.

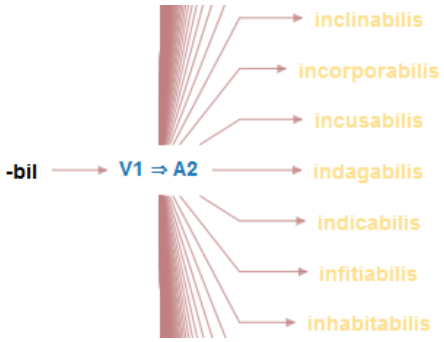


Figure 2. Derivation graph for a WFR.

5 Conclusion and Future Work

The building process of the word formation lexicon for Latin is ongoing. We still have to fully exploit the potential of querying the lexical basis of Lemlat to automatically detect candidates for WFRs. Furthermore, a substantial amount of manual work is needed to pick up morphotactically obscure formations, like those resulting from compounding.

The word formation lexicon is meant to enhance Lemlat by providing its processing with word formation analysis of input data, thus building a wide lexical resource and NLP tool for Latin morphology, which will be made available through CLARIN infrastructure (www.clarin.eu).

References

Marion Baranes and Benoît Sagot. 2014. A Language-Independent Approach to Extracting Derivational Relations from an Inflectional Lexicon. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ELRA, Reykjavik, Iceland, 2793–2799.

Egidio Forcellini. 1940. *Lexicon totius latinitatis*. Typis Seminarii, Padova.

Michele Fruyt. 2011. Word Formation in Classical Latin. J. Clarckson (ed.), *A Companion to the Latin Language*, Wiley-Blackwell, Chichester/Malden, Mass, 157–175.

Karl Ernst Georges and Heinrich Georges. 1913-1918. *Ausfuhrliches Lateinisch-Deutsches Handwôrterbuch*. Hahn, Hannover.

Peter GW. Glare. 1982. *Oxford Latin Dictionary*. At the Clarendon Press, Oxford.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2): 153–198.

Otto Gradenwitz. 1904. *Laterculi vocum latinarum*. Hirzel, Leipzig.

Charles F. Hockett. 1954. Two Models of Grammatical Description. *Words*, 10: 210–231.

Paul Rockwell Jenks. 1911. *A manual of Latin word formation for secondary schools*. DC Heath & Company, Harvard.

Renato Oniga. 1988. *I composti nominali latini: una morfologia generativa* (Vol. 29). Pàtron, Bologna.

Marco Passarotti. 2004. Development and perspectives of the Latin morphological analyzer LEMLAT. *Linguistica Computazionale*, XX-XXI: 397-414.

Marco Passarotti and Francesco Mambrini. 2012. First Steps towards the Semi-automatic Development of a Wordformation-based Lexicon of Latin. *Proceedings of the Eighth International Conference on Language Resources and*

Evaluation (LREC'12). ELRA, Istanbul, Turkey, 852–859.

Magda Ševčíková and Zdeněk Žabokrtský. 2014. Word-Formation Network for Czech. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. ELRA, Reykjavik, Iceland, 1087–1093.

Luigi Talamo, Chiara Celata and Pier Marco Bertinetto. 2016. DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure*, 9(1): 72–102.

Cornelis Joost Van Rijsbergen. 1979. *Information retrieval*. Butterworths, London, 2nd edition.

Britta D. Zeller, Jan Snajder and Sebastian Padó. 2013. DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, Sofia, Bulgaria, 1201-1211.

SIMPITIKI: a Simplification corpus for Italian

Sara Tonelli
Fondazione Bruno Kessler
satonelli@fbk.eu

Alessio Palmero Aprosio
Fondazione Bruno Kessler
aprosio@fbk.eu

Francesca Saltori
Fondazione Bruno Kessler
fsaltori@fbk.eu

Abstract

English. In this work, we analyse whether Wikipedia can be used to leverage simplification pairs instead of Simple Wikipedia, which has proved unreliable for assessing automatic simplification systems, and is available only in English. We focus on sentence pairs in which the target sentence is the outcome of a Wikipedia edit marked as 'simplified', and manually annotate simplification phenomena following an existing scheme proposed for previous simplification corpora in Italian. The outcome of this work is the SIMPITIKI corpus, which we make freely available, with pairs of sentences extracted from Wikipedia edits and annotated with simplification types. The resource contains also another corpus with roughly the same number of simplifications, which was manually created by simplifying documents in the administrative domain.

Italiano. *In questo lavoro si analizza la possibilità di utilizzare Wikipedia per selezionare coppie di frasi semplificate. Si propone questa soluzione come un'alternativa a Simple Wikipedia, che si è dimostrata inattendibile per studiare la semplificazione automatica ed è disponibile solo in inglese. Ci concentriamo soltanto su coppie di frasi in cui la frase target è indicata come il frutto di una modifica in Wikipedia, indicata dagli editor come un caso di semplificazione. Tali coppie sono annotate manualmente secondo una classificazione delle tipologie di semplificazione già utilizzata in altri studi, e vengono rese liberamente disponibili nel corpus SIMPITIKI. La risorsa include anche un secondo corpus, contenente circa lo stesso numero di semplificazioni, realizzato intervenendo manualmente su alcuni documenti nel dominio amministrativo.*

1 Introduction

In recent years, the shift of interest from rule-based to data-driven automated simplification has led to new research related to the creation of simplification corpora. These are parallel monolingual corpora, possibly aligned at sentence level, in which source and target are an original and a simplified version of the same sentence. This kind of corpora is needed both for training automatic simplification systems and for their evaluation. For English, several approaches have been evaluated based on the Parallel Wikipedia Simplification corpus (Zhu et al., 2010), containing around 108,000 automatically aligned sentence pairs from cross-linked articles between Simple and Normal English Wikipedia. Although this resource has boosted research on data-driven simplification, it has some major drawbacks, for example its availability only in English, the fact that automatic alignment between Simple and Normal versions shows poor quality, and that only around 50% of the sentence pairs correspond to real simplifications (according to a sample analysis performed on 200 pairs by Xu et al. (2015)). In this work, we present a study aimed at assessing the possibility to leverage a simplification corpus from Wikipedia in a semi-automated way, starting from Wikipedia edits. The study is inspired by the work presented in Woodsend and Lapata (2011), in which a set of parallel sentences was extracted from Simple Wikipedia revision history. However, the present work is different in that: *(i)* we use the Italian Wikipedia revision history, demonstrating that the approach can be applied also to languages other than English and on edits of Wikipedia that were not created for educational purposes, and *(ii)* we

manually select the actual simplifications and label them following the annotation scheme already applied to other Italian corpora. This makes possible the comparison with other resources for text simplification, and allows a seamless integration between different corpora.

Our methodology can be summarised as follows: we first select the edited sentence pairs which were commented as 'simplified' in Wikipedia edits, filtering out some specific simplification types (Section 3). Then, we manually check the extracted pairs and, in case of simplification, we annotate the types in compliance with the existing annotation scheme for Italian (Section 4). Finally, we analyse the annotated pairs and compare their characteristics with the other corpora available for Italian (Section 5).

2 Related work

Given the increasing relevance of large corpora with parallel simplification pairs, several efforts have been devoted to develop them. The most widely used corpus of this kind is the Parallel Wikipedia Simplification corpus (Zhu et al., 2010), which was automatically leveraged by extracting normal and simple Wikipedia sentence pairs. However, Xu et al. (2015) have recently presented a position paper, in which they describe several shortcomings of this resource and recommend the research community to drop it as the standard benchmark for simplification. Other alternative approaches, suggesting to further refine the selection of normal – Simple parallel sentences to target specific phenomena like lexical simplification, have been also proposed (Yatskar et al., 2010), but have had limited application. The fact that Simple Wikipedia is not available for languages other than English has proved beneficial to the development of alternative resources. Manually or automatically created corpora have been proposed among others for Brazilian Portuguese (Pereira et al., 2009), German (Klaper et al., 2013) and Spanish (Bott and Saggion, 2011). As for Italian, the only available corpus containing parallel pairs of simplified sentences is presented in Brunato et al. (2015). We borrow from this study the annotation scheme for our corpus, so that we can make a comparison between the two resources. We include in the comparison also another novel corpus, made of manually simplified sentences in the administrative domain, which we release together with the Wikipedia-based one.

3 Corpus extraction

The extraction of the pairs has been performed using the dump for the Italian Wikipedia available on a dedicated website.[1] This huge XML file (more than 1 TB uncompressed) contains the history of every operation of editing in every page in Wikipedia since it has been published for the first time. In particular, the Italian edition of Wikipedia contains 1.3M pages and is maintained by around 2.500 active editors, who made more than 60M edits in 15 years of activity. The Italian language is spoken by 70M people, therefore there are on average 35 active editors per million speakers, giving to the Italian Wikipedia the highest ratio among the 25 most spoken languages around the world.

We parse the 60M edits using a tool in Java developed internally and freely available on the SIMPITIKI website.[2] The user who edits a Wikipedia page can insert a text giving information on why he or she has modified a particular part of the article. This action is not mandatory, but it is included most of the times. We first select the edits which description includes word such as "semplificato" (simplified), "semplice" (simple), "semplificazione" (simplification), and similar. Then, the obtained set is further filtered by removing edits marked with technical tags such as "Template", "Protected page", "New page". This eliminates, for instance, simplifications involving the page template and not the textual content. The text in the Wikipedia pages is written using the Wiki Markup Language, therefore it needs to be cleaned. We use the Bliki engine[3] for this task. Finally, the obtained list of cleaned text passages is parsed using the Diff Match and Patch library,[4] identifies the parts of each article where the text was modified. With this process, we obtain a list of 4,356 sentence pairs, where the differences between source and target sentence are marked with *deletion* and *insertion* tags (see Figure 1).

4 Corpus annotation

We manually annotate pairs of sentences through a web interface developed for the purpose and freely available for download.[2] Differently from

[1] https://dumps.wikimedia.org/
[2] https://github.com/dhfbk/simpitiki
[3] http://bit.ly/bliki
[4] http://bit.ly/diffmatchpatch



Figure 1: Annotation interface used to mark simplification phenomena in the SIMPITIKI corpus.

corpora specifically created for text simplification, in which modifications are almost always simplifications, annotating Wikipedia edits is challenging because the source sentence may undergo several modifications, being partly simplifications and partly other types of changes. Therefore, the interface includes the possibility to select only the text segments in the source and in the target sentence that correspond to simplification pairs, and assign a label only to these specific segments. It also gives the possibility to skip the pair if it does not contain any simplification.

A screenshot of the annotation tool is displayed in Figure 1. On the left, the source sentence(s) are reported, with the modified parts marked in red (as given by the Diff Match and Patch library). On the right, the target sentence(s) were displayed, with segments marked in green to show which parts were introduced during editing. A tickbox next to each red/green segment could be selected to align the source and target segments that correspond to a modification. The annotation interface provides the possibility to choose one of the simplification types proposed in a dropdown menu ('Conferma'), or to skip the pair ('Vai Avanti'). The second option was given to mark the sentences where a modification did not correspond to a proper simplification. For example the last edit shown in Fig. 1 reports in the original version 'Contando *esclusivamente* sulla capacità del mare', which was modified into 'Contando *soprattutto* sulla capacità del mare'. Since this change affects the meaning of the sentence, turning *exclusively* into *mainly*, but not its readability, the pair was not annotated.

In order to develop a corpus which is compliant with the annotation scheme already used in previous works on simplification, we followed the simplification types described in (Brunato et al., 2015). The tagset is reported in Table 1 and comprises 6 main classes (Split, Merge, Reordering, Insert, Delete and Transformation) and some subclasses to better specify the Insert, Delete and Transformation operations. The labels are available in the dropdown menu on the annotation interface and can be used to tag selected pairs of sentences.

Class	**Subclass**
Split	
Merge	
Reordering	
Insert	Verb Subject Other
Delete	Verb Subject Other
Transformation	Lexical substitution (word) Lexical substitution (phrase) Anaphoric replacement Verb to Noun (nominalization) Noun to Verb Verbal voice Verbal features

Table 1: Simplification classes and subclasses. For details see Brunato et al. (2015).

5 Corpus analysis

So far, annotators viewed 2,671 sentence pairs, 2,326 of which were skipped because the target sentence was not a simplified version of the source one. 345 sentence pairs with 575 annotations are currently part of the SIMPITIKI corpus, and all

Class	Subclass	# wiki	# PA	Total
Split		20	18	38
Merge		22	0	22
Reordering		14	20	34
Insert	Verb	11	5	16
Insert	Subject	5	1	6
Insert	Other	58	21	79
Delete	Verb	12	1	13
Delete	Subject	17	1	18
Delete	Other	146	31	177
Transformation	Lexical Substitution (word level)	96	253	349
Transformation	Lexical Substitution (phrase level)	143	184	327
Transformation	Anaphoric replacement	14	3	17
Transformation	Noun to Verb	3	32	35
Transformation	Verb to Noun (nominalization)	2	0	2
Transformation	Verbal Voice	2	1	3
Transformation	Verbal Features	10	20	30
Total		575	591	1166

Table 2: Number of simplification phenomena annotated in the Wikipedia-based and the public administration (PA) corpus

phenomena presented in the annotation scheme proposed by (Brunato et al., 2015) are currently covered.

As a comparison, we analyse also the content of the annotated corpora described in (Brunato et al., 2015), which represent the only existing corpora for Italian simplification. These include the *Terence corpus* of children stories, which was specifically created to address the needs of poor comprehenders, and contains 1,036 parallel sentence pairs, and the *Teacher corpus*, a set of documents simplified by teachers for educational purposes, containing 357 sentence pairs. Besides, we include in the comparison also another corpus, which we manually created by simplifying documents issued by the Trento Municipality to rule building permits and kindergarten admittance. This corpus was simplified following the instructions in (Brunato et al., 2015) but pertains to a different domain, i.e. public administration (PA). The wikipedia-based and the PA corpus have a comparable size (575 vs. 591 pairs), but the simplification phenomena have a different frequency, as shown in Table 2.

In Fig. 2 we compare the distribution of the different simplification types across the four corpora. The graph shows that the same phenomena such as subject deletion, nominalizations, transfer of verbal voice tend to be rare across the four datasets. Similarly, the three top-frequent simplification types, i.e. delete-other, word transformation and phrase transformation, are the same across the four datasets. However, in the Wikipedia-based corpus, word transformation is less frequent than in the other document types, while phrase transformation is much more present. This may show that the 'controlled' setting, in which the Terence and the Teacher corpora were created, may lead educators to put more emphasis on word-based transformations to teach synonyms, while in a more 'ecological' setting like Wikipedia the performed simplifications are not guided or constrained, and phrase-based transformations may sound more natural. As for the PA documents, transformation phenomena are probably very frequent because of the technical language characterised by domain-specific words, which tend to be replaced by more common ones during manual simplification. In this corpus, noun-to-verb transformations are particularly frequent, since nominalizations are typical phenomena of the administrative language affecting its readability (Cortelazzo and Pellegrino, 2003).

While the *Terence* corpus contains on average 2.1 annotated phenomena per sentence pair, *Teacher* 2.8 and the *PA corpus* 2.9 , the Wikipedia-based corpus includes only 1.6 simplifications for each parallel pair. As expected, corpora that were explicitly created for simplification tend to have a higher concentration of simplification phenomena than corpora developed in less controlled settings.

As for non simplifications discarded during the

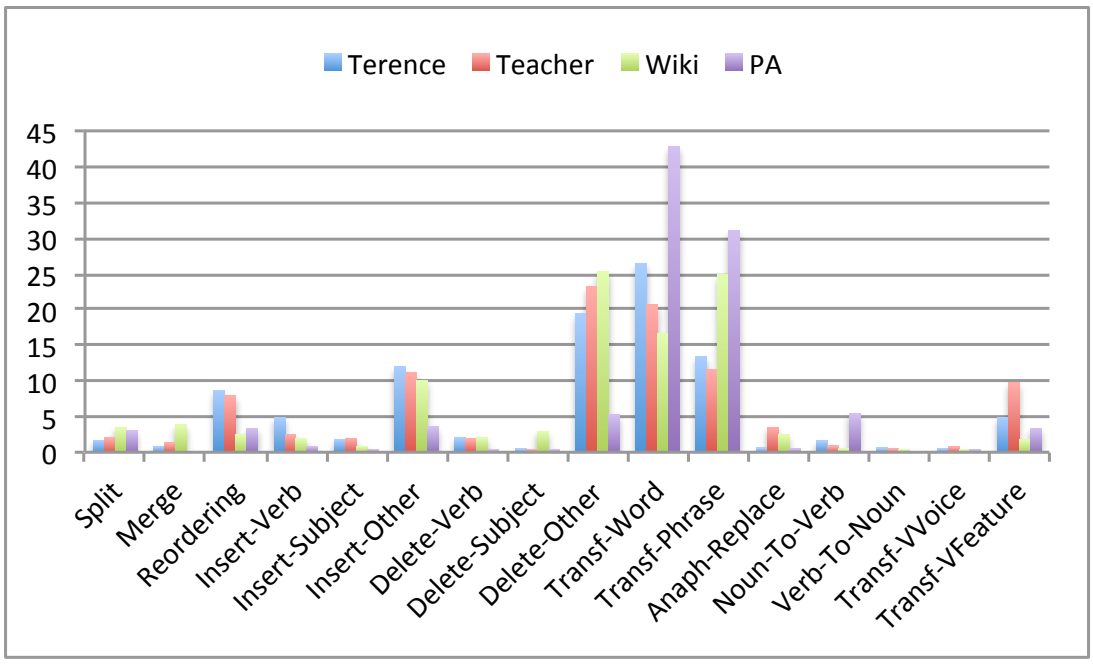


Figure 2: Distribution of the simplification phenomena covered in the Terence, Teacher and Wikipedia-based and Public Administration corpora.

1. Lo psicodramma è *stato il precursore di tutte le* forme di psicoterapia di gruppo 2. Lo psicodramma è *in relazione con altre* forme di psicoterapia di gruppo
1. Partigiani non comunisti e giornalisti *democratici* furono uccisi per il loro coraggio 2. Partigiani non comunisti e giornalisti furono uccisi per il loro coraggio
1. Il dispositivo di memoria di massa utilizza memoria allo stato solido, *ovvero basata su un semiconduttore* 2. Il dispositivo di memoria di massa *basata su semiconduttore* utilizza memoria allo stato solido

Table 3: Examples of parallel pairs which were not annotated as simplifications.

creation of the Wikipedia-based corpus, they include generalizations, specifications, entailments, deletions, edits changing the meaning, error corrections, capitalizations, etc. (see some examples in Table 3). These types of modifications are very important because they may represent negative examples for training machine learning systems that recognize simplification pairs.

6 Conclusions and Future work

We presented a study aimed at the extraction and annotation of a corpus for Italian text simplification based on Wikipedia. The work has highlighted the challenges and the advantages related to the use of Wikipedia edits. Our goal is to propose this resource as a testbed for the evaluation of Italian simplification systems, as an alternative to other existing corpora created in a more 'controlled' setting. The corpus is made available to the research community together with the tools used to create it. The SIMPITIKI resource contains also a second corpus, of comparable size, which was created by manually simplifying a set of documents in the administrative domain. This allows cross-domain comparisons of simplification phenomena.

In the future, this work can be extended in several directions. We plan to use the simplification pairs in this corpus to train a classifier with the goal of distinguishing between simplified and not-simplified pairs. This could extend the gold standard with a larger set of "silver" data by labelling all the remaining candidate pairs extracted from Wikipedia. Besides, the SIMPITIKI methodology is currently being used to create a similar corpus for Spanish, using the same annotation interface. The outcome of this effort will allow multilingual studies on simplification.

Finally, we plan to evaluate the Ernesta system for Italian simplification (Barlacchi and Tonelli, 2013) using this corpus. Specifically, since different simplification phenomena are annotated, it would be interesting to perform a separate evaluation on each class, as suggested in (Xu et al.,

2015).

Acknowledgments

The research leading to this paper was partially supported by the EU Horizon 2020 Programme via the SIMPATICO Project (H2020-EURO-6-2015, n. 692819).

References

Gianni Barlacchi and Sara Tonelli. 2013. ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part II*, pages 476–487, Berlin, Heidelberg. Springer Berlin Heidelberg.

Stefan Bott and Horacio Saggion. 2011. An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, MTTG '11, pages 20–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and Annotation of the First Italian Corpus for Text Simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41, Denver, Colorado, USA, June. Association for Computational Linguistics.

M. Cortelazzo and F. Pellegrino. 2003. *Guida alla scrittura istituzionale*. Laterza, New York, US.

David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a German/Simple German Parallel Corpus for Automatic Text Simplification. In *In: Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations, Sofia, Bulgaria*, pages 11–19.

Tiago F. Pereira, Lucia Specia, Thiago A. S. Pardo, Caroline Gasperin, and Ra M. Aluisio. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *In: 10th Conference on Intelligent Text Processing and Computational Linguistics, Mexico City*, pages 59–70.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the Sake of Simplicity: Unsupervised Extraction of Lexical Simplifications from Wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 365–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China, August. Coling 2010 Organizing Committee.

(Almost) Automatic Conversion of the Venice Italian Treebank into the Merged Italian Dependency Treebank Format

Linda Alfieri
FICLIT, University of Bologna, Italy
lindalfieri1988@gmail.com

Fabio Tamburini
FICLIT, University of Bologna, Italy
fabio.tamburini@unibo.it

Abstract

English. This paper describes the automatic procedure we developed to convert an Italian dependency treebank into a different format. We defined about 4,250 formal rules for rewriting dependencies and token tags as well as an algorithm for treebank rewriting able to avoid rule interference. At the end of this process a large portion of the whole treebank was automatically converted, with very few errors, leaving only a small amount of work to be done manually.

Italiano. *Questo contributo descrive la procedura automatica sviluppata per convertire un treebank italiano in un formato diverso. Abbiamo definito circa 4.250 regole formali di riscrittura per le strutture a dipendenza e i tag dei token e un algoritmo per la conversione del treebank in grado di evitare l'interferenza tra le regole. Al termine del processo una consistente sezione dell'intero treebank è stata automaticamente convertita, con un numero ridotto di errori, lasciando solo una piccola quantità di lavoro da svolgersi manualmente.*

1 Introduction

The availability of large annotated language resources is a prerequisite for the development of reliable automatic annotation tools using machine learning techniques.

Automatic tools able to enrich real texts with sentence syntactic structures are central instruments in Natural Language Processing (NLP) pipelines for a reliable annotation of text corpora. Modern NLP parsers heavily depend on complex training phases performed by examining manually annotated treebanks. Data sparsity, especially for low-resourced languages, seriously affect parsers performances, forcing scholars to annotate more and more data.

Since 2012 the state-of-the-art for Italian treebanks were not so satisfactory: three different projects and institutions produced three treebanks using different background theories, different formats and also different syntactic structures. They were the Italian Syntactic Semantic Treebank - ISST (Montemagni and Simi, 2007), the Turin University Treebank - TUT (Bosco et al., 2000) and the Venice Italian Treebank - VIT (Delmonte et al., 2007). Table 1 outlines the main characteristics of these treebanks at that time.

	ISST	**TUT**	**VIT**
Size (approx.)			
tokens	80,000	104,000	320,000
sentences	3,100	3,400	10,200
Type	Depend.	Depend.	Phr. Str.

Table 1: Italian treebanks in 2012.

ISST and TUT were used as gold standards in various evaluation campaigns (CoNLL2007 and EVALITA series), but only in 2012 the research groups developing such treebanks started to integrate them into a unique resource. In 2012 the Merged Italian Dependency Treebank - MIDT - was created and released by fusing the two resources (Bosco et al., 2012) and in the following years this project evolved such resource inserting it into the big Universal Dependency - UD - project (Nivre, 2015; Attardi et al., 2015), through another intermediate step, the Italian Stanford Dependency Treebank - ISDT (Bosco et al., 2013). During this process some other annotated texts were added to the treebank leveraging its size to around 315,000 tokens and 12,700 sentences (UD_Italian, v1.3).

This paper describes the latest effort for the Italian treebank merging: the conversion, harmoni-

sation and integration of the written sections of VIT, not previously included into ISST, with the other two resources for reaching a global amount of about 600,000 tokens and 23,000 sentences syntactically annotated. For practical issues we decided to convert VIT into the MIDT format and then use the set of already designed automatic procedures and checking programs to transform it into the final UD format.

There are other notable works aimed at treebank conversion in various languages, for example we can cite (Bos et al., 2009) for Italian.

2 The Venice Italian Treebank

The Venice Italian Treebank was created by the Laboratory of Computational Linguistics of the Department of Language Sciences, University of Venice (Delmonte et al., 2007). The theoretical framework behind VIT syntactic representation is the X-bar theory, thus the early version of the treebank expresses syntactic information as trees.

At a later time, one of the authors converted the treebank from phrase-structure to dependency structures (Delmonte, 2009), but this was not distributed. This version of VIT was the starting point for the conversion described in this paper.

3 The Merged Italian Dependency Treebank

The Merged Italian Dependency Treebank was created as a first attempt to merge two existing Italian resources, namely the TUT and a special version of the ISST treebank named ISST-TANL (Bosco et al., 2012) and represents the starting point for all subsequent attempts to convert and harmonise this resource to different standards, first the Stanford Dependencies[1] and last the Universal Dependencies[2].

4 VIT Conversion

The main part of the VIT conversion process was completely automatic. Using the Semgrex package[3] (Chambers et al., 2007) from the StanfordNLP group, we set up a set of procedures that, starting from the definition of conversion rules, automatically converted the VIT into the MIDT format. This procedure has been developed specifically for our conversion problem, but can be used, in principle, to convert any dependency treebank represented using the CoNLL format in a different format that does not require re-tokenisation steps.

4.1 The Semgrex language

Semgrex represents nodes in a dependency graph as a (non-recursive) attribute-value matrix. It then uses regular expressions for subsets of attribute values. For example, `{word:amo;tag:/N.*/}` refers to any node that has a value 'amo' for the attribute 'word' and a 'tag' starting with 'N', while '{}' refers to any node in the graph. The most important part of Semgrex is that it allows you to specify relations between nodes or group of nodes. For example, '`{}=1 <subj {}=2`' finds all the pairs of nodes connected by a directed 'subj' relation. Logical connectives can be used to form more complex patterns and node naming (the '=' assignments) can help retrieve matched nodes from the patterns.

Unfortunately Semgrex is simply a query language and, in its original form, cannot be used to rewrite dependency (sub)graphs. In order to extend the possibility of Semgrex, we then modified the original application to manage pairs of patterns: the first is used to search into the treebank for the required subgraphs, and the second is used to specify how the retrieved subsgraphs have to be rewritten. For example the pattern pair `{tag:det}=1 >arg {tag:noun}=2 --> {tag:ART}=1 <DET {tag:NN}=2`, what we called a 'Semgrex rule', changes the direction of the dependency and, at the same time, changes the words tags and relation label. The starting and final patterns have to contain the same number of nodes and dependency edges. Node naming has been the fundamental trick to introduce such extension allowing for node matching between patterns.

4.2 Conversion Procedure

For converting VIT into MIDT format, we manually defined about 4,050 Semgrex rules each capturing a specific syntactic configuration in VIT and transforming it into the MIDT schema and about 150 rules for residual tag rewriting. We spent about six months for writing the entire set of rules.

We have defined a set of new rewriting operations on a general dependency treebank:

- DEL_REL(graphID, depID, headID): deletes

[1] http://nlp.stanford.edu/software/stanford-dependencies.shtml

[2] http://universaldependencies.org/

[3] http://nlp.stanford.edu/software/tregex.shtml

a dependency edge between two graph nodes;

- INS_REL(graphID, depID, headID, label): inserts a new labelled dependency edge between two graph nodes;

- REN_TAG(graphID, nodeID, tag): replace the tag of a specific graph node.

The conversion task has been implemented as a three-steps process:

- first of all, each Semgrex rule is always applied to the original treebank producing a set of matching subgraphs that have to be rewritten;

- for each match, a set of specific operations for rewriting the subgraph corresponding to the processed matching are generated and stored;

- last, the whole set of operations produced processing the entire set of Semgrex rules, each applied to the original treebank, is sorted by graphID, duplicates are removed and every operation is applied graph by graph respecting the following order: first dependency deletions, second dependency insertions and lastly tag renaming.

This way of processing the original treebank and transforming it into the new format should guarantee that we do not experience rule interference during the conversion, because the generation of the rewriting operations due to the Semgrex rules application is decoupled from the real treebank rewriting.

5 Some Linguistic Issues

The set of rules manually written for converting VIT dependency structures can be subdivided into two macro-classes: (a) rules that do not modify the structures and (b) rules that need to modify the dependencies, both in term of edge direction and in term of different structuring between the involved nodes.

Regarding the rules that do not modify the dependency structures, they simply rename the dependency label using a 1:1 or an N:1 look-up table, as VIT, with respect to MIDT, typically involves more specific dependency types. Figure 1 outlines some simple examples of such kind of conversions.

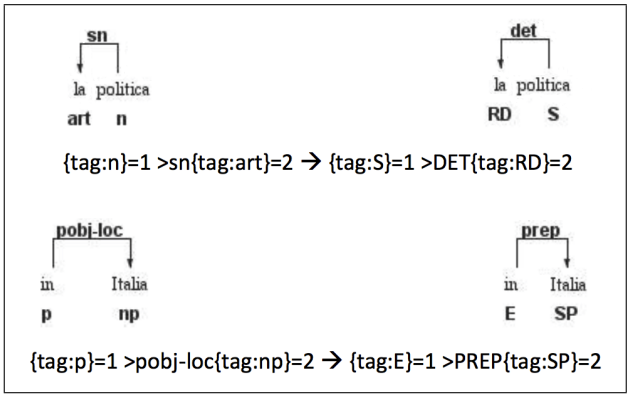


Figure 1: Some simple examples of rules that do not modify the dependency structures.

There are, of course, other kind of operations on subgraphs that require also the rewriting of the dependency structure. A good example concerns relative clauses in which the role of the relative pronoun and, as a consequence, the connections of the edge expressing the noun modification are completely different in the two formalisms. Figure 2 shows one example of this kind of rewriting.

Cases of coordination presented several problems: in VIT the head of the coordinated structure is linked to the connective and then the two (or possibly more) coordinated structures can be linked with a wide range of different dependency types (e.g. between phrases - *sn*, *sa*, *savv*, *sq*, *sp*, predicative complements - *acomp*, *ncomp*, adjuncts - *adj*, *adjt*, *adjm*, *adjv*, subjects - *subj*, objects - *obj*, etc.) leading to a large number of different combinations. Moreover, each dependency combination has to be further specified by the different token tags. MIDT represents coordinate structures in a different way: the connective and the second conjunct are both linked to the first conjunct that is connected to the head of the coordinated structure.

Figure 3 shows one example: the first formal rule represents an abstract rule pattern that has to be filled with all the real tag combinations found in VIT, generating a huge number of different rules, one of them outlined by the second complete formal rule. This process generated more than 2,800 different rules for handling all the coordinated structures in VIT.

There is also a need for a third kind of rules for rewriting single PoS-tags that might have remained unchanged during the main conversion process.

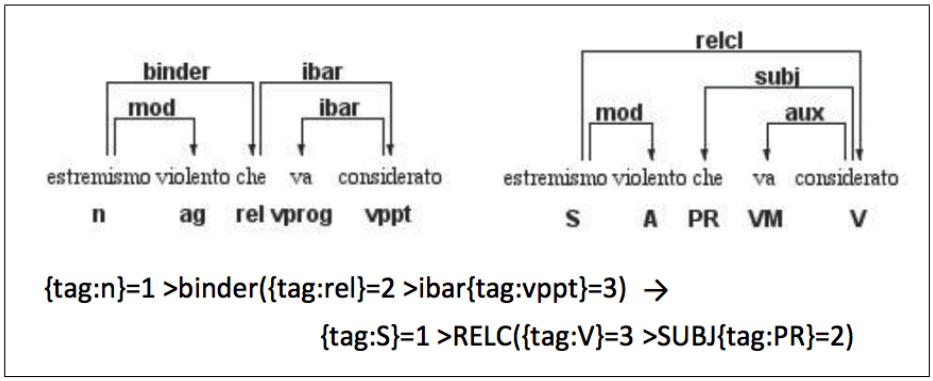


Figure 2: An example of rule that rewrite the dependency structure.

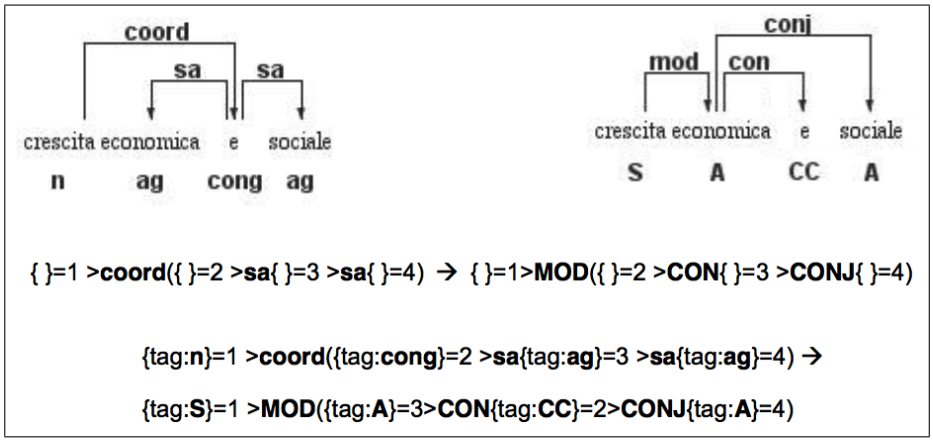


Figure 3: An example of coordination structures in VIT and MIDT and the conversion rule.

One further point deserves some discussion. In VIT, articulated prepositions are represented as two different tokens both linked with a common head: the preposition is tagged *part*/*partd*/*partda* and usually connected to the head with some kind of modification relations and the article is always tagged *art* and linked to the head with a *det* relation. In MIDT articulated prepositions are represented by a single token. As we said before, our process does not allow re-tokenisation rules. Given that MIDT is only an intermediate format and the goal is to convert VIT into the UD standard that requires two tokens for this phenomenon, we decided to avoid any re-tokenisation and to convert such structures linking the preposition to the head and the article to the preposition by introducing a new, dummy, relation label 'REL_EA'.

6 Evaluation

Applying all the 4,250 Semgrex rules, we obtained a converted treebank in which 228,534 out of 280,641 dependency relation were automatically converted, giving a global coverage of 81.4%.

To test the effectiveness of the conversion procedure and the conversion rules we randomly selected 100 sentences (2582 dependency relations to be converted) from the treebank and manually checked every newly created dependency relation, both in term of the connected nodes and the assigned label.

We obtained the following results: among the 2008 relations that have been automatically converted we found 125 wrongly converted dependency relations. So, on this sample, we obtained a coverage of 2008/2582 = 77.8%, slightly less than on the whole treebank, with a conversion error rate = 125/2008 = 6.2%.

7 Discussion and Conclusions

This paper presents the procedure we developed to convert VIT, one Italian treebank, into a different format. Most of the described conversion procedure rely on an automatic algorithm based on

formal rules that is able to automatically convert the 81.4% of the treebank. This procedure can be, in principle, adaptable to any conversion between different dependency treebank formats.

The formal rules has been manually defined by using a well known dependency search procedure, Semgrex from StanfordNLP group, properly extended to handle rewriting rules and the final result was manually evaluated to test the effectiveness of the written rules obtaining a very small error rate.

To the best of our knowledge, there is no general purpose tool available to automatise this task for dependency graphs. We can find some powerful converters in literature but they are usually tied to specific pair of tagsets (often tailored to the Penn treebank) (Johansson and Nugues, 2007; Choi and Palmer, 2010), and cannot be easily adapted to general needs, or are devoted to tree manipulation, for example the tool 'Tregex' (Levy and Andrew, 2006).

Even if the described procedure can convert a large part of the treebank automatically with a very small quantity of errors, the conversion certainly needs a careful manual analysis to complete the task and check the new treebank for remaining mistakes. The VIT treebank contains a lot of specific and peculiar dependency subgraph for representing phenomena in a very detailed way. Trying to capture all these different variations into formal rules can result in a very large rule set mostly composed of rule that handle single cases. We stopped the production of new rules when this situation arose.

Acknowledgments

We wish to thank Rodolfo Delmonte and Maria Simi for their precious suggestions and explanations for the analysis of linguistic phenomena and for defining the conversion process.

References

Giuseppe Attardi, Simone Saletti, and Maria Simi. 2015. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *Proc. of 2nd Italian Conference on Computational Linguistics - CLiC-it 2015*, pages 25–30, Trento.

Johan Bos, Cristina Bosco, and Alessandro Mazzei. 2009. Converting a dependency treebank to a categorial grammar treebank for Italian. In *Proc. of 8th International Workshop on Treebanks and Linguistic Theories - TLT8*, Milano.

Cristina Bosco, Vincenzo Lombardo, Daniela Vassallo, and Leonardo Lesmo. 2000. Building a treebank for Italian: a data-driven annotation schema. In *Proc. 2nd International Conference on Language Resources and Evaluation - LREC 2000*, pages 99–105, Athens.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2012. Harmonization and Merging of two Italian Dependency Treebanks. In *Proc. of LREC 2012, Workshop on Language Resource Merging*, pages 23–30, Istanbul.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *Proc. of ACL Linguistic Annotation Workshop & Interoperability with Discourse*, Sofia.

Nathanael Chambers, Daniel Cer, Trond Grenager, David Hall, Chloe Kiddon, Bill MacCartney, Marie-Catherine de Marneffe, Daniel Ramage, Eric Yeh, and Christopher Manning. 2007. Learning Alignments and Leveraging Natural Logic. In *Proc. of the Workshop on Textual Entailment and Paraphrasing*, pages 165–170.

Jinho Choi and Martha Palmer. 2010. Robust Constituent-to-Dependency Conversion for English. In *Proc. of 9th International Workshop on Treebanks and Linguistic Theories - TLT9*, Tartu, Estonia.

Rodolfo Delmonte, Antonella Bristot, and Sara Tonelli. 2007. VIT - Venice Italian Treebank: Syntactic and Quantitative Features. In *Proc. Sixth International Workshop on Treebanks and Linguistic Theories*.

Rodolfo Delmonte. 2009. Treebanking in VIT: from Phrase Structure to Dependency Representation. In Sergei Nirenburg, editor, *Language Engineering for Lesser-Studied Languages*, pages 51–81. IOS Press, Amsterdam, The Netherlands.

Richard Johansson and Pierre Nugues. 2007. Extended Constituent-to-dependency Conversion for English. In *Proc. of NODALIDA 2007*, Tartu, Estonia.

Roger Levy and Galen Andrew. 2006. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *Proc. of 5th International Conference on Language Resources and Evaluation - LREC 2006*, Genoa, Italy.

Simonetta Montemagni and Maria Simi. 2007. The italian dependency annotated corpus developed for the conll-2007 shared task. Tech. report, ILC-CNR.

Joakim Nivre. 2015. Towards a Universal Grammar for Natural Language Processing. In *Proc. of 16th International Conference Computational Linguistics and Intelligent Text Processing - CICLing 2015*, pages 3–16, Cairo, Egypt.

Enrichring the Ita-TimeBank with Narrative Containers

Alice Bracchi
Università degli Studi di Pavia
C.so Strada Nuova 65
27100 Pavia
alice.bracchi@gmail.com

Tommaso Caselli
Vrije Universiteit Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
t.caselli@vu.nl

Irina Prodanof
Università degli Studi di Pavia
C.so Strada Nuova 65
27100 Pavia
irina.prodanof@gmail.com

Abstract

English. This paper reports on an annotation experiment to enrich an existing temporally annotated corpus of Italian news articles with Narrative Containers, annotation devices representing temporal windows in text and marking up very informative temporal relations between temporal entities. The annotation has shown that the distribution of Narrative Containers is sensitive to the text genre and may be used to facilitate the creation of informative timelines.

Italiano. *Questo lavoro illustra i risultati di un esperimento di annotazione per l'identificazione di Contenitori Narrativi, ovvero marcatori di "finestre" temporali in un testo, come strategia per arricchire un corpus di articoli di quotidiano in lingua italiana, già annotato con informazioni temporali. L'annotazione ha mostrato che la distribuzione dei Contenitori Narrativi è legata al genere testuale e può essere usata per facilitare la creazione di linee temporali di eventi più informative.*

1 Introduction

Research in Temporal Processing has seen an increasing interest thanks to the availability of annotation schemes and corpora in multiple languages (Pustejovsky et al., 2003; Bittar et al., 2011; Caselli et al., 2011; Saurı and Badia, 2012), and the organization of evaluation campaigns (TempEval (Verhagen et al., 2007; Verhagen et al., 2010; UzZaman et al., 2013), Clinical TempEval (Bethard et al., 2015; Bethard et al., 2016), Cross-Document TimeLine (Minard et al., 2015), Temporal QA (Llorens et al., 2015)), and EVENTI (Caselli et al., 2014)). This has established best practices, common evaluation frameworks, international standards (e.g. ISO-TimeML (Pustejovsky et al., 2010)), and approaches to solve such a complex task. However, the expression of time in text/discourse is by no means obvious and the automatic extraction of timelines is not a solved task yet. One of the limits of current annotation frameworks and corpora relies mainly in the sparseness of the available temporal relations and in the fine-grained values used to classify the temporal links. For instance, in the TempEval-3 corpus the ratio between temporal relations and event plus temporal expressions is 0.8 (Bethard et al., 2014) for 13 temporal values. In the EVENTI corpus, the ratio is even smaller, only 0.19 for 13 temporal values. [1] Furthermore, in some cases annotation guidelines are not informative enough concerning what types of temporal links to annotate, or they force the annotation of temporal relations between pairs of events when they should not be annotated. Attempts to overcome these limits have focused on three main strategies: i.) annotating particular sets of temporal relations (Kolomiyets et al., 2012); ii.) elaborating detailed annotation guidelines for each kind of temporal relations (event-temporal expression pairs, event-event pairs, and temporal expression-temporal expressions pairs); and iii.) developing densely connected temporal graphs, where all valid relations among the temporal entities (events and temporal expressions) are marked up, including inferred relations based on transitive properties of the temporal relations (e.g. if event A is BEFORE event B and event B IS_INCLUDED in event C, then event A is BEFORE event C) (Bethard et al., 2014). We

[1] The smaller ratio for the Italian data is also due to specific restrictions on the annotation of the temporal relations as reported in the EVENTI Annotation Guidelines and explained in Section 2.

consider these solution as partial as they are not able to address the issue of identifying and extracting *informative* timelines, i.e. a set of maximally informative temporal links where relevant events in a text/discourse are correctly anchored to time, and then chronologically ordered. This paper reports on the first annotation effort to enrich existing resources for Temporal Processing in Italian by adopting a document-level approach rather than a sentence-level one. Following the proposal of **Narrative Containers** (NCs) (Pustejovsky and Stubbs, 2011), as embedding intervals where events occur, we developed an annotation scheme for their identification on the EVENTI corpus (Caselli et al., 2014) [2], as a strategy to increase the informativeness of the existing annotations and, possibly, improve systems's temporal awareness.

The remainder of this paper is structured as follows: the EVENTI corpus will be shortly introduced in Section 2, with a particular emphasis on the available temporal relations. Section 3 will present the notion of Narrative Container and the proposed annotation scheme. In Section 4 the results of a pilot annotation on the EVENTI dataset will be reported. Finally, conclusion, future work, and a pointer to the annotated data and guidelines will be reported in Section 5.

2 Temporal Relations in the EVENTI Corpus

The EVENTI corpus, released in the context of the EVALITA 2014[3] workshop, consists of 3 datasets: the Main task training data, the Main task test data, and the Pilot task test data. The corpus has been annotated with a simplified version of the It-TimeML Annotation Guidelines (Caselli et al., 2011), an adapted version to Italian of the TimeML Guidelines. Four tags have been used to annotate the data: `EVENT`, `TIMEX3`, `SIGNAL`, and `TLINK`.

The `EVENT` tag is used to annotate all lexical items which may realize an event mention. It includes verbs, nouns, adjectives, and prepositional phrases. The tag is enriched with 8 attributes expressing tense, (grammatical) aspect, part-of-speech, mood, modality, verb form, TimeML class, and polarity.

[2] `https://sites.google.com/site/eventievalita2014/`

[3] `http://www.evalita.it/2014/tasks/eventi`

The `TIMEX3` tag is used for the annotation of temporal expressions (timexes), expressing the type, the value and whether the timex is absolute or relative (e.g. "2015-05-18" *vs.* "yesterday"[ieri]).

The `SIGNAL` tag is employed to mark any linguistic elements, such as prepositions (e.g. *in* [in]), adverbs (e.g. *before* [prima]), or conjunctions (e.g. *when* [quando]), which support the identification and classification of a temporal relation between target entities (e.g. events and timexes).

Finally, the `TLINK` tag is used to annotate temporal relations. In the EVENTI task, the subset of possible temporal relations has been restricted to three subtypes of intra-sentence relations, namely: i.) pairs of syntactic main events in the same sentence; ii.) pairs of syntactic main event and subordinate event in the same sentence; and iii.) pairs of event and timexes. All 13 temporal relation values from It-TimeML (BEFORE, AFTER, IS_INCLUDED, INCLUDES, SIMULTANEOUS, I(MMEDIATELY)_AFTER, I(MMEDIATELY)_BEFORE, IDENTITY, MEASURE, BEGINS, ENDS, BEGUN_BY and ENDED_BY) have been used.

The Main task datasets, which have been enriched with Narrative Containers, add up to 130,279 tokens, divided into 103,593 tokens for training and 26,686 for test. They contain 21,633 `EVENT`s (17,835 in training and 3,798 in test), 3,359 `TIMEX3` (2,753 in training and 624 in test), 1,163 `SIGNAL`s (923 in training and 231 in test), and 4,561 `TLINK`s (3,500 in training and 1,061 in test).

3 Adding Narrative Containers to News Articles

The notion of Narrative Container (NC) was first introduced by Pustejovsky and Stubbs (2011) to deal with some aspects of Temporal Processing, such as sensitivity to the text genre and interaction with discourse relations, not addressed in the TimeML Guidelines nor in the TimeBank corpus. NCs were proposed as a temporal window, providing left and right boundaries, to when events not anchored to timexes could have happened, thus overcoming issues related to linking of events with the Document Creation Time (DCT), i.e. when a text was written or published. In particular, standard TimeML markup imposes that all events have

a link with the DCT but fail to specify that each event should also be annotating to its actual temporal anchor, i.e. to its moment of occurrence. As reported in Pustejovsky and Stubbs (2011), in example 1, TimeML guidelines will order both event mentions, e_1 and e_2, to the DCT with a BEFORE relation, anchor e_1 to the timex "yesterday" (*t*) but will fail to provide the anchoring of e_2:

1. The bomb **exploded**$_{e1}$ *yesterday*$_{t2011-09-09}$ and **killed**$_{e2}$ three people. [DCT=2011-09-10]

A further justification to the introduction of NCs is related to the different informational status of temporal relations. Assuming the informativeness of a temporal link as a function of the information contained in the individual links and their closure, an anchoring relation, that is a relation between a timex and an event explicitly stating when the event occurred as the one between e_1 and timex "yesterday" in example 1 (i.e. a temporal value of INCLUDES or IS_INCLUDED), is assumed to be more informative than an ordering relations, i.e. a precedence relation between two events.

To the best of our knowledge, the only corpus which extensively adopts the notion of NC and has available annotated data is the THYME corpus of clinical narratives (Styler IV et al., 2014). Our task is the first attempt at tackling temporal containment annotation over news articles in Italian.

A NC enables an accurate reproduction of the way events in text cluster around temporal reference points, explicitly or implicitly realized in the document, as the narration unfolds. NC relations are thus anchoring relations between pairs of events or events and temporal expressions. They are marked with an additional link tag, i.e. `CONTAINS`, to distinguish them from standard `TLINK`s. Each NC relation admits two components: i.) the narrative anchor, i.e. an element pointing to a specific temporal dimension shared by other events or timexes within the text; and ii.) the anchored element(s), i.e. events which satisfy the anchorability requirements (see Section 3.1 for details) and participate in an NC relation. Timex anchors are chosen on a transparency basis (i.e. granularity and nature of the timex), whereas Event anchors are chosen according to their relevance and salience for the timeline.

Two sub-types of NCs can be identified:

- **Temporal Containers (TCs)**: they correspond to the timexes in the text which clearly anchor the events in analysis on a timeline; the relation can hold both at intra- and inter-sentence level. Example 2 from our annotated corpus shows a timex (*2001*) and the events it anchors (e1–e4):

  2. [...] *la Sonata* **composta**$_{[e1]}$ *nel* **2001**$_{\mathbf{[TCanchor]}}$, *il cui primo esecutore* ***fu***$_{[e2]}$ *lo stesso Lucchesini. In questa* ***esecuzione***$_{[e3]}$ *si* ***ritrovavano***$_{[e4]}$ *già tutte le doti musicali di Lucchesini [...].*

- **Event Containers (ECs)**: they correspond to event mentions which function as a temporal anchor for other event mentions. ECs can be useful in cases where no anchoring timex is available or to model event-subevent relations. Example 3 shows a sentence with no explicit temporal expression, where the anchoring of events (e1–e3) is possible only with respect to the event (*ricognizione*).

  3. *[...] Durante la* **ricognizione**$_{\mathbf{[ECanchor]}}$, *il tenente ha dato* ***disposizioni***$_{[e1]}$ *per il presidio, e nella* ***fase***$_{[e2]}$ *iniziale ha* ***ordinato***$_{[e3]}$ *ai sottoposti di fare rapporto al campo base.*

Figure 1 serves as a visual representation of the NC as annotated in example 2. By means of NCs, a document timeline will result in an ordered succession of NCs rather than of isolated events. This is the NC resulting from the following sentence, taken from the annotated corpus.

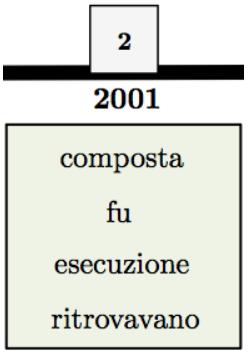


Figure 1: Visual representation of a NC for the sentence in Example 2.

Naturally, the NC represented here is only a visual aid picturing the conceptual outcome of

applying CONTAINS relations between the anchor (here, the TIMEX3 *2001*) and anchored elements (here, EVENTs *composta, fu, esecuzione,* and *ritrovavano*)

3.1 Event Anchorability Requirements

The set of events which can be anchored has been restricted to factual events. The identification of eligible anchorable events has been manually conducted at this stage of the annotation. We adopted the definition of **factuality** as proposed in the FactBank (Saurí, 2008) and which is based on the double axis of *polarity* (positive *vs*. negative) and *certainty*. For the sake of our annotation task, only positive and certain events can be anchored. Events in the future were generally not annotated as they normally do not have a certain status. However, those events with an established schedule (e.g. deadlines, meetings), or whose future temporal window is assumed to be certain, such as festivities, have been annotated in anchoring relations as well.

We excluded all events which are presented as **subjective** (i.e. judgements, opinions). In example 4, *esplosione* is a factual event and was anchored as such, whereas *sbagliato* describes it through the grid of the writer's judgement, who states that the explosion was a mistake, and thus not anchored.

4. ***L'esplosione***$_{e1}$ *è avvenuta a mezzanotte di lunedì [...]. Insomma, gli attentatori hanno* ***sbagliato***$_{e2}$ *obiettivo.*

Finally, **generic** events, i.e. events which acquire some kind of attributive value towards discourse participants, expressing persistent properties or reiterated, habitual activities, were not anchored.

4 The EVENTI-NC Corpus

The EVENTI-NC corpus includes documents from both the training and the test sections of the Main task of the EVENTI corpus. It includes 58 annotated articles, for a total of 24.259 tokens, covering roughly 11% of the EVENTI corpus; Table 1 shows the number of EVENTs and TIMEX3 involved in our annotation.

Table 2 reports the number of annotated containers in our corpus, and their distribution according to their type. TCs make up for almost 63% of the total number of NCs, against the 37% of ECs.

General EVENTI-NC statistics	
Annotated tokens	24.259
Annotated articles	58
EVENT markables	3645
TIMEX3 markables	595

Table 1: Overview of corpus statistics.

Annotated NCs		
Type	Number	%
ECs		
Verbal anchors	61	19.5
Nominal anchors	55	17.6
Total EC n.	**116**	**37.1**
TCs		
Text-consuming TIMEX3s	160	51.1
Empty TIMEX3s	37	11.8
Total TC n.	**197**	**62.9**
Total NC n.	**313**	

Table 2: Distribution of Narrative Containers in the corpus.

It is interesting to notice that 11.8% of TCs is realized by empty TIMEX3s, i.e. temporal expressions which do not correspond to lexical items but can be inferred and which are necessary to for assigning a correct value to a timex.

4.1 Distribution of Narrative Containers anchors

We conducted an in-depth analysis of the NC anchors following two parameters: i.) the properties of NC anchors on their own; and ii.) the sensitivity to the document genre, i.e. the news domain, on the line of Pustejovsky and Stubbs (2011).

Concerning the first parameter, we first investigated the incidence of verbal anchors as opposed to nominal anchors. Whereas there appear to be no tendency towards verb or nouns being more likely to anchor other events, it is interesting to take a look within these categories. Out of all the verbal anchors, 42.9% are reporting verbs or verbs employed in a declarative context. We observed that there is a preference for ECs to correspond to the the event with the highest degree of topicality in the article, or the most important event (climax

event). For example, one article[4] reports on President Clinton's surgery in 2004: the largest EC in the document is anchored by *intervento* (*surgery*), with a total of 12 anchored items.

Sensitivity to text genre can be easily observed with TC anchors. 25% of them anchor events in a timespan that can be measured as ± 1 day with respect to the DCT. Anchors for these containers are mostly represented by non absolute temporal expressions, such as temporal adverbs (e.g. "ieri" [yesterday], "domani" [tomorrow], among others) and by the DCT itself, which represents 11% of the TC anchors.

Genre-sensitivity might also be the factor behind the average number of NCs in the corpus. The documents have an average of 5.17 NCs, and even for more lengthy articles, the textual anchors were rarely more than 7. The average of 5.17 NCs/article might be due to the fact that newspaper articles usually refer to a limited number of facts, whose core is usually made of a handful of recent happenings; whereas the fluctuating relationship between length and NC number usually depends both on the content of the article and on the granularity of the selected NC.

5 Conclusion

This paper reports on a first proposal of an annotation scheme and accompanying annotated data for NCs in Italian news articles. The NC annotation is an additional layer on top of already available data for Temporal Processing in Italian. It addresses pending issues (e.g. the annotation of the temporal relations between event and the Document Creation Time) and increases the informativeness of the document timelines. Overall, we observed that there is a preference for NC to be realized by timexes in a limited time span. However, NCs may also be realised by events. In this case, nouns and verbs have a similar distribution with a preference for events which have a central role in the news or facilitate the clustering of the information (e.g. reporting events). Such a behaviour is different with respect to clinical narratives where nominal events are more frequently selected as NC (Bethard et al., 2016). This suggests that different text genres present different ways of organizing events on a timeline. The introduction of the factuality parameter to select the anchoring events is a strategy to clean timelines and to move Temporal Processing from a single document to a cross-document task.

Future work will aim at assessing the reliability of the proposed scheme via an inter-annotator agreement study and at completing the annotation of the entire EVENTI corpus. Finally, the annotated data and guidelines are publicly available [5] to encourage additional testing and experiments.

[4] adige20040709_id405401.txt

[5] https://sites.google.com/site/ittimeml/documents/narrative-container-data.zip?attredirects=0&d=1

References

Steven Bethard, Nathaniel Chambers, Bill McDowell, and Taylor Cassidy. 2014. An annotation framework for dense event ordering. In *52nd Annual Meeting of the Association for Computational Linguistics*, 52, Baltimore, MD, USA, June. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, James Pustejovsky, and Mark Verhagen. 2015. SemEval-2015 Task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814. Association for Computational Linguistics.

Steven Bethard, Savova Guergana, Leon Derczynski, Wei-Te Chen, James Pustejovski, and Mark Verhagen. 2016. SemEval-2010 Task 13: TempEval-2. In *Proceedings of SemEval-2016*, SemEval '13, pages 977–987, Stroudsburg, PA, USA. Association for Computational Linguistics.

André Bittar, Pascal Amsili, Pascal Denis, and Laurence Danlos. 2011. French Timebank: an ISO-TimeML annotated reference corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 130–134. Association for Computational Linguistics.

Tommaso Caselli, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. 2011. Annotating Events, Temporal Expressions and Relations in Italian: the It-TimeML Experience for the Ita-TimeBank. In *Proceedings of the Fifth Law Workshop (LAW V)*, pages pages 143–151, Portland, Oregon. Association for Computational Linguistics.

Tommaso Caselli, Rachele Sprugnoli, Manuela Speranza, and Monica Monachini. 2014. EVENTI: EValuation of Events and Temporal INformation at Evalita 2014. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*. Pisa University Press.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2012. Annotating narrative timelines as temporal dependency structures. In *Proceedings of the International Conference on Linguistic Resources and Evaluation*.

Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. 2015. SemEval-2015 Task 5: QA TempEval—Evaluating temporal information understanding with question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 792–800.

Anne-Lyse Minard, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, Ruben Urizar, and Fondazione Bruno Kessler. 2015. Semeval-2015 task 4: Timeline: Cross-document event ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 778–786.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160. Association for Computational Linguistics.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The TimeBank corpus. In *Corpus linguistics*, volume 2003, page 40.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An international standard for semantic annotation. In *Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Roser Saurı and Toni Badia. 2012. Spanish TimeBank 1.0. *LDC catalog ref. LDC2012T12*.

Roser Saurí. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. thesis, Brandeis University.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Pustejovsky, and James Allen. 2013. Semeval-2010 task 13: Tempeval-2. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, SemEval '13, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 Task 15: TempEval Temporal Relation Identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 75–80, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 57–62, Stroudsburg, PA, USA. Association for Computational Linguistics.

Increasing information accessibility on the Web: a rating system for specialized dictionaries

Valeria Caruso*, Anna De Meo*, Vincenzo Norman Vitale°
*Università degli Studi di Napoli 'L'Orientale',
°Università degli Studi di Napoli Federico II
vcaruso@unior.it, ademeo@unior.it,
vincenzon.vitale@studenti.unina.it

Abstract

English. The paper illustrates the features of the *WLR* (Web Linguistic Resources) portal, which collects specialized online dictionaries and asses their suitability for different functions using a specifically designed rating system. The contribution aims to demonstrate how the existing tool has improved the usefulness of lexicographical portals and how its effectiveness can be further increased by transforming the portal into a collaborative resource.

Italiano. *Questo contributo descrive le caratteristiche del portale WLR (Web Linguistic Resources) che raccoglie dizionari specialistici della Rete e ne stima l'utilizzabilità per diverse funzioni, avvalendosi di uno specifico sistema di valutazione. Viene quindi mostrato come questo strumento incrementi l'utilizzabilità dei portali lessicografici finora sviluppati e come la sua efficacia possa essere ulteriormente migliorata trasformandolo in risorsa collaborativa.*

1 Introduction

This paper sketches out the current features and an upcoming new application of a rating system designed to assess online specialized dictionaries. The system evaluative parameters are managed through a relational database accessible for free online at the *Web Linguistic Resources* (*WLR*) site. These parameters are used to identify the best available dictionaries to satisfy different types of information needs experienced by the Internet surfers, while the assessment procedure has been designed to be flexible and can be readapted to estimate the supportive value of other resources as well, like grammars or corpora. On the other side, once the score assignment for each dictionary feature has been decided, grades are given automatically by the database.
The assessment procedure is straight and strictly operationalized (Swanepoel, 2008, 2013), and it can be used as a guided process to collect data provided by the users themselves. The system is in fact going to be updated and transformed in a collaborative (Carr, 1977) dictionary portal, collecting forms that have been filled in by the Web surfers themselves.

2 Information overload on the Internet

The *WLR* dictionary portal has been designed as a tool that can offer assistance to solve different problems concerning specialized knowledge and lexicon that Web users might experience on different occasions in their lives. For example, if they need to understand specific concepts belonging to some technical fields, like a journalist who needs to acquire specific information about different topics during his/her professional activity. Or translators, who need both concise explanations of concepts and cross linguistic correspondences in order to understand specialized texts and translate them. Dictionaries can offer, in fact, proper assistance in a wide variety of different occasions, provided that they are reliable and efficient tools. The enormous inventory of specialized online dictionaries counts already reference works for top professionals in one field, like the authoritative *The New Palgrave Dictionary of Economics*, but also different hybrid[1] tools addressed to school children, like the entertaining *Math Spoken Here!*, which has been conceived to assist in learning and homework activities.

[1] For the concept of hybridization in electronic lexicography, see Granger 2011.

Surfing the Web it is possible to experience the tremendous amount of specialized dictionaries that are available for the most different fields. Compared to these resources, the number of general language vocabularies is but a few drops in the ocean. This state of affairs is however unsurprising, since similar disproportions were the rule in the paper dictionary era (Tarp, 2010), when vocabularies were not so easily accessible and one could not directly experience the real composition of the lexicographical production. The availability of these resources on the Internet has however overturned the proportion between the user, who is in need of lexicographical assistance, and the number of specialized resources he can consult, thus causing such an information overload that the user is either forced to resort to one of the usual *Wikipedia* pages, or to abandon the search completely. In both cases the user is stressed by the demanding activity of finding a source of information, rather than solving his/her information voids.

3 Solutions for integrated information access

Information overproduction on the Web has become one of the tasks of electronic lexicography since the advent of the first metalexicographical sites, called 'dictionary collections' (Engelberg and Müller-Spitzer, 2013), offering lists of links to different dictionaries. This practice has rapidly evolved into steadier solutions that have served also the opposite aim of a controlled integration of lexicographical data, made possible by the 'dictionary portals' (Engelberg and Müller-Spitzer, 2013) of well-established publishing houses, which have implemented the integration among their vocabularies in order to better meet the information needs of their users. In the *Pons* or *Cambridge* dictionary sites, for example, it is possible to access different vocabularies by filling in a single search mask and selecting the desired resource from a menu.

According to Engelberg and Müller-Spitzer (2013), dictionary portals "have followed [the] course from the single lexicographic product to the general lexicographic information service" that was predicted by Arnold (1979) and Kay (1983) as far as thirty years ago, thus creating a new type of dictionary. The possibility to cross-link well-structured informative resources, such as dictionaries, has in fact broaden the possibility of users to be informed promptly, by querying a single search engine that gives access to many dictionaries.

The right of ownership to the inventoried dictionaries is one of major restrictions determining the kind of access to the lexicographical information, thus influencing the portal typology. In the classification proposed by Engelberg and Müller-Spitzer (2013), dictionaries issued by the same publishing house may form 'integrated dictionary nets', if every vocabulary has been compiled with "a common concept of data modelling and structuring", thus allowing users to retrieve lemmata with similar properties from the different dictionaries inventoried, such as in the OWID. On the contrary, portals having no rights of ownership to the dictionaries, called 'dictionary collections', generally offer simple lists of links to external resources. Only a few of them are also provided with query systems that carry out searches in the lemma lists or in the whole text of the inventoried resources (see *OneLook*).

3.1 The *WLR* database assessment system

In addition to the types listed by Engelberg and Müller-Spitzer (2013), the *WLR* site increases the typologies of 'dictionary collections' by offering inventories of vocabularies that have been evaluated on the basis of the kind of data they contain (Caruso & De Meo, 2014). The assessment is carried out by a multi-parametric searchable database, which inventories dictionary features and assigns scores in order to display lists of resources that are more suited for two different types of parameters. It is in fact possible to search for dictionaries assisting with specific tasks, or 'lexicographical functions' that the dictionary should be able to fulfill (Tarp 2008), like acquiring new knowledge on a specific topic, solving communicative issues, or giving assistance with translations or learning tasks. These parameters can be set in *WLR* database by choosing the corresponding option in the 'Kind of assistance' box of the search form. Additionally, the user can set his/her level of expertise in the specialized field considered, and thus select the layman, semi-expert or expert profile in the 'Expertise level' box.

The rating system used in the *WLR* site is intended to increase the effectiveness and efficacy of portals, making dictionary collections less time wasting and more useful also for the less experienced dictionary users, since they avoid the display of "long lists" that show "results from trustworthy sources and downright amateurish concoctions all mixed up" (de Schryver 2003: 157). The evalua-

tion system relies in fact on the presence or absence of 58 types of data, addressing all the component parts of dictionaries (Caruso & De Meo, 2014): from the host site and the general organization (or macrostructure), to the mediostructure and microstructure, for which both linguistic and encyclopedic data are taken into consideration. Additionally, explicit guidelines are followed for the score assignment system: characterizing data for a specific parameter receive one or two points, according to their degree of relevance. Negative scores (-1, -2) are instead given to contradictory data. Similarly, each lexicographical parameter considered ('Kind of assistance' and 'Expertise level') can reach the same maximum score: for example, the different types of users may have no more than 24 points. In the meanwhile, for contradictory profiles, such as laymen and experts, the score distribution cannot be the same.

All this things considered, one can affirm that the *WLR* site aims to support different types of users decreasing the information overload that occurs while consulting rich inventories of non-integrated resources, such as dictionary collections sites. Additionally, the *WLR* rating system is in line with the parameters identified by Swanepoel (2008; 2013) to carry out dictionary evaluations that are scientifically grounded, i.e. assessments that explicitly state the analytic principles they use and the way these are applied, together with instructions to measure the compliance or non-compliance to these same principles.

Additionally, the portal wires together fragments of the huge repository of specialized knowledge available on the Internet (Caruso 2014), hosting dictionaries of around 60 different fields, such as oenology, mathematics and medicine.

4 How to make effective searches

Recent studies have underlined that electronic dictionaries are special types of information systems (Tarp, 2008; Bothma, 2011; Gows, 2011; Heid, 2011) and evaluative parameters borrowed from the Information Science are used in the literature on electronic lexicography topics. In particular, the quality of one dictionary can be assessed on the basis of its usefulness for a task completion, like finding a specific collocate while writing a text. Therefore, the dictionary is considered to be effective if it provides "the right data and the right amount of data to the user" (Heid 2011: 290). On the contrary, it is efficient if gives quick access to the data needed.

The *WLR* database developed so far assures that the search for a dictionary is less time wasting for the user but it does not guarantee that the data provided by one dictionary are correct or correctly stated. Contrarily, the quality of data is always paramount, and users' searches would be more effective if they could avoid to consult vocabularies whose data are unreliable.

For example, the following Spanish oenological dictionary (*Infoagro.com - Diccionario del vino*) explains that 'ácido' is a "green wine" whose colour seems to be a consequence of a bed fermentation:

> [1] "Ácido: Vino verde. Producto de una mala fermentación maloláctica, una uva en mal estado o recolectada antes de tiempo."

On the contrary, many other dictionaries explain the same term as denoting a sour wine, or a wine that is high in acidity, like in following entry (*Diccionario del vino.com*):

> [2] "Ácido: 1.- Vino cuya acidez sobrepasa la media de la región. La acidez puede ser debida a un exceso de ácidos organicos o a un desequilibrio entre los sabores del vino.
> 2.- Vinos con PH inferior a 3,2"

In order to carry out more efficient searches using the current release of the *WLR* database, one can look for dictionaries compiled exclusively by authoritative institutions, thus restricting the search to 'Institutional' and 'Specialized' host sites, two features that users can select in the database search form. However, even the dictionaries edited by the most authoritative institutions offer examples of bad explanations that can be misleading for the user, or even difficult to interpret (Caruso & De Meo, 2014). For example, the *Talking Glossary of Genetics*, published by the National Human Genome Research Institute, in the *Chromosome* definition explains that: "Humans have 23 pairs of Chromosomes (…), and one pair of sex chromosomes, X and Y". Stated this way the definition is incorrect, since only male humans have an XY pair of chromosomes, while females have an XX pair. Effective lexicographical definitions should obviously provide more complete descriptions and should avoid incorrect generalizations like this.

Assessing data quality poses however many methodological and theoretical problems regarding the

terms and the definition features that must be rated (see Caruso & De Meo, 2014) by the system. For example, the number of the assessed lemma must remain the same despite the number of dictionary entries? Which definition features are suited to estimate whatever concept belonging the specialized fields as different as, for example, figurative arts and finance? Furthermore, at least one expert for each specialized field considered should verify the information provided, which is probably the most serious obstacle to future developments of the project. However, a completely different solution has been imagined, as will be shown in a moment.

4.1 The database as a data validation tool

The *WLR* database has been conceived as a flexible tool that allows its administrators to add or change labels in the three component parts that make up the repository system, which are called 'categories', 'features' and 'rating system'. The first component, or 'category', lists the types of inventoried linguistic resources: only dictionaries have been assessed so far, but other supportive instruments to solve linguistic issues could be added to the database, like corpora or grammars. To each category the administrator assigns different descriptive features, which is the second component of the rating system, and can be both binary or multivalve. The 'dictionary' category has 58 feature (see Caruso, 2014 for a complete list), some of them can only be present or not, thus are binary, like *Cultural Notes*, others are multivalue and thus need further specifications, like the *Kind of Dictionary*, which must be set choosing among different choices: *Monolingual dictionary*, *Monolingual word list*, *Multilingual dictionary*, *Multilingual word list*, *Plurilingual dictionary*[2]. Lastly, grades are assigned to each of these values according to the methodology described above. The administrator can decide to set different evaluative parameters for each category taken into account: for example, if grammars were added to the repository, the language proficiency level could be a suitable evaluation parameters for it.

Once however that the grades distribution has been decided, the database assigns points automatically and independently from any actions performed by the compiler of the evaluation forms, who can set only the values of the different features. Likewise, if the score assignment is changed, the inventoried dictionaries will immediately change their evaluations. The automatization of grades assignment guarantees no errors in the final score computation, however, the selection of values that describe the dictionary features are of crucial relevance for the accuracy of the evaluation.

Under this respect, the inventoried resources must be analysed carefully, because most of the times specialized online dictionaries lack strict lexicographical organization and display different data types unsystematically: for example, basic information on the word form might be given exclusively in some of the entries of one dictionary, independently of any significant paradigmatic variation of the language considered. For similar cases, the compiler must set the 'sometimes' value in the corresponding feature of the evaluation form, and the record of the data that are sporadically given by the dictionary will make the evaluation procedure more reliable.

Actually, the current development of the project is improving the existing database components with an additional part that keeps track of where unsystematic data, like those mentioned above, are present in the dictionary. This addition will make the assessment procedure extremely reliable, since the less evident features can be registered, making the evaluation accuracy easily verifiable.

With this new database component, the evaluation forms will be fillable by anyone and the *WLR* database will become a collaborative portal. This, hopefully, will make the number of the inventoried resources increase, and it will offer other additional developments.

While compiling the forms, in fact, users could also contribute to verify the quality of the data provided, signalling for each dictionary feature if any wrong information is given. For each inconsistency the user should indicate one alternative data and the source of information from which this was driven. On the other hand, the database will offer warning signals that indicate the presence of problematic data within one dictionary.

Acknowledgements

We wish to thank Gianluca Monti for managing the first version of the *WLR* database and site.

The present research has been sustained by academic grants from the University of Naples 'L'Orientale'.

[2] For the concept of Plurilingual Dictionary, see Caruso, 2011.

References

Arnold, D. I., 1979, "Synonyms and the College-Level Dictionary", Dictionaries, 1: 103–12.

Bothma, T.J.D, 2011, "Filtering and Adapting Data and Information in an Online Environment in Response to User Needs", in Fuertes-Olivera, P.A., Bergenholtz, H. (Eds.), 71-102.

Carr, M., 1997, "Internet Dictionaries and Lexicography", International Journal of lexicography, 10/3: 209-230.

Caruso V., 2011, "Online specialised dictionaries: a critical survey", in Kosem I., Kosem, K. (eds.) *Electronic lexicography in the 21st century: new applications for new users. Proceedings of eLex 2011*, Ljubljana: Trojina, Institute for Applied Slovene Studies, 66-75.

Caruso V., 2014, "A Guide (not only) for Economics Dictionaries", Hermes – Journal of Language and Communication in Business, 52: 75-91.

Caruso, V., De Meo, A., 2014, "A Dictionary Guide for Web Users", in Abel, A., Vettori, C. & Ralli, N. (eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus*, Bolzano: EURAC, 1087-1098.

De Schryver, G. M., 2003, "Lexicographers' Dreams in the Electronic-Dictionary Age", International Journal of Lexicography, 16/ 2: 143-199.

Engelberg, S., Müller-Spitzer, C, 2013, "Dictionary Portals", in Gouws, R. H., Heid, U., Schweickard, W., e Wiegand, H. E. (eds.), *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*. Berlin/New York: de Gruyter, 1023-1035.

Fuertes-Olivera, P.A., Bergenholtz, H. (Eds.), 2011, *e-Lexicography: The Internet, Digital Initiatives and Lexicography*, London, New York: Continuum.

Gouws, R.H, 2011, "Learning, Unlearning and Innovation in the Planning of Electronic Dictionaries", in Fuertes-Olivera, P.A., Bergenholtz, H. (Eds.), 17-29.

Granger, S., 2012, "Introduction: Electronic Lexicography from Challenge to Opportunity", in Granger, S. & Paquot, M. (Eds.), Oxford: OUP, 1-11.

Heid, U. 2011. 'Electronic Dictionaries as Tools: Towards an Assessment of Usability.' In P. A. Fuertes-Olivera and H. Bergenholtz (eds.), 287–304.

Kay, M., 1983, "The dictionary of the future and the future of the dictionary", in Zampolli, A. & Cappelli, A. (eds.), *The Possibilities and Limits of the Computer in Producing and Publishing Dictionaries: Proceedings of the European Science Foundation Workshop*, Pisa: Giardini Editori, 161–74.

Swanepoel, Piet, 2008, "Towards a Framework for the Description and Evaluation of Dictionary Evaluation Criteria", Lexikos, 18: 207-231.

- 2013, "Evaluation of dictionaries", in Gouws, R. H., Heid, U., Schweickard, W., e Wiegand, H. E. (a cura di), *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*. Berlin/New York: de Gruyter, 587-596.

Tarp, S., 2008, *Lexicography in the borderland between knowledge and non-knowledge*, Tübingen: Niemeyer.

- 2010, "Beyond Lexicography: New Visions and Challenges in the Information Age", in Bergenholtz, H., Nielsen, S. & S. Tarp (eds.), *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*, Bertlin et.: Peter Lang, 17-32.

Online Dictionaries and resources

Cambridge Dictionaries, http://dictionary.cambridge.org/it/, accessed July 2016.

Diccionario del vino.com, http://www.diccionariodelvino.com/index.php/letra/a/.

Infoagro.com - Diccionario del vino, http://www.infoagro.com/viticultura/diccionario/diccionario.htm, accessed July 2016.

Math Spoken Here! An Arithmetic and Algebra Dictionary, http://www.mathnstuff.com/math/spoken/here/, accessed July 2016.

OneLook, www.onelook.com, accessed July 2016.

OWID (*Online-Wortschatz-Informationssystem Deutsch*), http://www.owid.de/, accessed July 2016.

Pons, www.pons.eu, accessed July 2016.

Talking Glossary of Genetics, https://www.genome.gov/glossary/, accessed July 2016.

The new Palgrave dictionary of economics, Basingstoke and New York: Palgrave Macmillan, http://www.dictionaryofeconomics.com, accessed July 2016.

Web Linguistic Resources (*WLR*), www.weblinguisticresources.org, accessed July 2016.

Parsing di corpora di apprendenti di italiano: un primo studio su VALICO

Elisa Corino[1]
Università di Torino
elisa.corino@unito.it

Claudio Russo
Università di Torino
clrusso@unito.it

Abstract

English. Modern learner corpora are now routinely PoS tagged, whereas syntactic parsing is much less frequent. This paper proposes a first attempt of parsing applied to a subcorpus of VALICO, in an effort to identify key elements to be further used to parse corpora of Italian as a foreign language in a proper way.

Italiano. *I moderni corpora di apprendenti sono ormai abitualmente annotati per PoS. Meno frequente è invece il parsing sintattico delle varietà di apprendimento. Questo contributo propone un primo tentativo di parsing applicato a un sottocorpus di VALICO, nel tentativo di individuare elementi chiave che possano servire a tracciare una rotta per l'etichettatura sintattica di corpora di italiano come lingua straniera/seconda (LS).*

1 Introduzione

È ormai prassi lemmatizzare e annotare i corpora di apprendenti per Parts of Speech (PoS) e molti sono gli esempi di etichettatura degli errori; non altrettanto diffuse sono invece le esperienze di parsing e annotazioni sintattiche di learner corpora.

Le ragioni sono molteplici, ma certamente riconducibili all'estrema imprevedibilità e marcatezza che caratterizza l'interlingua degli apprendenti. Nonostante la complessità dell'operazione, che richiede necessariamente una buona dose di intervento manuale, il parsing di un *learner corpus* può avere ricadute positive in più di un ambito: oltre al contributo per i contesti che si occupano in modo precipuo di linguistica computazionale, esso diventa estremamente utile per le ricerche su acquisizione e apprendimento, perché permette di individuare più facilmente errori, deviazioni dalla norma e distribuzione di categorie e strutture sintattiche altrimenti difficili da far emergere interrogando un corpus etichettato unicamente per parti del discorso.

Applicare processi di parsing alle varietà di apprendimento non è certo una novità: il trattamento sintattico delle interlingue degli apprendenti è da tempo il focus di sistemi di Computer Assisted Language Learning (CALL) e Intelligent Language Tutoring Systems (ILTS) (Vandeventer Faltin, 2003; Heift & Nicholson, 2001; Amaral & Meurers, 2011). In questi casi la violazione delle regole sintattiche, come ad esempio il mancato accordo tra soggetto e verbo, diventa spia della presenza di un errore da segnalare a chi produce il testo. Menzel & Schröder (1999) usano parametri che descrivono le dipendenze all'interno delle lingue degli apprendenti, alle quali vengono applicati vincoli pesati e procedimenti di *robust parsing* per definire la deviazione dalla norma.

La distanza da strutture canoniche e non marcate va tuttavia definita a partire da uno standard al quale le varietà di apprendimento vengono ricondotte. Ecco perché il parser del *learner corpus* deve essere preliminarmente applicato ad un corpus di nativi – possibilmente comparabile a quello degli apprendenti. Una prassi

[1] Il contributo è il risultato della collaborazione dei due autori; tuttavia a Elisa Corino vanno attribuiti i §§ 1, 2, 4 e 5; a Claudio Russo il § 3. Si ringraziano Cristina Bosco e Alessandro Mazzei per il prezioso aiuto e per aver messo a nostra disposizione gli strumenti elaborati dal loro gruppo di ricerca.

diffusa è poi quella di riportare le occorrenze errate ad una forma target, un procedimento manuale in genere usato da quei learner corpora annotati anche per errori. Alcuni tentativi sono stati fatti fino ad ora soprattutto per il tedesco (Nivre et al., 2007, Lüdeling 2008, Ott & Zai 2010), ma il terreno pare ancora inesplorato per quanto riguarda il panorama italiano.

Questo contributo si propone come un primo tentativo di applicare procedure di parsing a un corpus di apprendenti di italiano come lingua straniera, per definirne criticità e vantaggi sia dal punto di vista computazionale, sia rispetto allo studio delle varietà di apprendimento.

In particolare verranno presi in considerazione alcuni dati estratti dal corpus VALICO[2], selezionati a partire da sottocorpora definiti in base alla L1 degli apprendenti. Per questo studio pilota si è scelto di comparare i risultati del trattamenti di testi di apprendenti ispanofoni e germanofoni, con l'intento di individuare peculiarità sintattiche delle interlingue di discenti provenienti da lingue tipologicamente diverse: tipicamente romanza – e quindi vicina all'italiano l'una, rappresentativa dell'area germanica l'altra. Sono stati sottoposti al parser 12 testi (quattro per ciascuna delle prime tre annualità di studio di italiano) estratti in modo casuale dal sottocorpus germanofono e 12 testi derivati dal sottocorpus ispanofono, secondo gli stessi criteri di selezione. Per il gruppo tedesco sono state processate in totale 126 frasi, per il gruppo spagnolo 78.

Fine ultimo è stabilire quali deviazioni dalla norma dell'uno e dell'altro gruppo non sono etichettate correttamente, in modo da tracciare un percorso che possa portare alla definizione di regole di etichettatura per allenare il parser e migliorarne le prestazioni su questa varietà di lingua.

VINCA, il corpus di nativi appaiato a VALICO, servirà da corpus di riferimento per il parser a dipendenze sviluppato tra gli strumenti del Turin University Linguistic Environment (TULE).

2 Il corpus VALICO

VALICO è un corpus di scritti di apprendenti di italiano LS liberamente consultabile online (Marello et al. 2011, Marello & Corino i.s.), etichettato per PoS con il TreeTagger dell'IMS di Stoccarda e codificato in CQP (Heid 2007, 2009).

L'architettura del corpus permette di applicare le query a sottosezioni selezionate in base ai metadati contenuti nella *Header*, inseriti in una base dati (Colombo i.s.), tra questi la L1 degli apprendenti, l'annualità di studio di italiano, il luogo di produzione del testo, le altre LS conosciute.

Gli stimoli iconici dai quali è stato elicitato VALICO sono le stesse vignette somministrate agli autori di VINCA[3]; i due corpora sono quindi appaiati e comparabili per lessico, strutture sintattiche, organizzazione testuale (Corino & Marello 2009), si veda ad esempio la somiglianza tra (1a) e (1b) in relazione alla Fig. 1:



Fig. 1

(1a) […] la donna , arrabbiata , impreca contro di me dicendomi che quello che avevo appena picchiato è il suo ragazzo ! [VINCA]

(1b) Ha cominciato a sgridarme per farle una brutta passata e quello che riteneva molto grave era che avevo picchiato al suo ragazzo. [VALICO, L1 Spagnolo]

VINCA si presta quindi a fungere da corpus di controllo e modello per il parsing di VALICO.

In questa prima fase esplorativa delle modalità di etichettatura sintattica del corpus, è stato deciso di non ricorrere a ipotesi esplicite sul target delle produzioni degli apprendenti, così come avviene invece per altri learner corpora (ad es. FALKO, Lüdeling et al 2012). L'unica operazione di riconduzione a forme standard è stata fatta per identificare PoS e lemma di lessemi scorretti dal punto di vista ortografico, forme "creative" sia dal punto di vista lessicale che morfologico.

Già le forme segnalate dal TreeTagger come *unknnown* erano state corrette manualmente e ricondotte a PoS e lemma appropriato in relazione anche al co-testo nel quale esse occorrevano. Infatti è possibile interrogare VALICO, nella sua attuale forma, per lemmi e ottenere anche quelle occorrenze che morfologicamente

[2] Varietà di Apprendimento Lingua Italiana Corpus Online, liberamente consultabile all'indirizzo www.valico.org.

[3] www.valico.org/vignette

non rispondono alla norma linguistica standard (ad es. cercando il lemma *minacciare* otteniamo sì forme quali *minacciano*, *minacciato*, *minaccia*, ma anche *minaccava* e *minaciò*).

3 Dal PoS tagging al parsing

Il lavoro di disambiguazione e correzione degli *unknown* rilevati dal TreeTagger è tutt'ora in corso, anche se buona parte di essi sono già stati corretti e ad oggi tokenizzazione, PoS tagging e lemmatizzazione del corpus sono piuttosto affidabili.

Per questo primo tentativo di parsing del corpus si è tuttavia deciso di affidarsi – anche per la fase di PoS tagging - agli strumenti del TULE: un analizzatore morfologico, un tokenizzatore e un parser a dipendenze: tale scelta è giustificata dalla natura dei moduli del TULE, che, in quanto costituiti da regole di disambiguazione, non si avvalgono di addestramneti stocastici; in tale cornice, la preferenza del POS-tagger di TULE all'etichettatura già presente in VALICO ha permesso al paser di lavorare in condizioni ottimali, massimizzandone la precisione. Un'indagine a campione ha fatto emergere errori simili per entrambi gli strumenti, si veda ad esempio l'etichettatura della forma *dona* nella frase *è andato in freta ha ragiungere la sua dona è poi a preso questo uomo le ha fato male* (= donna, NOME), impropriamente riconosciuta come imperativo verbale da entrambi

dona (DONARE VERB MAIN IND PRES TRANS 3 SING) TULE
dona/VER:impe/donare è/VER:pres/essere TreeTagger

L'integrazione dei materiali già elaborati per il corpus e gli strumenti di parsing sarà necessariamente il prossimo passo nello sviluppo di questa ricerca ed eviterà di dover intervenire anche su forme che già sono state corrette nel processamento del corpus con il TreeTagger, si veda ad esempio

fidansato (FIDANSATO / NOUN COMMON M SING) VALICO TULE
fidansato (/NOM/fidanzato) VALICO 2016

Eliminerà inoltre a priori alcuni errori di annotazione sintattica e impedirà la generazione di errori dovuti al mancato riconoscimento delle forme "unknown" che bloccano la sequenza di processamento dei dati.

Ai fini del presente studio, è interessante sottolineare poi che il tokenizzatore qui utilizzato si basa su un automa deterministico a stati finiti che risulta robusto su sequenze di caratteri appartenenti a diverse lingue (tra cui rientrano inglese, spagnolo, hindi e italiano). Tale strumento compie una prima distinzione fra parole in generale (qui intese come sequenze di caratteri alfabetici), nomi propri, abbreviazioni/sigle, numeri in cifre, date in formati standard, segni di interpunzione, numerazioni di capitoli e paragrafi, anni (nelle forme contratte come '05 per 2005). Nella fase successiva di trattamento, il parsing si basa su una gerarchia precompilata di classi trasformate per generare gli alberi sintattici[4].

3.1 I dati di imprevedibilità morfosintattica

Le peculiarità dei dati linguistici sottoposti al parser hanno, come previsto, originato una serie di errori che hanno bloccato la sequenza di processamento in più momenti. I problemi non nascono soltanto in relazione alla flessione verbale o nominale, ma emergono anche laddove occorre un accumulo di clitici unitamente a forme ortografiche che deviano dalla norma. Si veda ad esempio la frase

(2) […], ma la ragazza bella è stata averecela.

Il verbo *avere* cliticizzato ha reso impossibile la disambiguazione e il tagging, anche a seguito della normalizzazione ortografica e di alcune progressive semplificazioni. È stato quindi necessario sostituire il token direttamente con il suo lemma, per permettere al TULE di procedere con il PoS tagging e, successivamente, con il parsing sintattico.

Fortunatamente, nei sottocorpora di VALICO qui considerati, tali errori si sono rivelati in quantità limitata: oltre al caso di *avercela*, sono state normalizzate in *È* tutte le istanze di *E'* e una istanza di *glielo* è stata normalizzata in *lui*[5], in un atto di correzione purtroppo invasivo e in futuro auspicabilmente evitabile.

[4] Si rimanda Lesmo (2009 e 2011) per la trattazione approfondita dell'implementazione del TULE.

[5] La frase originale recitava: "Il fratellino di Leo non capiva perchè e così lei ha spiegato glielo."

4 Prime osservazioni linguistiche: il gerundio e le preposizioni

Una prima verifica sull'etichettatura dei testi fa emergere due ordini ricorrenti di problemi che paiono essere indipendenti dalla L1 degli apprendenti: il primo è legato ad un uso peculiare del gerundio, assimilabile alle funzioni attributive delle participiali inglesi (fonte di interferenza per gli ispanofoni) e tedesche; l'altro dipende dalla presenza di preposizioni in posizione postverbale laddove invece il target richiede l'oggetto diretto.

L'uso anomalo del gerundio, come esemplificato da (2) e (3), pare ricalcato sulla struttura attributiva tipicamente inglese (participiale con la forma in *-ing*), ma possibile anche in tedesco, dove il participio specifica la condizione dell'antecedente dal quale dipende.

(2) Ieri al parco mio fratello stava leggendo il giornale quando ha visto un uomo portando una donna sopra i suoi ombrelli [L1 Spagnolo]

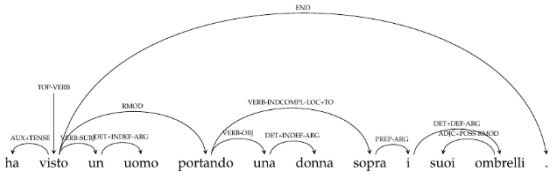


Fig. 2 Rappresentazione del segmento *visto un uomo portando una donna sopra i suoi ombrelli*

(3) Ieri al parco Giacomo è stato seduto sopra un banco nel parco e ha letto il giornale dello sport quando ha visto un uomo portando una donna che ha gridato molto forte. [L1 Tedesco]

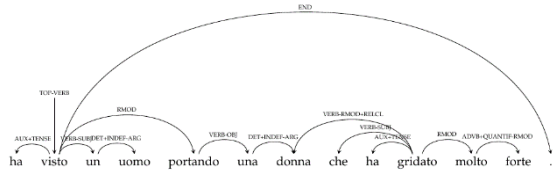


Fig. 3 Rappresentazione del segmento *quando ha visto un uomo portando una donna che ha gridato*

Come si osserva nelle Figg. 2-3[6], in entrambi casi il gerundio fa parte di una proposizione circostanziale (RMOD) al quale è attribuito un soggetto non espresso dalla struttura superficiale. Etichettare una serie di istanze verbali ravvicinate come entità circostanziali è un risultato che certo non sorprende; in casi come questi, tuttavia, il gerundio è il risultato di un'interferenza molto frequente nel corpus e non codifica la struttura rilevata dal parser.

La forma che meglio rispecchia in italiano la sfumatura attributiva dell'originale che ha causato il transfer è la frase relativa, pur considerando anche l'infinitiva come un'opzione possibile.

In VINCA effettivamente troviamo numerose occorrenze che descrivono la stessa situazione (si veda ad es. (4)), codificata in una relativa, e che come tali sono correttamente segnalate (VERBO-RMOD+RELCL, Fig. 3)

(4) Ieri al parco stavo leggendo il giornale, seduto su una panchina quando vedo passare davanti a me un energumeno che trascina a forza una donna sulle spalle. [VINCA]

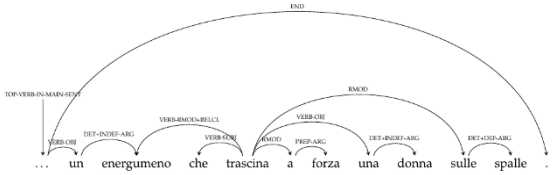


Fig. 4 Rappresentazione della frase relativa *un energumeno che trascina a forza una donna sulle spalle*

Una seconda importante causa di errore di etichettatura sintattica è la sovraestensione dell'uso delle preposizioni anche a quei casi in cui il verbo richiede l'oggetto diretto. (5) ne è un esempio emblematico:

(5) Per quello, il ragazzo di occhiali (che stava a un banco del parco) va golpeare al brutto ragazzo per liberare a la ragazza di bricci dil ragazzo che la portava. [L1 Spagnolo]

In casi come *va a golpeare al brutto ragazzo*, il parser interpreta come VERB-INDCOMP-LOC+TO il sintagma preposizionale, che invece dovrebbe essere ricondotto a un complemento diretto

[6] Le rappresentazioni degli alberi sintattici sono state generate dal software viewerTULE, implementato da L. Robaldo.

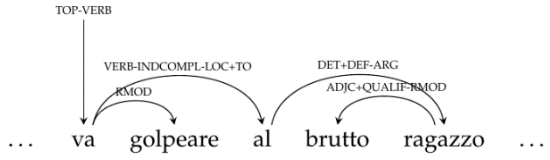


Fig. 5 Rappresentazione del segmento *va a golpeare al brutto ragazzo*

Tuttavia non si tratta di un comportamento consistente, poiché emergono anche casi come (6), in cui invece l'etichettatura è corretta (Fig. 6), nonostante l'agrammaticalità dell'enunciato.

(6) Lui stava contentissimo perche pensava che cossì aveva aiutato a la povera donna. [L1 Spagnolo]

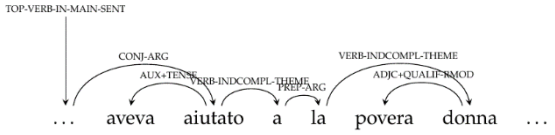


Fig. 6 Rappresentazione del segmento *aveva aiutato a la povera donna*

L'identificazione corretta della relazione sintattica sembra essere dovuta alla semantica del verbo e alla capacità del TULE di lemmatizzare correttamente la forma.

5 Conclusioni provvisorie e futuri sviluppi

L'esperienza maturata in questa prima fase del percorso di ricerca e implementazione del parsing sul corpus di apprendenti VALICO ha permesso di mettere in luce vantaggi e criticità degli strumenti utilizzati.

Dal punto di vista computazionale è emerso come sia necessario integrare risorse e definire "protocolli di allenamento" del parser orientati al trattamento di quella varietà particolare di lingua che è l'interlingua di apprendenti.

Una valutazione qualitativa ha mostrato che errori causati dall'omissione di elementi sintattici chiave per la definizione delle dipendenze porta a errori nella stessa etichettatura, così come avviene in occasione della sovraestensione di alcune forme a funzioni che esse non rivestono, o ancora della sinergia negativa tra forme morfologicamente complesse ed errori ortografici; altri errori invece sono trattati in modo aproblematico.

Le difficoltà di gestione dei sintagmi preposizionali e fenomeni diffusi di transfer linguistico soprattutto a livello di sistema verbale rivelano come l'individuazione di tendenze e la definizione di alcune indicazioni sul trattamento di tali costruzioni potrebbero incrementare notevolmente la precisione dell'etichettatura automatica.

Rispetto alla definizione di un'architettura che integri il parsing sintattico al PoS tagging, è necessario elaborare una chiara proposta di definizione di regole di etichettatura; inoltre resta da stabilire se sia veramente necessario allinearsi alle scelte operate per altri *learner corpora* in cui viene definita un'ipotesi target che segna la struttura obiettivo della produzione dell'apprendente e ne mette in luce la distanza dalla versione effettivamente riportata nel corpus.

Bibliografia

Amaral, L.; Meurers, D.(2011). On Using Intelligent Computer-Assisted Language Learning in Real-Life Foreign Language Teaching and Learning. ReCALL 23(1), 4–24. URL http://purl.org/dm/papers/amaral-meurers-10.html.

Colombo, S. (in stampa). Storia dell'architettura di VALICO. In E. Corino, C. Marello, VALICO e VINCA: corpora di apprendenti di italiano, Guerra, Perugia.

Corino, E.; Marello, C. (2009). Elicitare scritti a Partire da storie disegnate: il corpus di apprendenti Valico. In C. Andorno, S. Rastelli (a cura di), Corpora di Italiano L2: Tecnologie, metodi, spunti teorici, Perugia: Guerra.

Dickinson, M.; Ragheb, M. (2009). Dependency Annotation for Learner Corpora. In Proceedings of the Eighth Workshop on Treebanks and Linguistic Theories (TLT-8).

Heid, U. (2007). Il corpus WorkBench come strumento per la linguistica dei corpora. Principi ed applicazioni' in M. Barbera, E. Corino, C. Onesti (a cura di) (2007), *Corpora e linguistica in rete*. Guerra, Perugia, pp. 89-108.

Heid, U. (2009). Metadata for learner corpora. A case study on VALICO. In E. Corino, C. Marello (a cura di), VALICO. Studi di linguistica e didadttica. Perugia: Guerra, 151-165.

Heift, T; Nicholson, D. (2001). Web Delivery of Adaptive and Interactive Language Tutoring. *International Journal of Artificial Intelligence in Education* 12(4), 310–325

Lesmo, L. (2009). The Turin University Parser at Evalita 2009. In: Proceedings of Evalita '09, Reggio Emilia, Italy.

Lesmo, L. (2011). Use of semantic information in a syntactic dependency parser. In Magnini B., Cutugno F., Falcone M., Pianta E. (eds) "Evaluation of Natural Language and Speech Tools for Italian - Proceedings of Evalita 2011". LNCS/LNAI, Springer-Verlag

Lüdeling, A. et al. (2008). Syntactic Misuse, Overuse and Underuse: A Study of a Parsed Learner Corpus and its Target Hypothesis

Lüdeling, A. et al. (2012). Das Falko-Handbuch Korpusaufbau und Annotationen Version 2.01

Marello, C. et al (2011). I corpora VALICO e VINCA: stranieri e italiani alle prese con le stesse attività scritte. In La Piazza delle lingue L'italiano degli altri. Firenze, 27-31 maggio 2010. Atti, a cura di Nicoletta Maraschio e Domenico De Martino, Firenze, Accademia della Crusca, 2011 ("La Piazza delle lingue", 2). pp.49-61

Menzel, W.; Schröder, I. (1999). Error Diagnosis for Language Learning Systems [http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.4723&rep=rep1&type=pdf]

Nivre et al., (2007). MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. *Natural Language Engineering* 13(1), 1–41.

Ott, N.; Ziai, R. (2010). Evaluating Dependency Parsing Performance on German Learner Language. In Proceedings of the Ninth Workshop on Treebanks and Linguistic Theories (TLT-9), Tartu.

Vandeventer Faltin, A. (2003). Natural language processing tools for computer assisted language learning. In Linguistik online 17, 15/03 [http://www.linguistik-online.de/17_03/vandeventer.html]

Spammare senza pietà - Corpus based analysis of English, un-acclimatised verb loans in Italian and creation of a reference lexicon

Anna Fantini
Università degli studi di Pavia
anna.fantini01@universitadipavia.it

Abstract

English. We describe the lexical resource created to investigate the semantic changes of 90 English, un-acclimatised verb loans in Italian. Final results and interesting observations concerning the annotation task are discussed.

Italiano. *Descriviamo la risorsa lessicale creata per indagare in italiano il cambiamento semantico di 90 prestiti verbali inglesi non acclimatati. Illustriamo i risultati finali e le interessanti osservazioni emerse dall'esperimento di annotazione.*

1 Introduction

The case of language borrowing was investigated in depth by Gusmani (1983), who argues that a linguistic loan is an interference phenomenon, connected with contact and mutual influence of different languages. According to his study, the motivations behind the origin of a loan lie in the individual act of a speaker or of a group of speakers. The need to resort to a foreign alternative derives from the prestige held by the latter against an equivalent word in the mother tongue of the speaker (or from the absence all together of an alternative, as in our work: "*Se mi vede, Miki mi banna* (<to ban)" vs. *"*Se mi vede Miki mi bandisce*").

Facts show that language borrowing is particularly common among specialized languages, more so if they are linked to technical contexts.

This is extremely visible within the computer context. The main focus of this paper is the informal variety of Italian as used by communities of online video-gamers, computer experts and amateurs, forum users, etc.; a specialized language linked to technical context populated with partially integrated and un-acclimatised English verb loans.

These kinds of (mostly) lexical influences are so recent that their structure is hardly stable, and the process of integration – graphical, morphological, phonetic, and lexical – in the language is still in progress. For instance, they tend to retain the phonetic property of the original word, especially of the lexical root (to spawn > *spawnare* /spo'nare/).

The new word serves as an alternative – usually a hyponym – of an already existing term[1]. As for the concept of loan acclimatisation, the literature states that it involves the role of the new term in the target language. Therefore, Gusmani speaks of acclimatisation only with regard to the lexicon and its connection with speakers' usage: the more they familiarise with the loan, the more the latter gets acclimatised. It follows that – for very recent, scarcely integrated loans – the majority of speakers, as well as linguistic authorities, do not perceive the influence of English as an enrichment of the lexical heritage but mostly as a nuisance. If it is true that a number of reports describe the interference of English over Italian as an impoverishment, some attempts have also been made to study the less acclimatized loans themselves. It is thus of interest to examine why this kind of loan infiltrates the Italian language, how the speakers cope with the new word and what is the semantics of the loan in the target language. The aim of the present paper is to give a detailed account of how the meaning of a verb loan changes (and if it changes) in the target language and to offer a reliable source of lexical information in the form of an electronic lexicon built for the occasion. Section 2 details the method used to collect suitable data; section 3 illustrates the structure and functions of the lexicon; section 4 provides the results of our analysis as well as the annotation task performed with our data; section 5 discusses our findings and section 6 finally provides a conclusion.

2 Methods

In order to investigate the semantics of English un-acclimatized verb loans, we examined their occurrence in a monitor web-corpus created for

[1] E.g. *googlare* < to google as hyponym of *cercare.*

this purpose, following the guidelines and instructions of previous Corpus Linguistics works (Baroni and Kilgarriff, 2001; Lenci et al., 2012; McEnery, Xiao, Tono, 2006; Pomikálek, 2011; Pustejovsky and Stubbs, 2012). The corpus contains 6 transcriptions from a total of 194,07 minutes of audio material, collected with consensual but unaware recordings and then transcribed using the software Elan 4.9.6 (Wittenburg et al., 2006), plus 129 texts obtained through the Sketch Engine web-crawler, suitably set. We extracted a sum of 90 different verb lemmas (542 different word forms), for a total of 1327 occurrences. The annotation involves a POS level – limited to the sentence containing the loan – a loan-type level – describing three degrees of language integration[2] – a semantic type level[3] and a thematic role level[4]. The last two levels have been annotated using the tags proposed in Jezek and Nissim (2014) and Jezek and Vieu (2013) respectively.

Every text has been annotated using the Mae software (Stubbs 2011). An annotation task was conducted using a sample of the corpus (see section 4.2), its agreement result being only partially positive but interesting nonetheless from a linguistic point of view.

The next part of our research involved the analysis of the semantic patterns for each lemma[5], thus compiling one or more data-driven senses for every verb. The senses obtained were classified according to Verb Net's semantic class hierarchy. The assumptions underlying this investigation are grounded on Corpus Pattern Analysis (CPA) and Computational Lexicography (Hanks 2008; Hanks 2012; Jezek 2011).

Verb patterns have – in general – the following structure, where:

(1) *Spammare* 2b
Agent[PERSON] V_*spammare* (Theme[ARTEFACT | ABSTRACT]).

We have chosen all uppercase for the semantic type, and first letter uppercase for the thematic role, extended to every argument of the verb. Round brackets contain the possible optional arguments of the verb.

[2] Totally integrated, e.g. *spammare*; partially integrated (grafic), e.g. *trackare*; partially integrated (phonetics), e.g. *spawnare* /sp'nare/.

[3] E.g. Person, Artefact, Location, Abstract, etc.

[4] E.g. Agent, Patient, Goal, Source, Duration, ect.

[5] From Hanks, Pustejovsky 2005, a pattern is intended here as an argument structure with specifications of both the thematic roles and the semantic types of each argument positions.

3 The Lexicon

After extracting the semantic patterns for each lemma from the corpus, we stored the information in an electronic lexicon, built using the software Personal Lexicon 2.7.1, a language learning resource developed by Alexander Smith between 2007 and 2015. The software comes both in free and registered versions, the current lexicon has been compiled – and it will be consultable – using the free version.[6]

The lexicon is designed to give a precise account of every semantic feature and every meaning variation of the verb loans. As the reader will see observing Figure 1, each entry is characterized by the following elements (some of them pre-named in the software):



"Figure 1. The *spammare* 2b lexical entry"

- The entry citation form, with the number of the sense or of the sub-sense[7];
- The Pronunciation of the citation form;
- The Class (pre-named) as in the loan type which it belongs to (whether it is fully integrated or only partially integrated);
- The Root element, as in the lexical English root it comes from;
- The Theme (pre-named), as in the Verb Net class it was reduced to;
- The Definition box, containing the lexical definition and the verb pattern;
- The Related entries in the lexicon, all accessible through hyperlink;
- The Personal Examples, used to extract the pattern.

In figure 2 we show the Conjugations tab (pre-named) that includes all the syntactic complements of the verb and their semantic properties (thematic roles and semantic types).

[6] The resource is not yet available for public consultation..

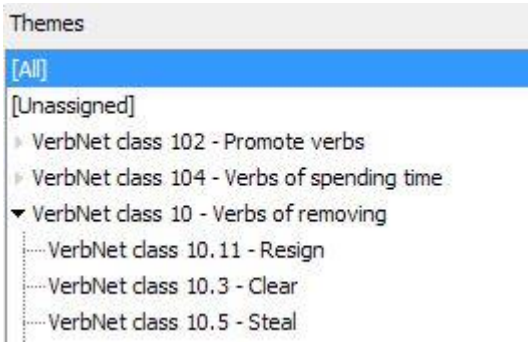
[7] Sub-senses corresponding to GRADIT's *accezioni*, the progressively numbered paragraphs a sense is pos-sibly divided into.



"Figure 2. Lexicon conjugation tab"

We listed the grammatical subject, the direct object, the indirect object and up to ten different indirect complements. Notice that the semantic type slot may specify more than one element, in which case we used the | separator.

In figure 3 we illustrate the Themes section of the lexicon, with a partial list of Verb Net classes and sub-classes used in the resource.



"Figure 3. Lexicon Theme section"

Clicking on each one produces the list of entries belonging to that particular class; this list appears in the third section of the lexicon, the Lexical Items column storing all the entries ordered alphabetically.

4 The Results

In this section we report the results of both the semantic analysis of the loans and the annotation task.

4.1 Quantitative considerations

The lexicon contains 157 senses for a total of 90 verbs. As shown in Table 1, the 157 senses have been classified into 3 groups according to three main criteria about the degree of semantic conservativeness of the loan:

1 The meaning remains the same as the original verb.

2 The meaning remains linked to the original one, but it diversifies to some degree.

3 The meaning changes to the point that it becomes a new meaning altogether

Group #	Type	Numbers
Group 1	Same sense	88
	New verb form	11
Group 2	Diversified sense	25
Group 3	New sense	26
	New v. and new sense	7
Senses	157	

"Table 1. Senses sorted according to their semantic behaviour"

Group 1 coincides with 63% of the total (99 senses out of 157), 11 senses have also new verb forms (*bishottare, autospottare*), 78 senses occur in 1 to 10 examples – they often have new verb forms (*riloggare*) or a very specific meaning (*rippare*)[8]. In this group there is the highest percentage (41%) of monosemic verbs.

Group 2 coincides with 15% of the total (25 senses out of 157), 19 senses occur only in 1 to 10 examples – they do not have a very specific meaning, but may be considered as hyponyms of Italian verbs (*farmare*2 of *sfruttare*)[9]. In this group there is the lowest percentage (16%) of monosemic verbs – their distribution proved not to be directly proportional to the senses' quantity.

Group 3 coincides with 25% of the total (33 senses out of 157), 24 senses occur only in 1 to 10 examples yet we also have the senses occurring in the highest number of examples (*droppare* 1[10] and 2[11] with 189 examples). In this group 27% of the verbs are monosemic.

4.2 The inter annotator agreement

The semantic annotation task was conducted following the methodology of Pustejovksy and Stubbs (2012); only a sample of 440 random occurrences was annotated by 9 groups of annotators, each constituted by 3 people. They were given guidelines explaining the method and the tagsets, and they were asked to separately annotate the semantic type and the thematic role of each verbal argument. We used Fleiss'*k* algorithm to calculate the agreement[12] (Artstein and Poesio, 2008), the values being interpreted according to Landis and Koch (1977). We already said that the results have been only partially positive, in particular – as for the thematic role – on-

[8] E.g. *"Non potendo accedere al CD-Rom non posso rippare niente"* .

[9] E.g. "*Ok, farmerò i campi di battaglia eterni*".

[10] E.g. "*Non molto tempo fa ho droppato i bracciali"*.

[11] E.g. "*Non è difficile droppare un computer privato"*.

[12] We choose Geertzen, J. (2012) online resource for agreement evaluation .

ly one group reached the 0.6 threshold considered acceptable with semantic annotation, the others showing moderate agreement and fair agreement (one group only). For the semantic type annotation, three groups reached the 0.6 value, four groups showed moderate agreement and two groups showed fair agreement. Nonetheless, we could make interesting linguistic considerations.

5 Discussion: the semantic behaviour of the loans

Let us consider the case of *spammare*[13]: all its three main senses are distributed among the three groups mentioned in Table 1, the semantic behaviour shows not only a certain degree of conservativeness, but also a great degree of diversification (just 18 occurrence out of 76 keep the original meaning) and thus of acclimatisation. Italian speakers apply a saving strategy: the monosemous loans are also the most conservative, while diversification often results in polysemous verbs – it seems that, once a semantic change starts, the speaker continues to use the loan until it reaches a definitive meaning, eventually becoming acclimatized. The cognitive effort behind this process is very high, but it also implies a certain linguistic confidence. Of course it is less arduous to produce a loan whose sense is strictly linked to the original verb's one, thus generating many monosemous loans. Nevertheless we wonder whether – aside from being economically convenient – is it also strategically and linguistically sensible to produce just monosemous loans instead of using semantically diversified ones. Is it sensible to keep numerous and specific loans, when there can be fewer and polysemous ones? Further investigations of English un-acclimatised verb loans may answer part of these questions.

5.1 Interesting observations about the annotation task

We feel that the only partially satisfying results may depend on the tricky lexical meaning of each loan. It is clearly easier to annotate the argument structure of a well-known verb like *potenziare*, rather than the one of the loan *overcloccare*[14] (*potenziare* and *overcloccare* being almost synonyms). The thematic role level is the most problematic, obtaining substantial agreement only in one case; the semantic type level on the other end is perceived as a less abstract, more transparent concept and the annotation is slightly better, with three groups over 0.6. What is really interesting is that the group which performed best with thematic roles is also the one which did worst with semantic types. Moreover, the groups which performed best with semantic types showed only fair to moderate agreement in thematic roles. We observe a – general and group wise – performance improvement with the semantic type level.

This is because assigning a thematic role requires a deeper reflection and some of the roles may be ambiguous (for example, Beneficiary and Goal). The creation and combination of more specific sub-types and sub-roles – targeted to this kind of verbs – could help resolve the ambiguity hindering agreement (for example, Person split into Authority and Subordinate)[15].

Furthermore the un-acclimatisation of the loans leads to somewhat different uses and different meanings among the speakers. This happens either between different communities, either between different speakers of the same community. Other significant observations emerged on the frequency of roles and types and on their co-occurrence: the most used roles are Agent (often erroneously) and Patient (often in the place of a more neutral Theme). The types most used are Person, Artefact and Abstract. The Agent-Person combination is the most frequent, even if the role is often wrongly assigned. Great uncertainty emerged in assigning the correct type to arguments whose referents are intangible informatics entities, e.g. nicknames, server, updates, etc. or characters of a game, e.g. boss, Pokémon, etc.

Last but not least, it was possible to already identify primitive verb classes, depending on the roles and the types assigned to verb arguments, i.e. verbs of change of state with a Patient role and possibly a Beneficiary role, or verbs of creation with a Result role and occasionally an Agent role.

---

[13] E.g. 1 "*Magari capiterà di spammare un oggetto"*
2 *"Iniziano a spammare pubblicità a tutti gli iscritti."*
3 *"Questa è un'abilità controversa [...] ma non va spammata"* .

[14] E.g. "*Prima di lanciare il tutto ho overcloccato la scheda video"*.

[15] From section 1 "*Se mi vede Miki mi banna*" (*bannare 1a)* Miki annotated as Agent-Authority and the personal pronoun as Beneficiary-Subordinate.

6 Conclusions

The peculiarities of each group of annotators lead to a thought-provoking analysis of the tagsets and the semantic notions themselves. The analysis of the semantic behaviour of the loans unveils deeper questions on the speaker's strategy: if it is easier to reuse a loan whose meaning already exists, why are there also loans with new or diversified meaning? Furthermore it came to our attention that yes, many loans are linked to the informatics/technical context (*rippare*, *sloggare*, etc.) but others can be considered as hyponyms of already existing Italian verbs, whose meaning is rather general (*cheattare* for *barare*, *whinare* for *lamentarsi*, etc.). It is possible to already discern synonymic and antonymic relations between the loans themselves: *craftare*, *farmare* 2, *spawnare* 1 or *sbuggare-buggare*. Finally, the most productive verb classes in our corpus are the verbs of change of state, the creation verbs and the verbs of killing.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555-596.

Joe Barcroft, Gretchen Sunderman and Norbert Schmitt. 2011. Lexis. In J. Simpson, ed. *The Routledge Handbook of Applied Linguistics*. Routledge Taylor & Francis Group. Oxon., pp. 597 – 610.

Marco Baroni. Italian tagset documentation. Available at:http://sslmit.unibo.it/~baroni/collocazioni/itwac.tag set.txt.

Marco Baroni and Adam Kilgarriff. 2001. *Large linguistically processed web corpora for multiple language.*

Raffaella Bombi. 2003. Anglicismi come banco di prova dell'interferenza linguistica. *Italiano e inglese a confronto, Firenze, Franco Cesati Editore.* 101-125.

Raffaella Bombi. 2005. *La linguistica del contatto: tipologie di anglicismi nell'italiano contemporaneo e riflessi metalinguistici.* (Vol. 11). Il calamo.

Tullio De Mauro. 1999–2007. *GRADIT - Grande Dizionario Italiano Dell'uso*. 6 vols. Torino: UTET.

Joseph L. Fleiss. 1971 . *Measuring nominal scale agreement among many raters*. Psychological Bulletin, 76:378-382.

Thierry Fontenelle. 2011. Lexicography. In J. Simpson, ed *The Routledge Handbook of Applied Linguistics*. Routledge, Taylor & Francis Group. Oxon. pp. 53 – 66.

Jeoren Geertzen. 2012. *Inter-Rater Agreement with multiple raters and variables*. Retrieved November 27, 2015, from https://nlp-ml.io/jg/software/ira/

Roberto Gusmani. 1983. *Saggi sull'interferenza linguistica (vol.1 e vol.2)*. Le lettere.

Patrick Hanks. 2012. How people use words to make meanings: Semantic types meet valencies. In A. Boulton & J. Thomas, eds. *Input, Process and Product: Developments in Teaching and Language Corpora*. Masaryk University Press.

Patrick Hanks. 2008. Lexical Patterns: from Hornby to Hunston and beyond. In E. Bernal & J. DeCesaris, eds.*Proceedings of Euralex 2008*. Barcellona, Universitat Pompeu Fabra.

Patrick Hanks and Elisabetta Jezek. 2010. What lexical sets tell us about conceptual categories. In *Lexis: E-journal in English lexicology,* 4: Corpus Linguistics and the Lexicon.

Patrick Hanks and James Pustejovsky. 2005. A Pattern Dictionary for Natural Language Processing. *Revue française de linguistique appliquée*, 10(2).

Martin Haspelmath. 2009. Lexical borrowing: concepts and issues. *Loanwords in the world's languages: a comparative handbook.*

Martin Haspelmath. 2008. Loanword typology: steps toward a systematic cross-linguistic study of lexical borrowability. *Aspects of language contact: new theoretical, methodological and empirical findings with special focus on Romancisation processes.* Mouton de Greuyter.

Elisabetta Jezek. 2011. *Lessico. Classi di parole, strutture, combinazioni*. Il Mulino, Bologna.

Elisabetta Jezek. 2010. Struttura argomentale dei verbi. In L. Renzi – G. Salvi (a cura di) *Grammatica dell'Italiano Antico*. Bologna: Il Mulino, 77-122.

Elisabetta Jezek and Malvina Nissim. 2014. *Linee guida per l'annotazione degli argomenti del verbo*

Elisabetta Jezek and Laure Vieu. 2013 . Lista di ruoli semantici per Senso Comune.

Adam Kilgarrif. 2001. Web as corpus. *Proceedings of corpus linguistics 2001*. Corpus Linguistics. Reading in a Widening Discipline.

Alan Kirkness. 2004. Lexicography. In A. Davies & C. Elder, eds. *The Handbook of Applied Linguistics*. Blackwell Handbooks in Linguistics. Blackwell Publishing, pp. 73 – 101.

Klaus Krippendorf. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications

Alessandro Lenci, Simonetta Montemagni and Vito Pirrelli. 2012. *Testo e computer. Elementi di linguistica computazionale*. Carocci editore.

Beth Levin. 1993. *English verb classes and alternations: a preliminary investigation*. University of Chicago Press.

Anthony McEnery, Richard Xiao and Yukio Tono. 2006. *Corpus based language studies. An advanced resource book*. Routledge, Taylor & Francis Group. Oxon. Routledge Applied Linguistics.

Jan Pomikálek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. thesis. Brno, Repubblica Ceca: Masaryk University. Available at: http://is.muni.cz/th/45523/fi_d/phdthesis.pdf.

James Pustejovsky and Amber Stubbs. 2012. *Natural language annotation for machine learning*. O'Reilly.

Renata Savy. 2006. Specifiche per la trascrizione ortografica annotata dei testi raccolti. In *Progetto CLIPS. Corpora e Lessici dell'Italiano Parlato e Scritto.*

Amber Stubbs. 2011. *MAE and MAI: Lightweight Annotation and Adjudication Tools.*

Michael Stubbs. 2004. Language Corpora. In A. Davies & C. Elder, eds. *The Handbook of Applied Linguistics*. Blackwell Handbooks in Linguistics. Blackwell Publishing, pp. 125 – 152.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann and Han Sloetjes. 2006. ELAN: a Professional Framework for Multimodality Research. In: *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.*

Taming Sense Sparsity: a Common-Sense Approach

Antonio Lieto, Enrico Mensa and Daniele P. Radicioni
Dipartimento di Informatica
Università degli Studi di Torino
Corso Svizzera 185, 10149 – Torino ITALY
{lieto,mensa,radicion}@di.unito.it

Abstract

English. We present a novel algorithm and a linguistic resource named CLOSEST after 'Common SEnse STrainer'. The resource contains a list of the main senses associated to a given term, and it was obtained by applying a simple set of pruning heuristics to the senses provided in the NASARI vectors for the set of $15K$ most frequent English terms. The preliminary experimentation provided encouraging results.

Italiano. *In questo lavoro presentiamo un algoritmo e una risorsa linguistica,* ClOSeSt*, che contiene i sensi più rilevanti per i* $15K$ *termini più frequenti del dizionario inglese. L'algoritmo implementato utilizza una risorsa esistente che codifica conoscenza di tipo enciclopedico, e poggia sulla nozione di senso comune per filtrare i possibili sensi associati a ciascun termine. La valutazione preliminare ha fornito risultati incoraggianti in merito alla qualità dei sensi estratti.*

1 Introduction

Many NLP tasks involve word sense disambiguation (WSD) and word sense induction (WSI), and require using lexical resources such as WordNet (Miller, 1995) and BabelNet (Navigli and Ponzetto, 2010) that provide a rich mapping of terms (or word *forms*) onto the corresponding senses (word *meanings*). These widely used resources provide in fact subtle distinctions between the possible senses of a term. It is largely acknowledged that while fine-grained sense distinctions are necessary for some precise tasks (such as machine translation), for other sorts of applications (such as text categorization and information extraction) coarse-grained sense inventories are preferable (Palmer et al., 2004). In these cases, fine-grained distinctions may be unnecessary and even detrimental to WSD and WSI, so that in the last few years many efforts concentrated on clustering senses. Most works focused on producing coarser-grained sense inventories, to the ends of grouping together the closest (partially overlapped) senses of a word; to these ends, various techniques have been carried out, that are briefly surveyed in Section 2.

Differently from existing approaches, we propose a simple yet effective method that relies on recently developed resources that are assumed to also grasp *common-sense knowledge* (Camacho-Collados et al., 2015; Lieto et al., 2016a), which is assumed to be both widely accessible and elementary knowledge (Minsky, 2000), and to reflect *typicality traits* encoded as prototypical knowledge (Rosch, 1975). The research question presently addressed is thus: To what extent can we individuate few principal —common-sense— senses for a term, and in how far is it possible to approximate human performance? Although it is known that even human annotators provide quite different response when annotating text with senses (Palmer et al., 2004), we presently explore the hypothesis that wide-coverage resources are sufficient to individuate the main senses associated to English terms.

2 Related Work

In order to attain coarse-grained senses, different approaches have been proposed, based on some sort of semantic underspecification (Buitelaar, 2000; Ng et al., 2003; Palmer et al., 2007), on existing dictionaries and on exploiting handcrafted sense hierarchies (Navigli, 2006), on syntactic and semantic properties (such as selectional restrictions on verb arguments) (Artale et al., 1998; Palmer et al., 2004), on linguistically motivated heuristics (Mihalcea and Moldovan,

2001), or on distributional similarity among word senses (Agirre and De Lacalle, 2003). Further approaches have been proposed that rely on an adjustable nearest neighbour schema for clustering senses according to the sense granularity actually required by the application at hand (McCarthy, 2006). A popular testbed for experimenting these and other approaches is represented by the sense-annotated *corpora* Senseval-2 and 3 (Edmonds and Cotton, 2001; Mihalcea and Edmonds, 2004).

The problem of annotating a term with the appropriate sense is a challenging one, to such an extent that by no means "two lexicographers working independently are guaranteed to derive the same set of distinctions for a given word" (Palmer et al., 2004). It has been raised that this issue can be overcome to some extent by adopting a more flexible annotation schema, where senses are described in a *graded fashion*: in this way, the applicability of a sense can be assessed on an ordinal scale, rather than in 'crisp' fashion. This sort of annotation would allow to better interpret human annotations, in particular for coarse-grained groups (Erk et al., 2013). A related and complementary issue is that of *clusterability*, that measures in how far word meanings can be partitioned. In this setting, whereas highly clusterable lemmas can be grouped based on traditional clustering techniques, less clusterable lemmas require more sophisticated soft-clustering algorithms to computational systems, and more time and expertise to human annotators (McCarthy et al., 2016).

This work is framed in the context of a long-term project aimed at investigating conceptual categorization (Lieto et al., 2015; Lieto et al., 2016b) based on a hybrid strategy (Evans and Frankish, 2009) complementing formal ontologies with the geometrical framework of Conceptual Spaces (CS) (Gärdenfors, 2014). In particular, we are building a knowledge base to collect conceptual information encoded in a CS-based representational format to provide a uniform interface between the linguistic and the conceptual level, where CSs representations are fully endowed with BabelNet identifiers (Lieto et al., 2016a).[1] This trait will make it possible to link the present work to existing initiatives like Senso Comune (Oltramari and Vetere, 2008; Chiari et al., 2010), that provides about 2000 fundamental Italian terms (De Mauro, 1999) with an ontological description.

3 The CLOSEST Algorithm

The rationale underlying the CLOSEST algorithm is that the main (most frequent) senses gained more room than marginal senses in our lexical and conceptual system and in general in our utterances. This phenomenon determines words and phrases *availability* and *saliency* (Vossen and Fellbaum, 2009), that are arguably grasped by encyclopedic resources, as well. Herein, more central senses are typically featured by richer (i.e., longer vectors) and less specific information, richer semantic connections with other concepts, and heavier feature weights. Although it may happen that some sense spans over (or even subsumes) another one, we are not primarily trying to cluster senses in agglomerative fashion, e.g., by resorting to some superclass of the considered concept; rather, we select the most relevant ones (a term is seldom associated to more than few, say three or four, senses) and we discard the other ones.

The CLOSEST algorithm takes in input a term t and provides a set of possibly related senses.[2] The algorithm first retrieves the set of senses $S = \{s_1, s_2, \ldots, s_n\}$ that are possibly associated to t: such set is obtained by directly querying NASARI. The output of the algorithm is a result set $S^{\trianglelefteq} \subseteq S$. In order to attain $S^{\trianglelefteq}$ we devised a process of incremental filtering, that is arranged into two main phases:

1. *LS-Pruning*. Pruning of less salient senses: senses with associated poor information are eliminated. Senses salience is determined both in absolute terms and in relation to the most salient sense.

2. *OL-Pruning*. Pruning of overlapping senses: if senses with significant overlap are found, the less salient sense is pruned.

Senses are represented as NASARI vectors, that are the vectorial counterpart of BabelNet synsets; concepts (basically, WordNet synsets and Wikipedia pages) are described through vector representations, whose features are synset IDs themselves. Feature weights are computed

[1] The integration of different semantic models such as CSs and the distributional semantics underlying NASARI is still an open issue; we provided an initial solution to this problem in (Lieto et al., 2016a).

[2] The present investigation is restricted to nouns, but no theoretical limitation prevents us from extending the approach to verbs and adjectives.

through the metrics of lexical specificity, by exploiting a semantics-based dimensionality reduction (Camacho-Collados et al., 2015). Each sense is associated with exactly one NASARI vector, so that pruning a sense amounts to pruning a vector.

LS-Pruning. To analyze the senses in S, we inspect each vector $\vec{v}_{ts}$ related to sense s for the term t. The first pruning occurs when no enough information is found, that is when $\vec{v}_{ts}$ contains less than a fixed number of elements (Table 1). Then, in order to determine the next vectors to be pruned, we compute the weight of each vector ($\overline{W}(\vec{v}_{ts})$), the longest vector and the heaviest one among those associated with t ($L(\vec{v}_t)$ and $H(\vec{v}_t)$, respectively). The weight of a NASARI vector $\overline{W}(\vec{v}_{ts})$ is computed by averaging the weight of the features (i.e., the synsets) contained herein. The definitions for these measures are illustrated in Equations 1–3.

$$L(\vec{v}_t) = \arg\max_{s \in S} (\text{len}(\vec{v}_{ts})) \quad (1)$$

$$\overline{W}(\vec{v}_{ts}) = \frac{1}{\text{len}(\vec{v}_{ts})} \cdot \sum_j w_{sj} \quad (2)$$

$$H(\vec{v}_t) = \arg\max_{s \in S} \left(\overline{W}(\vec{v}_{ts})\right). \quad (3)$$

The decision on whether to prune or not a vector is based on a simple criterion: $\vec{v}_{ts} \in S$ is pruned if both its length is below a given fraction of the length of the longest one $L(\vec{v}_t)$, *and* its weight is lower than a given fraction of the heaviest one, $H(\vec{v}_t)$. The parameter settings adopted in the present work are illustrated in Table 1.

OL-Pruning. The second phase of the algorithm aims at detecting overlapped senses. The overlap between vectors that survived the *LS-Pruning* is computed thanks to the information provided in NASARI. The heuristics used in this phase is as follows: the overlap between two vectors $Ovl(\vec{v}_{ti}, \vec{v}_{tj})$ is computed as a fraction of the length of the shortest vector between the two considered, as indicated in Equation 4).

$$Ovl(\vec{v}_{ti}, \vec{v}_{tj}) = \frac{\vec{v}_{ti} \cap \vec{v}_{tj}}{\text{len}(shortest(\vec{v}_{ti}, \vec{v}_{tj}))} \quad (4)$$

The overlapping is checked for every pair $\langle \vec{v}_i, \vec{v}_j \rangle$ (with $i \neq j$) and when an overlap is detected higher than a fixed threshold (see Table 1), the shortest vector between the two is pruned.

At the end of this phase, we have the set $S^{\lhd}$ where only the most salient vectors survived and where, among overlapped vectors, the most salient one has been retained.

3.1 Building the CLOSEST resource

Overall the system handled about $2.69M$ NASARI vectors. Some $207K$ vectors associated to Named Entities were discarded, as not directly related to common-sense concepts; the remaining vectors contained overall $6.9M$ unique words.

The top (most frequent) $15K$ nouns were extracted from the Corpus of Contemporary American English (COCA) which has been built from composite and balanced sources, including spoken, fiction, magazine, newspaper, academic text.[3] Over $6K$ terms were discarded, since they are associated in NASARI either to 1 sense (about $1K$ terms) or to no sense at all (over $5K$ terms), which actually reduced the input size to about $8.7K$ terms; overall $32.6K$ senses were retrieved (on average, 3.7 senses per term), corresponding to such input terms.

The figures featuring the processing phases are reported in Table 1: over $4K$ senses were filtered in the first step of the *LS-Pruning* phase, based on the length of the vector $\vec{v}_{ts}$, and $7.4K$ senses were further discarded in the second step. Finally, in the *OL-Pruning* phase, $5.6K$ vectors were canceled based on overlapping accounts, thus overall yielding $17.5K$ deleted and $15.1K$ survived vectors.[4] The polysemy rate was reduced from the 3.74 senses per term initially featuring NASARI down to 1.73 senses per term, which is in line with the degree of polysemy detected in the Collins English Dictionary for English nouns by WordNet authors (Fellbaum, 1990).

4 Evaluation

A preliminary experimentation has been devised to assess the correctness and completeness of the extracted senses: that is, the question addressed was whether *i)* all senses extracted for the input term are salient (and actually judged as the main senses), and *ii)* all the relevant senses were preserved in CLOSEST. To these ends, 15 volunteers were recruited and interviewed through an on-line questionnaire to evaluate, on a human common-sense judgement basis, the set of senses extracted by the system for 20 terms.

[3] http://corpus.byu.edu/full-text/.

[4] CLOSEST is available at http://goo.gl/7B61Oz.

	condition	threshold values	pruned senses	pruning phase
prune $\vec{v}_{ts}$ IF	$\text{len}(\vec{v}_{ts}) \leq \alpha$	$\alpha = 5$	$4,389$	*LS-Pruning*
	$\left(\frac{\text{len}(\vec{v}_{ts})}{L(\vec{v}_t)} < \beta\right)$ AND $\left(\frac{\overline{W}(\vec{v}_{ts})}{\overline{W}(H(\vec{v}_t))} < \gamma\right)$	$\beta, \gamma = .40$	$7,460$	*LS-Pruning*
	$Ovl(\vec{v}_{ts}, \vec{v}_{tu}) \geq \delta$	$\delta = .20$	$5,676$	*OL-Pruning*
		filtered out senses	17,525	
		retained senses	15,134	

Table 1: Pruning of senses in the three steps, along with the number of senses pruned at each step.

Stimuli. The list of 20 terms was algorithmically selected from the aforementioned COCA corpus (see footnote 3) by selecting terms herein with index 1, 51, 101, and so forth. In this way we selected highly frequent terms that are expected to be part of common-sense for those who participated in our experimentation.[5]

Experimental design and procedure. The participants were asked *a)* to assess each and every sense extracted by the system and associated to each input term by indicating whether it was acceptable as one of the principal senses for the term at hand. Additionally, they were requested *b)* to indicate any further sense they reputed essential in order to complete the common-sense pool of senses for the given term.

Results. Overall 42 senses (corresponding to the 20 mentioned terms) were assessed through the experimentation: each sense was rated 15 times, thus resulting in 630 judgements: 24% of senses were not found appropriate, according to a common-sense judgement, thereby determining a 76% accuracy as regards as question *a)*. However, if we consider senses refused by at least 10 participants, only 5 senses were refused (12%), that actually correspond to very specific senses (e.g., the sense 'Net (textile)' for the term 'network'; 'Session (Presbyterianism)', 'session house' for the term 'session').

As regards as question *b)*, results are more difficult to interpret, due to the sparsity of the answers: out of the 59 added senses, only in 8 cases the added sense has been indicated by two or three participants (and never more): in such cases it emerged, for example, that the sense 'manners' was relevant (and missing, in the CLOSEST resource) for the input term 'education'; the sense 'social network' is relevant for the term 'network'; and 'meeting' for 'session'.

However, although encouraging results emerged from the experimentation, further experiments are needed to assess the CLOSEST resource in a more extensive and principled way, also in consideration of the many factors that were presently neglected, such as, e.g., age, education, occupation of the participants, their native language, *etc.*.

5 Conclusions

In this paper we have illustrated the CLOSEST algorithm to extract the most salient (under the common-sense perspective) senses associated to a given term; also, we have introduced the CLOSEST resource, which has been built by starting from the $15K$ top frequency English terms. The resource currently provides senses in a flat manner, but, if required, senses can be organized in a sorted fashion by extending the metrics used for filtering. Our work relies on a recently developed resource such as NASARI that is multilingual in nature.[6] Consequently, different from most previous approaches, CLOSEST can be linked to various existing resources aimed at grasping common-sense to complete the ideal chain connecting lexicon, semantics and formal (ontological) description. The experimentation revealed a reasonable agreement with human responses, and pointed out some difficulties in fully assessing this sort of resource. These issues, along with improvements to the heuristics implemented by the algorithm and a different evaluation based on a shared NLP task, will be addressed in future work.

[5] The full list of the considered terms includes: time, side, education, type, officer, ability, network, shoulder, threat, investigation, gold, claim, learning, session, aid, emergency, bowl, pepper, milk, resistance. The printed version of the online questionnaire is available at the URL `http://goo.gl/w9TNQT`.

[6] An interesting question may be raised on this point, about the conceptual alignment in a *inter*-linguistic perspective, which is a well-known issue, e.g., for applications in the legal field (Ajani et al., 2010).

References

Eneko Agirre and Oier Lopez De Lacalle. 2003. Clustering WordNet Word Senses. In *RANLP*, volume 260, pages 121–130.

Gianmaria Ajani, Guido Boella, Leonardo Lesmo, Marco Martin, Alessandro Mazzei, Daniele P Radicioni, and Piercarlo Rossi. 2010. Multilevel legal ontologies. In *Semantic Processing of Legal Texts*, pages 136–154. Springer.

Alessandro Artale, Anna Goy, Bernardo Magnini, Emanuele Pianta, and Carlo Strapparava. 1998. Coping with WordNet Sense Proliferation. In *First International Conference on Language Resources & Evaluation*.

Paul Buitelaar. 2000. Reducing Lexical Semantic Complexity with Systematic Polysemous Classes and Underspecification. In *NAACL-ANLP 2000 Workshop: Syntactic and Semantic Complexity in Natural Language Processing Systems*, pages 14–19. Association for Computational Linguistics.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. NASARI: a Novel Approach to a Semantically-Aware Representation of Items. In *Proceedings of NAACL*, pages 567–577.

Isabella Chiari, Alessandro Oltramari, and Guido Vetere. 2010. Di Cosa Parliamo quando Parliamo Fondamentale? Lessemi, Accezioni, Sensi e Ontologie. In *Lessico e Lessicologia. Atti del Convegno della Societ di Linguistica Italiana*, pages 177–194, Roma, September. Bulzoni.

Tullio De Mauro. 1999. *Grande Dizionario Italiano dell'Uso*. UTET, Turin, Italy.

Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France, July. Association for Computational Linguistics.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

Jonathan St BT Evans and Keith Ed Frankish. 2009. *In Two Minds: Dual Processes and Beyond*. Oxford University Press.

Christiane Fellbaum. 1990. English Verbs as a Semantic Net. *International Journal of Lexicography*, 3(4):278–301.

Peter Gärdenfors. 2014. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press.

Antonio Lieto, Daniele P. Radicioni, and Valentina Rho. 2015. A Common-Sense Conceptual Categorization System Integrating Heterogeneous Proxytypes and the Dual Process of Reasoning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 875–881, Buenos Aires, July. AAAI Press.

Antonio Lieto, Enrico Mensa, and Daniele P. Radicioni. 2016a. A Resource-Driven Approach for Anchoring Linguistic Resources to Conceptual Spaces. In *Proceedings of the 15th International Conference of the Italian Association for Artificial Intelligence*, Genoa, Italy, December. Springer.

Antonio Lieto, Daniele P Radicioni, and Valentina Rho. 2016b. Dual PECCS: a Cognitive System for Conceptual Representation and Categorization. *Journal of Experimental & Theoretical Artificial Intelligence*, pages 1–20.

Diana McCarthy, Marianna Apidianaki, and Katrin Erk. 2016. Word Sense Clustering and Clusterability. *Computational Linguistics*.

Diana McCarthy. 2006. Relating WordNet Senses for Word Sense Disambiguation. *Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, 17.

Rada Mihalcea and Phil Edmonds. 2004. SENSEVAL-3: Overview. In *Proceedings Senseval-3 3rd International Workshop on Evaluating Word Sense Disambiguation Systems. ACL, Barcelona, Spain*.

Rada Mihalcea and Dan I Moldovan. 2001. Automatic Generation of a Coarse Grained WordNet. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources*.

George A Miller. 1995. WordNet: a Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Marvin Minsky. 2000. Commonsense-based interfaces. *Communications of the ACM*, 43(8):66–73.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225. Association for Computational Linguistics.

Roberto Navigli. 2006. Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 105–112. Association for Computational Linguistics.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 455–462. Association for Computational Linguistics.

Alessandro Oltramari and Guido Vetere. 2008. Lexicon and Ontology Interplay in Senso Comune. *OntoLex 2008 Programme*, page 24.

Martha Palmer, Olga Babko-Malaya, and Hoa Trang Dang. 2004. Different Sense Granularities for Different Applications. In *Proceedings of Workshop on Scalable Natural Language Understanding*.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making Fine-Grained and Coarse-Grained Sense Distinctions, both Manually and Automatically. *Natural Language Engineering*, 13(02):137–163.

Eleanor Rosch. 1975. Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General*, 104(3):192–233.

Piek Vossen and Christiane Fellbaum, 2009. *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, chapter Universals and idiosyncrasies in multilingual WordNets. Trends in linguistics / Studies and monographs: Studies and monographs. Mouton de Gruyter.

Building a computational lexicon by using SQL

Alessandro Mazzei
Dipartimento di Informatica
Universit degli Studi di Torino
Corso Svizzera 185, 10149 Torino
mazzei@di.unito.it

Abstract

English. This paper presents some issues about a computational lexicon employed in a generation system for Italian (Mazzei et al., 2016). The paper has three goals: (i) to describe the SQL resources produced during the construction of the lexicon; (ii) to describe the algorithm for building the lexicon; (iii) to present an ongoing work for enhancing the lexicon by using the syntactic information extracted from a treebank.

Italiano. *Questo lavoro descrive la costruzione di un lessico computazionale per la generazione automatica dell'italiano (Mazzei et al., 2016). Il lavoro ha tre obiettivi: (i) descrivere alcune risorse SQL prodotte funzionalmente alla costruzione del lessico; (ii) descrivere l'algoritmo per la costruzione del lessico; (iii) presentare un lavoro in divenire per migliorare il lessico che usa l'informazione sintattica estratta da un treebank.*

1 Introduction

A number of free large multilingual resources covering Italian have been released, e.g. MultiWordnet, UniversalWordnet, BabelNet (Pianta et al., 2002; de Melo and Weikum, 2009; Navigli and Ponzetto, 2012). Moreover, several lexical corpora have been built specifically for Italian, as the detailed map of the Italian NLP resources produced within the PARLI project shows[1]. Unfortunately most resources are designed to represent lexical semantics rather than morpho-syntactic relations among the words. As a consequence, these resources cannot be employed in statistical or rule-based natural language morho-syntactic analyzer or generator.

A notable exception is the PAROLE-SIMPLE-CLIPS lexicon, that is a four-level (i.e. phonological, morphological, syntactical, semantic) general purpose lexicon composed by $53,044$ lemmata (Ruimy et al., 1998). Unfortunately, a strong limitation for the usage of PAROLE-SIMPLE-CLIPS is the licence, since it is not freely available for research or commercial use.

Rule-based natural language realization engines, that are systems performing linearisation and morphological inflections of a proto-syntactic input tree (Gatt and Reiter, 2009), need wide coverage morpho-syntactic information as knowledge-base. In other words, to perform realization, that is the last step of natural language generation (Reiter and Dale, 2000), one needs two main kinds of linguistic knowledge: (i) the grammatical/syntactical knowledge that specifies the syntactic rules of the language and which is usually encoded into formal rules; (ii) the morphological and lexical knowledge, which is usually encoded into a computational lexicon. In the porting of the SimpleNLG system to Italian (henceforth *SimpleNLG-IT*) (Mazzei et al., 2016), we have used the grammar (Patota, 2006) as the linguistic reference for the syntax: we have encoded the Italian syntactic inflections and word ordering by using IF-THEN-ELSE rules in Java. However, since Italian has a high number of irregularities for verb and adjective inflections, we needed for a specifically designed computational lexicon too. We needed for a lexicon that has both a good coverage and a detailed account of the morphological irregularities.

In order to build this specific lexicon, that we have called *SimpleLEX-IT*, we have decided to merge three free resources for Italian, namely *Morph-it!* (Zanchetta and Baroni, 2005), the *Vo-*

[1] http://parli.di.unito.it/resources_en.html

cabolario di base della lingua italiana (De Mauro, 1985) and, for some specific issues, Wikipedia. The differences between the three resources can be referred to both the reasons for which the authors developed them and the adopted methodology and approach they applied in their development: the first is a hand-made list of *basic* words; the second one is an extensional corpus based morphological lexicon; the third one is a collection of encyclopedic entries about irregular verbs in Italian.

This paper is organized as follows: in Section 2 we describe the conversion of the three lexical resources used into a relational database; in Section 3 we provide some details about the algorithm used to build SimpleLEX-IT; in Section 4 we describe a work in progress to enrich the lexicon by using the syntactic information extracted from a treebank; finally, Section 5 closes the paper with conclusions.

2 Using relational database for representing linguistic data

In order to merge different lexical resources we needed to convert them in a common computational representation. We used a relational database[2] (SQL henceforth) since all the three resources are originally provided as text files, organized as tables or simple list.

The first resource that we exploited for populating SimpleLEX-IT is Morph-it! (Zanchetta and Baroni, 2005). The dataset released in the Morph-it! project consists of a lexicon organized according to the inflected word forms, with associated lemmas and morphological features. The lexicon is provided by the authors as a text file where the values of the information about each lexical entry are separated by a tab key. It is an alphabetically ordered list of triples form-lemma-features. An example of the annotation for the form *corsi* (*ran*) is:

```
corsi correre-VER:ind past+1+s
```

where the features are the part of speech (PoS, `VER`b), the mood of the verb (`ind`icative), the tense (`past`), the person (`1`), and the number (`s`ingular). The last released version of Morph-it! (v.48, 2009-02-23) contains $505,074$ different forms corresponding to $35,056$ lemmas. It has been realized starting from a large newspaper corpus, nevertheless it is not balanced and a small number of also very common Italian words are not included in the lexicon, e.g. *sposa* (bride), *ovest* (west) or *aceto* (vinegar). Morph-it! represents extensionally the Italian language by listing all the morphological inflections, i.e. adjective, verbs, nouns inflections are represented as a list rather than by using morphological rules. We converted Morph-it! in SQL by exploiting its original feature structure: we used one single attribute to represent one single feature[3]. We used one table to collect all the lemmata and seven tables, with a different number of attributes, to collect the various inflected forms:

- the table *lemmata* is formed by 3 attributes: a lemma, its PoS and its ID (integer). This table contains $34,725$ records. A number of lemmata belonging to the original version of Morph-it! have been excluded in our conversion: proper nouns, emoticons and cardinals beginning with a digit (e.g. *15mila*).
- the tables *det_demo_table, pro_demo_table, pronou_table* are used to collect inflected form of demonstrative determiners (116 records, 4 attributes: *ID_word, form, ID_lemma, number, gender*), demonstrative pronouns (95 records, 5 attributes: *ID_word, form, ID_lemma, number, gender*), personal pronouns (63 records, 7 attributes: *ID_word, form, ID_lemma, person, number, gender, clitics*).
- the tables *adv_table, adj_table, nou_table, ver_table* are used to collect inflected form of adverbs ($1,594$ records, 3 attributes: *ID_word, form, ID_lemma*), adjectives ($72,367$ records, 6 attributes: *ID_word, form, ID_lemma, kind, number, gender*), nouns ($35,618$ records, 5 attributes: *ID_word, form, ID_lemma, number, gender*) and verbs ($392,139$, 8 attributes: records: *ID_word, form, ID_lemma, mode, time, person, number, gender*) respectively.

The second resource we exploited for populating SimpleLEX-IT is the "Vocabolario di base della lingua italiana" (*VdB-IT* henceforth), a collection of around $7,000$ words created by the linguist Tullio De Mauro and his team[4] (De Mauro, 1985). The development of this vocabulary has been mainly driven by the distinction between the

[2]We used the PostgreSQL database .

[3]Morph-IT! is provided with a script that allows for a naive conversion into SQL that use one single table and one single attribute for all the features.

[4]The second edition of the vocabulary has been announced (Chiari and De Mauro, 2014) and it is going to be released (p.c.).

most *frequent* words (around 5.000) and the most *familiar* words (around 2.000). VdB-IT is therefore organized in the following three sections:

- the *vocabolario fondamentale* (fundamental vocabulary), which contains 2,000 words featured by the highest frequency into a balanced corpus of Italian texts (composed of novels, movie and theater scripts, newspapers, basic scholastic books); *amore* (love), *lavoro* (work), *pane* (bread) are in this section.
- the *vocabolario di alto uso* (vocabulary of high usage), which includes other 2,937 words with high frequency, but lesser than the *vocabolario fondamentale*; *ala* (wing), *seta* (silk), *toro* (bull) are in this section
- the *vocabolario di alta disponibilità* (vocabulary of high availability), is composed of 1,753 words not often used in written language, but featured by a high frequency in spoken language, which are indeed perceived as especially familiar by native speakers; *aglio* (garlic), *cascata* (waterfall), *passeggero* (passenger) are in this section.

The list of lemmata of VdB has been converted in SQL by using one single table, called *lemmadema* (6540 records), which have two attributes, i.e. an ID (integer) and the lemma.

The third resource that we used for the lexicon creation is Wikipedia. Our reference grammar (Patota, 2006) reports a partial list of the principal Italian irregular verbs, but we decided to use the larger list of verbs reported in Wikipedia[5] (*VerIrr* henceforth). Another linguistic distinction for Italian verbs reported in Wikipedia[6] (*VerInc* henceforth) has been exploited in the lexicon: the *incoativi* verbs are a subclass of the third conjugation that have a special behavior in the present time (e.g. *capire*). So, in order to produce the correct conjugation of these verbs in SimpleNLG-IT, they needed to be marked in the lexicon. Both these lists of verbs have been converted in SQL by using two distinct tables which have two attributes, i.e. an ID (integer) and the verb in the infinitive form. The two tables are *verbiirregolari* (858 records) and *verbiincoativi* (726 records).

A notable advantage of the SQL representation for linguistic resources is the possibility to extract intrinsic information with simple queries. Indeed, we found that Morph-it! and VdB share 4,086 nouns and 1,448 verbs, but there are 245 lemmas belonging to VdD and not belonging to Morph-it!: most of these words are nouns, for instance *lavapiatti, chimica, inclinta*, but we found too a systematic difference for verbs. Indeed, VdB consider as *proper* reflexive a number of verbs, for instance *avvantaggiarsi, sdraiarsi*. In contrast, these verbs are are treated as *improper* reflexive in Morph-it!, which annotates *avvantaggiare* and *sdraiare* as their lemmata.

[5] https://it.wikipedia.org/wiki/Verbi_irregolari_italiani

[6] https://it.wikipedia.org/wiki/Verbi_incoativi

3 Building SimpleLEX-IT 1.0

In this section we describe the algorithm used to build the computational lexicon SimpleLEX-IT, which is based on the three resources described in the Section 2, and that has been used in SimpleNLG-IT.

A computational lexicon can be split in two major classes: open and closed classes. The closed class, that are usually composed by function words (i.e. prepositions, determiners, conjunctions, pronouns, etc.) is one to which new words are very rarely added. In contrast, the open classes, that is usually composed by lexical words (i.e. nouns, verbs, adjectives, adverbs), accept the addition of new words. We adopted the same strategy of (Vaudry and Lapalme, 2013): we built by hand the closed part of the Italian lexicon and we built automatically the open part by using the available resources.

In order to build the open class for SimpleLEX-IT we needed both a large coverage and a detailed account of morphological irregularities, also considering their high frequency in Italian. Moreover, in order to have good time execution performance in the realiser (cf. (De Oliveira and Sripada, 2014)), a trade-off between the size of the lexicon and its usability for our task must be achieved, which consists in assuming a form of word classification where fundamental Italian words are distinguished from the less-fundamental ones. In order to balance completeness and efficiency in SimpleLEX-IT, we put in the lexicon the open classes words belonging to the intersection of VdB-IT and Morph-it!.

We reported in Algorithm 1 the process used for the insertion and the annotation of the words belonging to the open classes in SimpleLEX-IT. Note that in order to recognize proper reflexive verbs, we check if the infinitive form of the verb

```
foreach adverb ∈ Morph-it! ∩ VdB-IT do
    Add the adverb in normal form into SimpleLEX-IT
end
foreach adjective ∈ Morph-it! ∩ VdB-IT do
    Add the adjective in normal form (masculine-singular) and in
    feminine-singular, masculine-plural, feminine-plural forms, into
    SimpleLEX-IT
end
foreach noun ∈ Morph-it! ∩ VdB-IT do
    Add the noun in normal form (singular), the plural form, and the
    gender into SimpleLEX-IT
end
foreach verb ∈ Morph-it! ∩ VdB-IT do
    if verb ∈ VerIrr then
        Add all the inflections for the indicativo presente,
        congiuntivo presente, futuro semplice, condizionale,
        imperfetto, participio passato, passato remoto into
        SimpleLEX-IT
    else
        if verb ∈ VerInc then
            Set active the incoativo feature in the entry
        end
        if the verb is properly reflexive (i.e. "...rsi") then
            Set active the reflexive feature in the entry
        end
        Add the verb in normal form into SimpleLEX-IT
    end
end
```

Algorithm 1: The algorithm for building the adverbs, adjectives, nouns and verbs in SimpleLEX-IT

has the postfix "*rsi*", since MorphIT! contains this inflection as its normal form. In Table 1 we reported some statistics about SimpleLEX-IT composition. Most of the lexicon is composed by nouns (58%), followed by verbs (21%), adjectives (19%), and adverbs (2%).

PoS	Number	%
Adverb	146	2
Verb (irr.)	283	4
Verb (reg.)	1168	17
Adjective	1333	19
Noun	4092	58
Total	7022	100

Table 1: Number of adverbs, adjectives, nouns and verbs in SimpleLEX-IT.

4 Work in Progress: adding information from a treebank

The Universal Dependency Treebank (UDT) is a recent project that releases freely available treebanks for 33 languages (in this work, version 1.2) (Nivre et al., 2016). Each UDT is split in three sections, *train*, *dev* and *test*, which can be exploited in the evaluation of NLP/NLG systems.

We are working on the idea of adding more information in SimpleLEX-IT by using UDT-IT, i.e. the Italian section of UDT. A specific case that we are currently considering regards auxiliary verbs. The current version of SimpleNLG-IT does not manage auxiliary verbs: in order to produce some complex verb tense, e.g. *passato prossimo*, the user needs to give in input to the realiser the correct auxiliary, i.e. *essere* (to be, e.g. *Io sono nato a Napoli*) or *avere* (to have, e.g. *Io ho amato la scuola.*). Our reference grammar reports complex rules based on lexical semantics in order to choose the correct auxiliary verb and, unfortunately, these rules have many exceptions. So, we can use UDT-IT to empirically decide the correct auxiliary in SimpleNLG-IT. By following this idea, we converted UDT-IT in SQL by exploiting its original feature structuree. We used one table to collect information about the sentences, and one table to collect information about the words:

- the table *sentence_ud* is formed by 4 attributes: an ID (integer), the original treebank (i.e. TUT, ISST, etc.), the original ID, the section (i.e. DEV, TRAIN, TEST).
- the table *words_ud* is used to collect all the words of the UDT-IT. It uses 21 attributes: one attribute *id_sentence*, contains the id of the sentence in the table *sentence_ud*, and 20 attributes correspond to the featured used in the UD annotation.

In order to find the correct auxiliary for a specific verb in UDT-IT, we need to exclude passive, reflexive and modal verb constructions in the query. We found 512 verbs of SimpleLEX-IT that are used in UDT-IT with an auxiliary. It is interesting to note that 60 verbs are used both with the auxiliary *essere* and with the auxiliary *avere*: this is grammatical for some verbs (e.g. *vivere*), but more often we found an annotation error in the UDT-IT.

Finally, note that another possible use of UDT-IT regards the evaluation of the lexicon. In future work we plan to quantify the coverage of SimpleLEX-IT by using the TEST section of the UDT-IT.

5 Conclusions

In this paper we have presented some issues about the computational lexicon SimpleLEX-IT. We described the algorithm used to build the lexicon, three SQL resources produced as side effects of the lexicon building and a work in progress about the extraction of syntactic information from UD-IT.

All the resources described in this paper can be downloaded at `https://github.com/alexmazzei/SimpleLEX-IT`.

References

Isabella Chiari and Tullio De Mauro. 2014. The New Basic Vocabulary of Italian as a linguistic resource. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors, *1th Italian Conference on Computational Linguistics (CLiC-it)*, volume 1, pages 93–97. Pisa University Press, December.

Tullio De Mauro. 1985. *Guida all'uso delle parole*. Libri di base. Editori Riuniti.

Gerard de Melo and Gerhard Weikum. 2009. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

Rodrigo De Oliveira and Somayajulu Sripada. 2014. Adapting simplenlg for brazilian portuguese realisation. In *Proc. of INLG 2014*.

Albert Gatt and Ehud Reiter. 2009. SimpleNLG: A Realisation Engine for Practical Applications. In *Proc. of ENLG 2009*, ENLG '09.

Alessandro Mazzei, Cristina Battaglino, and Cristina Bosco. 2016. SimpleNLG-IT: adapting SimpleNLG to Italian. In *Proc. of INLG 2016*. TO APPEAR.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1:A Multilingual Treebank Collection. In *Proc. of LREC'16*, may. TO APPEAR.

Giuseppe Patota. 2006. *Grammatica di riferimento dell'italiano contemporaneo*. Guide linguistiche. Garzanti Linguistica.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, January.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press, New York, NY, USA.

Nilda Ruimy, Ornella Corazzari, Elisabetta Gola, Antonietta Spanu, Nicoletta Calzolari, and Antonio Zampolli. 1998. The European LE-PAROLE project: the Italian Syntactic Lexicon. In *Proceedings of the First International Conference on Language resources and Evaluation* , pages 241–248.

Pierre-Luc Vaudry and Guy Lapalme. 2013. Adapting simplenlg for bilingual english-french realisation. In *Proc. of ENLG 2013*.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the italian language. *Corpus Linguistics 2005*, 1(1).

FB-NEWS15: A Topic-Annotated Facebook Corpus for Emotion Detection and Sentiment Analysis

Lucia C. Passaro, Alessandro Bondielli and **Alessandro Lenci**
CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica
University of Pisa (Italy)
lucia.passaro@for.unipi.it
alessandro.bondielli@gmail.com
alessandro.lenci@unipi.it

Abstract

English. In this paper we present the FB-NEWS15 corpus, a new Italian resource for sentiment analysis and emotion detection. The corpus has been built by crawling the Facebook pages of the most important newspapers in Italy and it has been organized into topics using LDA. In this work we provide a preliminary analysis of the corpus, including the most debated news in 2015.

Italiano. *In questo lavoro presentiamo il corpus FB- NEWS15, un corpus italiano creato per scopi di sentiment analysis ed emotion detection. Il corpus stato costruito scaricando le pagine Facebook delle maggiori testate giornalistiche in Italia e successivamente organizzato in topic utilizzando LDA. In questo articolo forniamo una analisi preliminare del corpus, e mostriamo le notizie pi discusse nel 2015.*

1 Introduction

The use of Social Networks (SN) platforms like Facebook and Twitter has developed overwhelmingly in recent years. SN are exploited for different purposes ranging from the sharing of contents among friends and useful contacts to the news-gathering about different domains such as politics and sports (Ahmad, 2010; Ahmad, 2013; Sheffer and Schultz, 2010). Many journalists indeed use SN platforms for professional reasons (Oriella, 2013; Hermida, 2013).

Several recent studies provide insights on how the popularity of blogs and other user generated content impacted the way in which news are consumed and reported. Picard (2009) states that SN platforms provide an easy and affordable way to take part in discussions with larger groups of people and, consequently, the bond between SN and information is becoming increasingly stronger.

Mass information is gradually moving towards general platforms, and official websites are losing their lead position in providing information. As noted by Newman et al. (2012), even though the use of internet in the years 2009-2012 has grown, the same is not reflected in the consumption of online newspapers, probably because of the increasing use of SN for news diffusion and gathering. If on the one hand this apparent decline of the traditional news platforms may lead to a decline in quality and news coverage (Chyi and Lasorsa, 2002), on the other hand the rise of SN as platforms to spread news promotes a more fervid debate between users (Shah et al., 2005). This issue is central for the present work. In fact, user's comments very often contain their own opinions about a certain issue. In addition, because of the colloquial style of the comments, they contain large amounts of words and collocations with a high subjective content, mostly concerning the author's emotive stance.

Facebook is one of the most popular online SN in the world with 1 billion active users per month and it offers the possibility to collect data from people of different ages, educational levels and cultures. From a linguistic point of view, previous studies (Lin and Qiu, 2013) demonstrated that the language in Facebook is more emotional and interpersonal compared for example to the language in Twitter. Probably, this is due to the fact that in Facebook there is a stronger psychological closeness between the author and audience because of the different structure (bidirectional vs. unidirectional graphs) of the SNs.

In this paper we present the FB-NEWS15 corpus, a new Italian resource for sentiment analysis and emotion detection. The FB-NEWS15 corpus can be freely downloaded at

colinglab.humnet.unipi.it/resources/under the Creative Commons Attribution License creativecommons.org/licenses/by/2.0.[1]

The debate among users in commenting news and posts on Facebook offers a lot of subjective material to study the way in which people express their own opinions and emotions about a target event. In fact, in FB-NEWS15 we find linguistic items expressing the whole range of positive and negative emotions. In analyzing a news corpus, however, it is not simple to aggregate the posts on the basis of a certain fact, since several posts relate to the same event. For this reason, we decided to organize the corpus into clusters of topically related news identified with Latent Dirichlet Allocation (LDA: Blei et al. (2003)). This approach allow us to infer the most debated news in the corpus, and, in a second step, to discover the readers' sentiment about a particular topic.

The paper is organized as follows: Section 2 describes the creation of the corpus, from crawling (2.1) to linguistic annotation (2.2), and finally provides basic corpus statistics (2.3). Section 3 reports on the automatic topic extraction with LDA.

2 FB-NEWS15

For the creation of the corpus we followed the most important Italian newspapers. Since we were interested in building a corpus as heterogeneous as possible, we decided to focus on major newspapers with different political orientations, and which have in general heterogeneous readers.

Facebook allow users to post states, links, photos and videos on their own wall. In general, users can be divided into two macro-categories: People and Pages. People are often individuals, and the interaction with them is usually bidirectional (user A can read what user B publishes if A and B have a *friendship relation*). Conversely, Pages are typically used to represent organizations, public figures (web stars), companies or, as in our case, newspapers. In this case, the relationship is unidirectional, in the sense that user A can access the timeline of the page P by putting a "Like" on P. Unlike a single-user, who usually publishes photos, videos and links about his private life, the timeline of a newspaper Facebook page, in general contains news titles with a link to the official website of the newspaper, where the user can read the full article. The corpus keeps tracks of the threefold hierarchical structure of Facebook, which includes the news posts by the newspaper, the users' comments to the posts and the replies to the comments. In this context, it is clear that the emotive content of the post is often neutral, but this post can inspire long discussions among readers, which can become useful material for sentiment analysis and emotion detection. Figure 1 shows a post, with some of its comments and replies.



Figure 1: Example of post in Facebook with the relative comments and replies.

In order to create the FB-NEWS15, we decided to download the timeline of the following newspapers, from 1 January 1 to 31 December 2015: La Repubblica, Il Giornale, L'Avvenire, Libero, Il Fatto Quotidiano, Rainews24, Corriere Della Sera, Huffington Post Italia.

2.1 Crawling

Facebook offers developers Application Programming Interfaces (APIs) for creating apps with Facebook's native functionalities. In order to develop the crawler, we exploited the Graph API, which provides a simple view of the Facebook social graph by showing the objects in the graph and the connections between them. The Graph API allows us to navigate through the graph of the social network, which is organized into nodes

[1] All data collected have been processed anonymously for scientific purposes, without storing personal information.

```
<doc user="<newspaper(string)>"
 id="<id_post(string)>"
 type="post"
 parent_post=""
 parent_comment=""
 date="AAAA-MM-DD HH:MM:SS"
 location=""
 likes="662"
 comments="54"
 shares="322">

Un business truffaldino [E ora
finitela con l'eco-balla dei
controlli sulle emissioni]

</doc>
```

Figure 2: Example of crawled text.

(Users, Pages, Photos and Comments) and Edges (Connections such as Friendship or Likes). The graph is navigated by exploiting HTTP requests, that may be implemented using any programming language. The native APIs offered by Facebook has some drawbacks: i) the maintenance of the app, since the APIs change over time, making it necessary to update the code of the crawler; ii) only public data can be accessed without requiring the user's consent; iii) Facebook places limitations on the number of requests through a given period of time. For each post, comment and reply, we stored the message (text), the story (presence of photos and links tags), its timestamp, the type (post, comment, reply), the parent post/comment, the number of likes, shares and replies (Figure 2).

2.2 Linguistic annotation

A very basic preprocessing phase has been applied to the corpus before linguistic annotation, to replace urls with the tag _URL_. The text has been subsequently feed to a pipeline of general-purpose NLP tools. In particular, it has been POS-tagged with the Part-Of-Speech tagger described in (Dell'Orletta, 2009) and dependency-parsed with the DeSR parser (Attardi et al., 2009). In addition, complex terms like *forze dell'ordine* (security force) or *toccare il fondo* (hit rock bottom) have been identified using the EXTra term extraction tool (Passaro and Lenci, 2015).

2.3 Corpus Analysis

Except for Avvenire and Rainews24, for which we downloaded very few data, the other newspapers are attested in the corpus in a balanced way. In general, the number of posts is very low compared to the number of comments and replies. The average number of posts for each newspaper is 27,341.25, while for comments and replies is respectively 2,016,243.38 and 576,498.5. Table 1shows the number of texts (including posts, comments and replies) in FB-NEWS15 for each Newspaper and Figure 2.3 shows their cumulative distribution for each Newspaper.

NEWSPAPER	N. OF TEXTS
La Repubblica	4558,829
Avvenire	91,824
Il Giornale	3,497,610
Libero	2,436,246
Il Fatto Quotidiano	4,900,314
Rainews24	369,834
Huffington Post	1,552,042
Corriere della Sera	3,553,966
OVERALL	20,960,665

Table 1: Number of texts aggregated by Newspaper in FB-NEWS.

Table 2 shows the total number of tokens for each page and the average number of texts, produced for each post for each page. We can notice that the most followed newspapers on Facebook are Il Fatto Quotidiano and La Repubblica.

NEWSPAPER	TOKENS	TEXTS/POSTS
La Repubblica	96,059,756	182.61
Avvenire	2,611,899	12.65
Il Giornale	64,345,260	77.93
Libero	41,166,457	81.87
Il Fatto Quotidiano	99,025,541	193.33
Rainews24	7,735,908	10.21
Huffington Post	32,587,065	84.06
Corriere della Sera	64,197,579	95.01
OVERALL	407,729,465	94.83

Table 2: Tokens and Texts/Posts ratio for page.

3 Topics in FB-NEWS15

FB-NEWS15 contains texts referring to a large variety of events. In order to organize the corpus into clusters of thematically related news, we used LDA (Blei et al., 2003). LDA represents documents as random mixtures over latent topics, where each topic is characterized by a distribution over words. These random mixtures express a document semantic content, and document similarity can be estimated by looking at how similar the corresponding topic mixtures are. For the topic identification we used the software Mallet (McCallum, 2002).

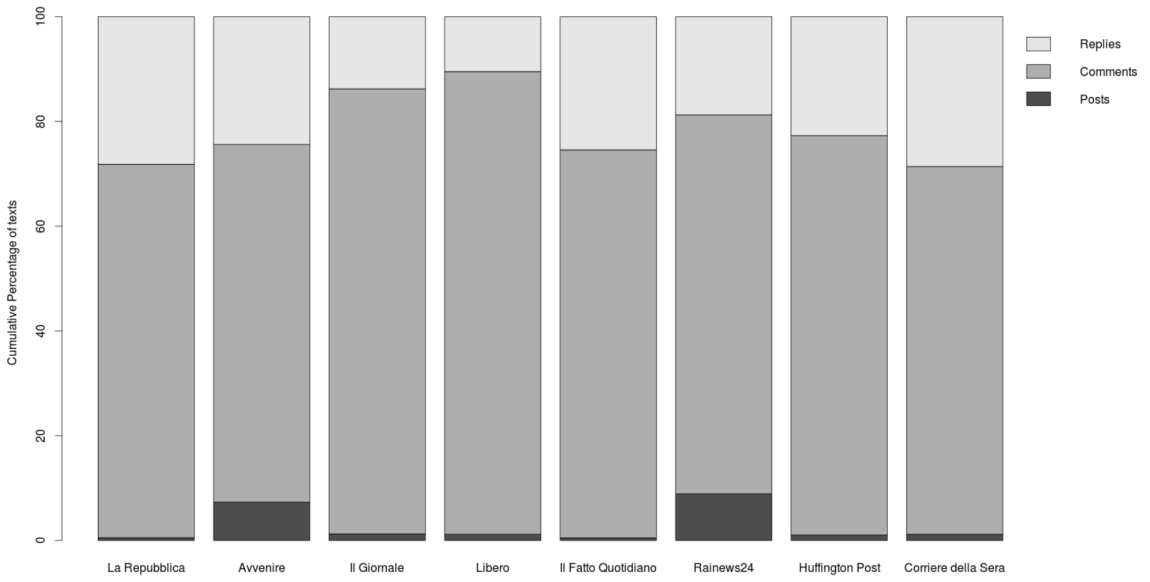


Figure 3: Cumulative distribution of posts, comments and replies in FB-NEWS15 for each Newspaper.

3.1 Selecting the vocabulary

Since we were interested in extracting the topics from the news articles, we have built the model on the portion of FB-NEWS15 containing the posts (FB-NEWS15_posts) published by the newspaper. In particular, we used entropy (Dumais, 1990) as a global term weighting and we selected for training the terms (nouns, adjectives, verbs and complex terms) with a high informative value (threshold fixed to 0.3), while using the remaining words as stopwords in Mallet (McCallum, 2002).

3.2 Extracting topics from posts

In order to determine the most debated topics in 2015, we used LDA to assign 50 topics to the posts in FB-NEWS15_posts and we navigated the graph to assign the topics to the comments and the replies. Later, we restricted the topics associated to a post P to the topics T having a probability higher than the 90^{th} percentile of the topic distribution of P. In this way, each post has been assigned, on average, to 3.06 topics. Finally, comments and replies have inherited the probability of belonging to the topic T from their parent post. Among the extracted topics ranked according to the sum of these probabilities we can find national and foreign politics, terrorism and church but also food, football, cinema and weather forecast. We report some topics below, with the number of texts and the relative ranking (i.e., rank 1 is given to the topic with the higher number of texts).

NATIONAL POLITICS (2,516,640 TEXTS, RANK 1): {*Renzi, presidente, premier, Mattarella, riforma, Alfano, senato, camera, Boschi, aula*} (Renzi, president, Mattarella, reform, Alfano, senate, chamber, Boschi, hall)

SCHOOL (1,707,145 TEXTS, RANK 2):{*scuola, giovane, studente, protesta, corso, mancare, sospendere, inglese, spiegare, lezione*} (school, young, protest, class, lack, suspend, English, explain, lesson)

CRIME (1,543,735 TEXTS, RANK 7): {*uccidere, polizia, arrestare, fermare, sparare, uomo, poliziotto, colpo, ferire, agente*} (kill, police, detain, stop, open fire, man, policeman, bump, wound, police officer)

ISIS (1,267,749 TEXTS, RANK 16): {*Isis, guerra, siria, minaccia, U.S.A., Libia, colpire, islamico, usare, jihadisti* } (Isis, war, Syria, threat, U.S.A., Libya, damage, islamic, use, jihadist)

FOOD (949,520 TEXTS, RANK 40): {*mangiare, ricetta, cibo, preparare, consiglio, evitare, perfetto, trucco, salute, semplice*} (eat,

recipe, food, prepare, advice, avoid, perfect, trick, health, simple)

FOOTBALL (606,560 TEXTS, RANK 50): {*seguire la diretta, guardare il video, campo, calcio, serie, Napoli, Milan, segnare, battere, partita*} (follow the live, look at the video, football field, football, league, Naples, Milan)

4 Conclusions and ongoing work

As one of the most widespread social networks, Facebook offers the possibility to collect opinionated pieces of texts from people of different ages, cultures and education. The composition of FBNEWS15, in which each comment is explicitly associated with a particular post, allows us to study the differences in terms of readers' perceptions about a particular topic. Differently from other social media like Twitter, Facebook contains larger texts including lot of subjective expressions that are very useful for the construction of sentiment and emotive lexicons.

Starting from previous works (Passaro et al., 2015; Passaro and Lenci, 2016), we plan to use this corpus to build lexical resources for sentiment analysis and emotion detection, which will include both words and complex terms. In addition, we plan to optimize the topic modeling phase and to investigate the possibility of using the extracted topics as a prior for inferring the sentiment orientation of a particular comment.

References

A. Ahmad. 2010. Is twitter a useful tool for journalists? *Journal of Media Practice*, 11(2):145–155.

A. Ahmad. 2013. Whats in a tweet? foreign correspondents use of social media. *Journalism Practice*, 7(1):33–46.

G. Attardi, F. Dell'Orletta, M. Simi, and J. Turian. 2009. Accurate dependency parsing with a stacked multilayer perceptron. In *EVALITA 2009 Evaluation of NLP and Speech Tools for Italian 2009*, LNCS, Reggio Emilia (Italy). Springer.

D. M. Blei, A. Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

H. Chyi and D. L. Lasorsa. 2002. An explorative study on the market relation between online and print newspapers. *Journal of Media Economics*, 15(2):91–106.

F. Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. In *EVALITA 2009 Evaluation of NLP and Speech Tools for Italian 2009*, LNCS, Reggio Emilia (Italy). Springer.

S. T. Dumais. 1990. Enhancing performance in latent semantic indexing (lsi) retrieval. Technical Report TM-ARH-017527.

A. Hermida. 2013. #journalism. reconfiguring journalism research about twitter, one tweet at a time. *Digital Journalism*.

H. Lin and L. Qiu. 2013. Two sites, two voices: Linguistic differences between facebook status updates and tweets. In P. L. Patrick Rau, editor, *Cross-Cultural Design. Cultural Differences in Everyday Life: 5th International Conference, CCD 2013, Held as Part of HCI International 2013*, volume 2, pages 432–440, Las Vegas (USA). Springer Berlin Heidelberg.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

N. Newman, W. H. Dutton, and G. Blank. 2012. Social media in the changing ecology of news: The fourth and fifth estate in britain. *Internet Science*, 7(1):6–22.

Oriella. 2013. The new normal for news. have global media changed forever? *The 6th Annual Oriella Digital Journalism Survey*.

L. C. Passaro and A. Lenci. 2015. Extracting terms with extra. In *Proceedings of the EUROPHRAS 2015 Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, pages 188–196, Malaga (Spain).

Lucia C. Passaro and Alessandro Lenci. 2016. Evaluating context selection strategies to build emotive vector space models. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), may.

L. C. Passaro, L. Pollacci, and A. Lenci. 2015. Item: A vector space model to bootstrap an italian emotive lexicon. In *Proceedings of the second Italian Conference on Computational Linguistics CLiC-it 2015*, pages 215–220, Trento (Italy).

R. Picard. 2009. Blogs, tweets, social media, and the news business. *Nieman Reports*, 63(3):10–12.

D. V. Shah, J. Cho, W. P. Eveland, and N. Kwak. 2005. Information and expression in a digital age. *Communication Research*, 32(10):531–565.

M. L. Sheffer and B. Schultz. 2010. Paradigm shift or passing fad? twitter and sports journalism. *International journal of Sport Communication*, 3(4):472–484.

Sardinian on Facebook: Analysing Diatopic Varieties through Translated Lexical Lists

Irene Russo
ILC CNR
Pisa
irene.russo@ilc.cnr.it

Simone Pisano
Università Guglielmo Marconi
Roma
s.pisano@unimarconi.it

Claudia Soria
ILC CNR
Pisa
claudia.soria@ilc.cnr.it

Abstract

English. Presence of regional and minority languages over digital media is an indicator of their vitality. In this paper, we want to investigate quantitative aspects of the use on Facebook of the Sardinian language. In particular, we want to focus on the co-existence of diatopic varieties. We extracted linguistic data from public pages and, through the translation of the most frequent words, we find out similarities and differences between varieties.

Italiano. *La presenza e l' uso delle lingue regionali e minoritarie sui mezzi digitali è un indicatore della loro vitalità. In questo lavoro vogliamo concentrarci sugli aspetti quantitativi del sardo usato su Facebook. In particolare, vogliamo analizzare le varietà diatopiche estraendo i dati linguistici dalle pagine pubbliche. Mediante la traduzione delle parole più frequenti abbiamo trovato similarità e differenze tra le varietà.*

1 Introduction

Everyday life makes an increasingly extensive use of digital devices that involve language use; for this reason, usability of a language over digital devices is a sign for that language of being modern, relevant to current lifestyles and capable of facing the needs of the XXI century. A positive correlation between presence in new technologies and better appreciation of a language has been repeatedly observed in the literature, see for instance (Eisenlohr, 2004) and (Crystal, 2010). Regional and minority languages (RMLs henceforth) are usually very poorly represented digitally (Soria, 2016).

Since poor digital representation of regional and minority languages further prevents their usability on digital media and devices, it is extremely important to enhance every bottom-up effort that can boost the quantity of available digital content. In fact, if the perception of the marginal role and limited applicability of RMLs persists, their attractiveness diminishes.

An increase in quantity of digital content available online represents today an opportunity for regional and minority languages. Online speakers can make visible the existence of a community that uses the language to interact; they can use online communication to converge toward a standard and they can instruct less skilled speakers toward better mastering of the rules of the language, especially when the language is not formally included in education. From the perspective of computational linguistics, the presence of digital content written in RMLs means that corpora can be built for them and basic tools (lemmatizers, spell checkers, lexicons etc.) can be developed.

The presence of RMLs over digital media and their usability through digital devices is often limited to instances of digital activism and/or by means of cultural initiatives focused on the preservation of cultural heritage.

In this paper we promote the first study we are aware of about the use on social networks (more specifically, Facebook) of Sardinian, an Italian minority language characterised by the coexistence of varieties and the difficulties for the promoted standard to emerge as unifying factor. Our starting hypothesis concerned the vitality on social networks of a language that is mainly spoken. With the help of a Sardinian linguist, we identi-

fied a small set of FB public groups where specific varieties of Sardinian are chosen as their main language plus groups where generic, not further defined Sardinian is used to communicate. We extracted messages from these pages and created a frequency lexicon for each variety. The most frequent 150 words have been translated by a Sardophone expert linguist with the aim of finding differences and commonalities between varieties. This preliminary analysis is the first step toward the use of computational linguistics methodologies in the promotion of a standard for Sardinian based on quantitative data.

2 Sardinian today: Main Varieties and Standardization Efforts

Sardinian is an autonomous Romance language spoken in the island of Sardinia. According to (Lupinu, 2007) it is known by approximately 68,4% of the population of the island. Ethnologue[1] lists four varieties for Sardinian: Northwestern Sardinian or Sassarese (100,000 speakers ca.), Campidanese (500,000 speakers ca.), Central Sardinian or Logudorese (500,000 speakers ca.) and Gallurese (100.000 speakers ca.)

The most important differences from a lexical, phonological and morphological point of view within Sardinian can be found between Central-Southern and Central-Northern dialects.

Scholars use to divide Sardinian in two main varieties: Logudorese and Campidanese, the first one spoken in the North and in the center of the island and the second one spoken in the South.

Logudorese and Campidanese can be related to two different pre-existing written standards: the so-called *Logudorese* (or *Logudorese illustre*) was used for the first time in a short poem at the end of the XV century (Manca, 2002), whereas what is known as Campidanese was the language of some religious plays at the end of the XVII Century (De Martini Abdullah Luca, 2006).

Today, Sardinian lacks of a generally agreed standard variety, although standardization efforts characterised the recent history of the Region.

The first attempt to introduce a written system based on an integration of phonetic, lexical and morphological features of modern Sardinian varieties was made in 2001, when the basic rules of LSU (*Limba Sarda Unificada*,Unified Sardinian Language) were presented (Blasco Ferrer, 2001). This proposal was sharply criticised by some sectors of the public opinion and strong disapproval came even from a part of native speakers, especially from the South, who considered this standard too much different from the language they spoke. It is a fact that it never became a model of official Sardinian.

In 2006, another model of written language was made official by the Regional Committee resolution n°16/14. This standard, called LSC (*Limba Sarda Comuna*, Common Sardinian Language)[2] made the effort of taking into account also the dialects of the transition region of the center mentioned earlier. Although regional administration recommended its use for written public documents it is still reluctantly accepted by some speakers, who perceive it as too distant from the varieties they speak.

In 2010, the Provincial Council of Cagliari took a different course choosing with the Provincial Committee resolution n°17 a linguistic norm[3] based on literary language of Southern poets and writers, in order to draw up acts, documents and even textbooks for primary children.

All these standardization efforts, politically guided or emerged bottom-up, clearly show that Sardinian speakers are aware of the role of standard orthography and grammar for the vitality and the survival of their language. On the one hand, they want to promote the idea of a unique language as a matter of identity; on the other, they dont want to lose local peculiarities by adopting standard rules that inevitably hide some local differences.

Social media are widely used by Sardinian speakers and they represent an interesting scenario for written but informal use of the language. An in-depth analysis of the type of language used by Sardinian speakers on social media is still missing. Certainly, use of everyday Sardinian in spoken and written (online) informal communication, is a sign of vitality of the language. Interaction is a powerful instrument for standardization, and the interactive modality offered by social media could reveal the emergence of coordination strate-

[1] www.ethnologue.com

[2] Regione Autonoma della Sardegna (2006), Limba Sarda Comuna. Norme linguistiche di riferimento a carattere sperimentale per la lingua scritta dellAmministrazione regionale, Cagliari, Regione Autonoma della Sardegna.

[3] Arrègulas po sortografia, sa fonètica, sa morfologia e su fueddàriu de sa bariedadi Campidanesa de sa lìngua sarda (Rules for orthography, phonetic, morphology and the vocabulary of Campidanese variety of Sardinian language)

gies toward a standard in speakers community as a natural need (Burghardt, 2016). To check this hypothesis, we started to analyse the use of different varieties of Sardinian that is being made on Facebook. According to the preliminary data of a recent survey, Facebook is the social media that is most used by Sardinian speakers, and where Sardinian is actively and extensively used[4].

3 Data Extraction and Analysis

We selected public pages and communities on Facebook that are rich in content and interactions between users. With the help of a Sardinian linguist we identified four mutually exclusive sets:

- pages where people communicate in LSC;
- pages where people communicate in Sardinian without further specification of the chosen variety;
- pages where people communicate in Campidanese;
- pages where people communicate choosing a local variety (in our case Nugoresu, local variety of Logudorese).

All the messages have been extracted from the json of the pages obtained through Facebook API. Lowercase texts have been tokenized splitting on whitespaces. Four frequency lists have been created, emoticons and symbols have been deleted. The 150 most frequent words have been translated in Italian by a Sardinian linguist that provided also PoS and morphological annotation plus all the available translations in case of polysemous words. We left in these lists Italian words because every cleaning procedure (lists of Italian words, PoS for Italian etc.) was risky: very frequent words in Sardinian can be found in Italian too (e.g. *a, chi, bonus, cosa*) with a different meaning.
Table 1 reports basic statistics about public pages and communities in the four sets listed above. Active users are the ones who wrote at least one message on the page. Number of active users and messages varies for each set but it was not possible to get a balanced sample.
In Table 2 the number of tokens and types for the four sets of Facebook groups analysed are reported. In Table 3 each possible pair of varieties is compared by checking the overlapping of translations into Italian. The second column reports how many Italian types are in common between two varieties. For example, among the most frequent 150 LSC word forms and the 150 most frequent *Sardu* word forms, 61 words have the same Italian translation. The third column contains the number of words with the same word forms in the two varieties compared, e.g. the Italian adjective *grande* has the same word form (*mannu*) in Nugoresu and Campidanese. This is a first attempt to understand if two varieties are close orthographically, considering the orthographic forms of the analysed words. We also report the number of content words found in each pair because we believe that in the future the overlapping at orthographic level should be analysed taking into account the distinction between content and function words.
The fourth column contains the number of the word forms related to the types in common which are different in the two varieties e.g. for the Italian word *è*, third singular person of verb *to be* in the present form, LSC has just one word *est*, while Campidanese has *est* and *esti*. In this case *esti* is counted as a different form and is included in the table under the fourth column.
Table 4 summarises for each pair how variability patterns are distributed, where pattern 1 to 2 means that there is one word form for variety *a* that correspond to two word forms for variety *b*. We know that the group *Sardu* contains data from more than one variety and we plan as future work a more detailed analysis. For the moment we note a clear overlapping because speakers of LSC contribute with posts and comments on pages where people communicate in Sardinian. For the same reason, when *Sardu* is one of the item in the pair we notice more variability patterns (see Table 4).
Concerning the comparisons between LSC and the two main varieties Campidanese and Logudorese, represented in our data by the local variety Nugoresu, we found evidence of the distance between the two main varieties with an overlapping of 41,5% in terms of word forms. LSC and Campidanese have an overlapping of 64,2% while LSC and Nugoresu have an overlapping of 83%. LSC emerges as a variety that tried to set a linguistic common ground and achieved this result, even if there is a bias toward Logudorese variety, one of

[4]Preliminary data of the DLDP Survey (www.dldp.eu) "Su Sardu: una limba digitale?". In July 2016, Facebook appears to be used by 98,1% of the respondents. Of those, 44% use Sardinian for writing and reading posts and messages, and 32,5% only for reading.

page name	type	#users	#active users	#messages	#variety
LSC, Limba Sarda Comuna: Sotziedade pro sa limba sarda comuna	Community	590	27	160	LSC
Iscritores in limba sarda	Public Group	331	49	916	LSC
Amigosde-sa-Limba-Sarda-Comuna	Community	1673	13	40	LSC
Solu in sardu	Public Group	15890	5701	373430	generic Sardinian
Solu poesias	Public Group	2018	158	1679	generic Sardinian
Scrieusu in campidanesu	Public Group	1984	576	17960	Campidanese
Cabuderra lngua e cultura	Public Group	116	1	18	Campidanese
Sos chi li piacheta faveddare e a iscrivere nugoresu	Public Group	984	438	1157	Nugoresu

Table 1: Basic statistics about data extracted.

FBgroup	tokens	types fr >10
LSC	71018	847
Sardu	3300408	18248
Campidanese	257110	2285
Nugoresu	379802	3412

Table 2: Basic statistics about token and types.

the complaints of Campidanese speakers (see par. 2).

4 Conclusion and Future Work

In this paper we address the following open question: could quantitative analysis of written data help Sardinian community to find out a common core (not specific of a variety) that could reinvigorate the idea of a standard? We plan future work on this issue, with the awareness that digital content on social media is both an opportunity and a challenge for this kind of analyses.

This paper is a first analysis of diatopic varieties of Sardinian through orthographical comparisons of word forms with the same meaning. Thanks to translated lists it was possible to look at commonalities and differences between varieties. Social media are a source of real data about language uses and the best observatory for regional and minority languages. Concerning Sardinian Facebook offers the possibility to test the distance between the proposed orthographic standard and the existing varieties. We will test the interplay between varieties with other methodologies to measure the distance and to find out usage patterns (e.g. Levenshtein distance for similar words).

This work is being carried out in the framework of the project DLDP (Digital Language Diversity Project, http://www.dldp.eu). DLDP is a three year project funded under the Erasmus+ programme. It aims at addressing the problem of low digital representation of EU regional and minority languages by giving their speakers the intellectual and practical skills to create, share, and reuse online digital content. DLDP fully embraces a bottom-up approach to language revitalization by addressing the speakers cognitive and practical skills as the cornerstone of effective revitalization initiatives.

Acknowledgments

This work is partially funded by the Erasmus + DLDP Project (Grant Agreement no. 2015-1-IT02-KA204-015090). The opinions expressed reflect only the authors view and the Erasmus+ National Agency and the Commission are not responsible for any use that may be made of the information contained.

References

Blasco Ferrer E., Bolognesi, R. et al. 2001. *Limba Sarda Unificada. Sintesi delle norme di base: ortografia, fonetica, morfologia, lessico*. Cagliari, Regione Autonoma della Sardegna.

Burghardt, M., Granvogl, D. and Wolff, C. 2016. Creating a Lexicon of Bavarian Dialect by Means of Facebook Language Data and Crowdsourcing. *Proceedings of LREC-2016*. Portoroz, Slovenia.

Crystal, D. 2010. *Language Death*. Cambridge University Press.

De Martini Abdullah Luca (ed.) 2001. *Libro de Comedias (by Antonio Maria da Esterzili)*. Cagliari, Cuec.

Eisenlohr, P. 2004. Language revitalization and new technologies: Cultures and electronic mediation and the refiguring of communities. *Annual Review of Anthropology*. 18(3):339361.

Lupinu, G., Mongili, A. , Oppo, A. , Spiga, R. , Perra, S. , Valdes, M. 2007. *Le lingue dei sardi: una ricerca sociolinguistica*. Assessorato alla Pubblica istruzione, beni culturali, informazione, spettacolo e sport, Regione Autonoma della Sardegna.

De Martini Abdullah Luca (ed.) 2002. *Sa Vitta et sa Morte, et Passione de sanctu Gavinu, Prothu et Januariu (by Antonio Cano)*. Cagliari, Cuec.

	common_types	types_with_same_word_forms	types_with_different_word_forms
LSC-Sardu	61	60 (21 content words)	32 (11 content words)
LSC-Campidanese	67	43 (14 content words)	44 (17 content words)
LSC-Nugoresu	65	54 (14 content words)	39 (17 content words)
Sardu-Campidanese	65	47 (15 content words)	64 (26 content words)
Sardu-Nugoresu	70	64 (16 content words)	65 (33 content words)
Campidanese-Nugoresu	81	34 (12 content words)	82 (27 content words)

Table 3: Comparison between Sardinian varieties.

	1 to 2	1 to 3	1 to 4
LSC - Sardu	8	4	1
LSC - Campidanese	3	0	0
LSC - Nugoresu	7	1	1
Sardu - Campidanese	5	1	5
Sardu - Noguruse	8	3	5
Campidanese - Nugoresu	2	0	1

Table 4: Comparison between Sardinian varieties.

Soria, C., Russo, I. , Quochi, V., Hicks, D., Gurrutxaga, A., Sarhimaa, A. and Tuomisto, M. 2016. Fostering digital representation of EU regional and minority languages: the Digital Language Diversity Project. *Proceedings of LREC-2016*. Portoroz, Slovenia.

Virdis, M. 1988. Sardisch: Areallinguistik / Aree linguistiche. Holtus G., Metzeltin, M., Schmitt, C. (eds.), *Lexicon der Romanistischen Linguistik* 4, Tubingen, Max Niemeyer, pp. 897-913.

Wagner, M. L. 1997. *La lingua Sarda. Storia, spirito e forma*. Nuoro, Ilisso.

A New Dataset for Source Code Comment Coherence

Anna Corazza
DIETI,
Univ. di Napoli Federico II
anna.corazza@unina.it

Valerio Maggio
Fondazione
Bruno Kessler
vmaggio@fbk.eu

Giuseppe Scanniello
Dept. of Mathematics, Information
Technology, and Economics
Univ. della Basilicata
giuseppe.scanniello@unibas.it

Abstract

English. Source code comments provide useful insights on a codebase and on the intent behind design decisions and goals. Often, the information provided in the comment of a method and in its corresponding implementation may be not coherent with each other (i.e., the comment does not properly describe the implementation). Several could be the motivations for this issue (e.g., comment and source code do not evolve coherently). In this paper, we present the results of a manual assessment on the coherence between comments and implementations of $3,636$ methods, gathered from 4 Java open-source software. The results of this assessment has been collected in a dataset that we made publicly available on the web. We also sketch here the protocol to create this dataset.

Italiano. *I commenti al codice sorgente forniscono informazioni utili sull'implementazione del codice e sulle intenzioni relative alle decisioni e agli scopi del progetto. Spesso, le informazioni presenti nel commento di un metodo e nella sua implementazione possono non essere coerenti (nel senso che il commento non dà una descrizione adeguata dell'implementazione). Ci possono essere diverse spiegazioni per questo (ad esempio, commenti e codice sorgente non sono stati modificati in modo coerente). In questo articolo, presentiamo i risultati di una valutazione manuale della coerenza tra commenti e implementazione di* $3,636$ *metodi, raccolti da* 4 *applicazioni open source in Java. I risultati di questa valutazione sono stati raccolti in un dataset che abbiamo pubblicato sul web. Accenniamo anche al protocollo seguito per la preparazione del dataset.*

1 Introduction

Natural language is used in different ways in the development process of a software system and therefore techniques of natural language processing (NLP) and information retrieval (IR) are more and more frequently integrated into software development and maintenance tools. Many automated or semi-automated techniques have been proposed to aid developers in the comprehension and in the evolution of existing systems, e.g., (Corazza et al., 2016; Scanniello et al., 2010).

Natural language information is provided in source code comments and in the name of identifiers. In the former case, standard natural language is usually adopted, although quite technical. Comments are written in English, even when developers have different mother-tongues. On the other hand, identifiers are typically constructed by composing multiple terms and abbreviations. Therefore, more sophisticated techniques are necessary to extract the lexical information contained in each identifier (Corazza et al., 2012).

Most of these techniques assume that the same words are used whenever referring to a particular concept (Lawrie et al., 2010). In many cases, this represents an oversimplification: methods are often modified without updating the corresponding comments (Salviulo and Scanniello, 2014). In these cases, comments might convey information unrelated or inconsistent with the corresponding implementation. Nevertheless, comments are extremely important because they are

expected to convey the main intent behind design decisions, along with some implementation details (e.g., types of parameters and of returned values).

Therefore, more sophisticated models are necessary to determine if there is coherence between the lexicon provided in comments and in its corresponding source code. Hence, there exists coherence between a lead comment of a method and its source code (also simply coherence, from here on) if that comment describes the intent of the method and its actual implementation.

In this work, we focus on the lead comment of methods. This kind of comments precedes the definition of a given method and is supposed to provide its documentation and details about the implementation. We discuss here a dataset we made publicly available on the web.[1] It contains annotations about the coherence of $3,636$ methods collected from 4 implementations of 3 open source projects written in Java. The defined protocol used for its creation is also sketched, to give researchers the opportunity to possibly extend it. Further details on this protocol can be found in (Corazza et al., 2015).

For the assessment of the quality of the annotation with special focus on computational linguistics applications, a few indexes have been considered (Eugenio and Glass, 2004; Artstein and Poesio, 2008; Mathet et al., 2015), among which the *kappa index* (Cohen, 1960) is the most widely adopted because of its favorable characteristics. The inter-annotator agreement has therefore been assessed by this parameter.

We expect that making freely available this dataset could give impulse to the research for approaches to assess the coherence between the implementation of a method and its lead comment (simply coherence, from here on). In fact, although no approach has been yet proposed in this regard, they could be of great help for software maintenance and evolution activities.

The paper is structured as follows. In Section 2, we discuss the methodology used to create our dataset. A description of the main characteristics of the dataset is given in Section 3, while in Section 4 the annotation is assessed. Some final considerations conclude the paper.

[1] www2.unibas.it/gscanniello/coherence/

2 Dataset Construction

To create our dataset, we adopted the perspective-based and the checklist-based review methods (Wohlin et al., 2012). The perspective is the one of the Researcher aiming at assessing the coherence between the lead comment of a method and its implementation. The process of creation is based on the following elements:

1. **Working Meetings.** We used meetings to determine the goals of our research work and the process to create the dataset.
2. **Dataset Creation.** We instantiated the defined process to create our dataset.
3. **Outcomes.** We gathered results during and after the creation of our dataset.
4. **Publishing Results and Dataset.** We shared our experience with the community in (Corazza et al., 2015) and released the dataset on the web.

The construction of the dataset has been completed in two main consecutive phases by using an ad-hoc web system implemented for the purpose:

- **Verify coherence**. Annotators verify by means of a checklist the coherence between the lead comments of a set of methods and their corresponding implementation.
- **Resolve conflicts**. The intervention of experts is required whenever the judgements of the annotators differ. In our case, two of the authors, with a background in software engineering, assumed the role of experts and examined the problematic cases. For each conflicting method, the experts should reach an agreement about the coherence or the non-coherence. Methods on which experts do not get a consensus are automatically discarded.

3 Dataset Description

Some descriptive statistics (e.g., number of classes and methods) of the software systems in our dataset are shown in Table 1.

- **CoffeeMaker**[2] is a software to manage inventory and recipes and to purchase beverages. We chose this software because it has

[2] agile.csc.ncsu.edu/SEMaterials/tutorials/coffee_maker/

Table 1: Descriptive statistics of the software systems in the dataset: N_f stands for the number of files, N_c for the number of classes, N_m for the number of methods and N_m^* for the number of methods with lead comments.

System	Version	N_f	N_c	N_m	N_m^*
CoffeeMaker	-	7	7	51	47 (92%)
JFreeChart 6.0	0.6.0	82	83	617	485 (79%)
JFreeChart 7.1	0.7.1	124	127	807	624 (77%)
JHotDraw	7.4.1	575	692	6414	2480 (39%)

a simple and clear design (being it developed for educational purposes).

- **JFreechart**[3] is a Java tool supporting the visualization of data charts (e.g., scatter plots and histograms). We included two versions of this software. As reported in Table 1, both these versions contain almost the 80% of methods with lead comments. This suggests an extensive use of comments, which is the main reason why we decided to include this software in the dataset.

- **JHotDraw**[4] is a framework for technical and structured graphics. Even if the source code of JHotDraw is scarcely commented (see Table 1), it is well-known in the software maintenance community due to its good Object-Oriented design (Scanniello et al., 2010).

In Figure 1, we report the implementation and the lead comment of `setTickLabelsVisible` (extracted from JFreeChart ver. 0.7.1, included in our dataset). According to our definition of coherence, we can assert that this method is coherent. On the other hand, the `save` method reported in Figure 2 provides a very poor and inadequate description of the design intent of the method, thus reflecting a lack of coherence with the underlying implementation.

Three annotators were involved in the dataset creation process. Two of them hold a Bachelor degree in Computer Science, and have very similar technical backgrounds. On the other hand, the third annotator can be considered more experienced than the other two since he holds a Master degree in Computer Science. We distributed the effort among the annotators so that each software

[3] `www.jfree.org/jfreechart/`

[4] `www.jhotdraw.org/`

```
/**
 * Sets the flag that determines whether or not
 * the tick labels are visible.
 * Registered listeners are notified of a
 * general change to the axis.
 *
 * @param flag The flag to set.
 */
public void setTickLabelsVisible(boolean flag) {
    if (flag!=tickLabelsVisible) {
        tickLabelsVisible = flag;
        notifyListeners(new AxisChangeEvent(this));}
}
```

Figure 1: The lead comment and the implementation of the method `setTickLabelsVisible` of JFreeChart 0.7.1

```
// GEN-FIRST:event_save
// Code for dispatching events from components
// to event handlers.
private void save(java.awt.event.ActionEvent evt) {
try {
  String methodName = getParameter("datawrite");
  if (methodName.indexOf('(') > 0) {
        methodName = methodName.substring(0,
                methodName.indexOf('(') - 1); }
  JSObject win = JSObject.getWindow(this);
  Object result = win.call(methodName, new
           Object[]{getData()});
} catch (Throwable t) {
  TextFigure tf = new TextFigure("Fehler:␣" + t);
  AffineTransform tx = new AffineTransform();
  tx.translate(10, 20);
  tf.transform(tx);
  getDrawing().add(tf); }
}
```

Figure 2: Non-Coherent method in JHotDraw 7.4.1

would be separately evaluated by at least two annotators. This allowed us to have multiple judgements for each method in the dataset, and to calculate the rate of agreement among annotators.

4 Annotation Assessment

The whole dataset creation process occurred from January, 15^{th} 2014 to June, 20^{th} 2014, for a total of 800 man-hours. This gives an estimation of the effort required to conduct the study presented in this paper, and provides an indication to the researcher interested in extending our dataset.

The annotators provided indications on the coherence of methods by assigning them one out of three following possible values: *Non-Coherent*, *Don't Know*, and *Coherent*.

In this scenario, we use the *kappa index* (Cohen, 1960) to obtain an assessment of the agreement among annotators, thus estimating the reliability of their evaluations. In fact, if annotators agree on a large number of methods, we can conclude

that their annotations are reliable. The kappa index is designed for categorical judgments and refers the agreement rate calculation to the rate of chance agreement:

$$kappa = \frac{p_o - p_c}{1 - p_c}, \quad (1)$$

p_o is the observed probability of agreement, while p_c is the chance probability of agreement. Both probabilities are estimated by the corresponding frequencies. By a simple algebraic manipulation, Equation 1 can be written as:

$$kappa = 1 - \frac{q_o}{q_c}, \quad (2)$$

where $q_o = 1 - p_o$ and $q_c = 1 - p_c$ and correspond to the observed and the chance probabilities of *disagreement*, respectively. Usually the index assumes values in $]0, 1]$[5], as it can be expected that the observed disagreement is less likely than chance. A null value signals that observed disagreement is exactly as likely as chance, while the kappa index assumes negative values in the unwanted case where disagreement is more likely than chance. Perfect agreement corresponds to $k = 1$. Values greater than 0.80 are usually considered as a cue of good agreement. Values in the interval $[0.67, 0.80]$ are considered acceptable (Cohen, 1960).

The classical formulation of the kappa index considers a binary classification problem (e.g., *Non-Coherent* or *Coherent*). However in our case, the neutral judgement (i.e., *Don't know*) is also allowed. Therefore, possible disagreements include the case where one of the two answers is the neutral one. In this case, it is possible to differently weigh the possible disagreements among annotators. In fact, disagreements due to the neutral answers are less serious than disagreements where judgments are totally divergent (i.e., *Coherent* and *Non-Coherent*, in our case). To this end, Cohen (Cohen, 1968) presents a variant of the kappa index, where in case of a disagreement, different weights can be applied. In case the same weight is assigned to all possible disagreement combinations, the original (unweighted) formulation is obtained. The formulation of the *Weighted Kappa* (WK) is the one in the equation (2), but for the computation of q_o and q_c the contributions are weighted according to the importance given to the corresponding disagreement cases. By contrast, we refer to the original formulation of the kappa index as *Unweighted Kappa* (UK). We assign to the *Don't know* response a weight that is half the weight assigned to the *Not-Coherent* (or *Coherent*) one. This is the same schema reported by Cohen (Cohen, 1968). Weighted and Unweighted kappa indexes are reported in Table 2.

[5]The notation means that 1 is included in the interval, while 0 is not.

Table 2: Agreement Rate of Judges as computed by the Cohen's Kappa Index.

System	UK	WK
CoffeeMaker	0.913	0.913
JFreeChart 6.0	0.939	0.918
JFreeChart 7.1	0.983	0.977
JHotDraw	0.824	0.684

The agreement between annotators is good on the first three systems, and acceptable for JHotDraw. However, on this system the difference between the values for UK and WK is large, thus providing a more accurate indication on the agreement of the evaluations on this software.

At the end of the first step of the dataset creation process, the number of methods on which annotators did not agree was 302, corresponding to the 8.3% of the total number of methods from all the systems in the dataset. Most of these methods are those in JHotDraw, as suggested by the kappa index values (see Table 2). These methods were reviewed by two of the authors. An agreement was reached on all of these methods, which were then included in the dataset. The total number of methods in the dataset is reported in Table 3 (i.e., $2,883$).

5 Conclusions and future work

In this paper, we have presented the early steps of our research on the coherence between the lead comment of methods and their implementations. In particular, we have provided a description of the problem settings, along with the experimental protocol defined to create our dataset. We made it publicly available on the web. We also sketched the results of quantitative analysis conducted on a codebase of $3,636$ methods, gathered from 4 different open-source systems written in Java.

There could be many possible future directions for our research. For example, it would be interesting to conduct an empirical study to investigate

Table 3: Descriptive Statistics of the Dataset: C stands for Coherent, NC for Non-Coherent.

System	C	NC	Total	Don't Know (Not included)
CoffeeMaker	27	20	**47**	0
JFreeChart 6.0	406	55	**461**	24
JFreeChart 7.1	520	68	**588**	36
JHotDraw	762	1025	**1787**	693
Total	**1715**	**1168**	**2883**	

the effect of maintenance operations on the coherence for multiple versions of the same system. As a step forward in this future research direction, we already included in the dataset two versions of the JFreeChart system. Our results and those by Fluri et al. (Fluri et al., 2007) represent a viable starting point. Finally, we would like to exploit the collected data as an evaluation set to assess the performance of approaches able to discern method coherence.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, December.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

J. Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*.

A. Corazza, S. Di Martino, and V. Maggio. 2012. LINSEN: An efficient approach to split identifiers and expand abbreviations. In *Proceedings of International Conference on Software Maintenance*, pages 233–242. IEEE Computer Society.

A. Corazza, V. Maggio, and G. Scanniello. 2015. On the coherence between comments and implementations in source code. In *41st Euromicro Conference on Software Engineering and Advanced Applications, EUROMICRO-SEAA 2015, Madeira, Portugal, August 26-28, 2015*, pages 76–83. IEEE Computer Society.

A. Corazza, S. Di Martino, V. Maggio, and G. Scanniello. 2016. Weighing lexical information for software clustering in the context of architecture recovery. *Empirical Software Engineering*, 21(1):72–103.

Barbara Di Eugenio and Michael Glass. 2004. The Kappa statistic: a second look. *Computational Linguistics*, 30(1).

B. Fluri, M. Wursch, and H.C. Gall. 2007. Do code and comments co-evolve? on the relation between source code and comment changes. In *Proceedings of the Working Conference on Reverse Engineering*, pages 70–79. IEEE Computer Society.

D Lawrie, D Binkley, and C Morrell. 2010. Normalizing Source Code Vocabulary. In *Proceedings of Working Conference on Reverse Engineering*, pages 3–12. IEEE Computer Society.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma γ for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479, September.

F. Salviulo and G. Scanniello. 2014. Dealing with identifiers and comments in source code comprehension and maintenance: Results from an ethnographically-informed study with students and professionals. In *Proceedings of International Conference on Evaluation and Assessment in Software Engineering*, pages 423–432. ACM Press.

G. Scanniello, A. D'Amico, C. D'Amico, and T. D'Amico. 2010. Using the kleinberg algorithm and vector space model for software system clustering. In *Proceedings of International Conference on Program Comprehension*, pages 180–189. IEEE Computer Society.

C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, and A. Wesslén. 2012. *Experimentation in Software Engineering*. Computer Science. Springer.

Tracing metaphors in time through self-distance in vector spaces

Marco Del Tredici
ILLC, Univ. of Amsterdam
Amsterdam, The Netherlands
marcodeltredici@gmail.com

Malvina Nissim
CLCG, Univ. of Groningen
Groningen, The Netherlands
m.nissim@rug.nl

Andrea Zaninello
Zanichelli editore
Bologna, Italy
azaninello@zanichelli.it

Abstract

English. From a diachronic corpus of Italian, we build consecutive vector spaces in time and use them to compare a term's cosine similarity to itself in different time spans. We assume that a drop in similarity might be related to the emergence of a metaphorical sense at a given time. Similarity-based observations are matched to the actual year when a figurative meaning was documented in a reference dictionary and through manual inspection of corpus occurrences.

Italiano. *Nel presente esperimento costruiamo spazi vettoriali progressivi nel tempo su un corpus diacronico dell'italiano e calcoliamo la distanza di alcuni termini rispetto a loro stessi in differenti periodi. L'ipotesi è che un calo di similitudine possa essere indicativo dell'acquisizione di un significato metaforico. Tale ipotesi è valutata attraverso una risorsa lessicografica esterna e l'annotazione manuale dei contesti dei termini nel corpus.*

1 Introduction

It is widely acknowledged that metaphors are pervasive in language use, and that their detection and interpretation are crucial to language processing (Group, 2007; Turney et al., 2011; Shutova, 2015).

One tricky aspect related to metaphors is their dynamic nature: new metaphors are created all the time. For example, in recent years the Italian term "talebano" ('Taliban'), previously only used to refer to the Islamic fundamentalist political movement founded in the Nineties in Afghanistan (Example 1), has come to define more generally someone who is extreme in his or her positions, for example regarding food, use of medicines, and the like (Example 2).[1]

(1) (*lit.*) l'operazione [...] ha permesso di arrestare un **talebano** esperto in esplosivi

(2) (*fig.*) [...] senza l'atteso top player, e di un allenatore **talebano** della tattica

If the metaphorical meaning becomes commonly used, it might get recorded in reference dictionaries, too. Indeed, for the case of "talebano" the Italian dictionary Zingarelli (Zingarelli, 1993–2017) has recorded the metaphorical extension ("che (o chi) è dogmatico, integralista") in the year 2009, while until then only the literal meaning was included.

Most of the computational work on metaphors has focused on their identification and interpretation using a variety of techniques and models, such as clustering (Shutova and Sun, 2013), LDA topic modeling (Heintz et al., 2013), tree kernels (Hovy et al., 2013), but all from a purely synchronic perspective.[2] The way metaphors develop across time, instead, and whether the shift of a word's literal meaning to a figurative one can be automatically detected and modelled is as of now a little investigated aspect.

As a contribution in this sense, we build on the basic observation that if a metaphorical meaning is acquired by a term at a certain point in time, the context of use of that term will, at least partially, change. In this paper we offer a proof of concept of this assumption, based on a selection of terms. (Dis)similarity of contexts is measured relying on the distributional semantics approach, and thus on the terms' vector representations, and the existence of a metaphoric shift is derived from the Zingarelli dictionary of Italian.

[1] All of the examples in this paper are from the newspaper *la Repubblica*, see Section 4.2.

[2] For a detailed survey on current NLP systems for metaphor modeling see (Shutova, 2015).

2 Approach

According to the principle of distributional semantics, the meaning of a word is represented by vectors that encode the contextual information of that word in a corpus (Turney et al., 2010). All vectors representing words are included in a distributional semantic space in which similar words are represented by vectors that are close in that space, while different words are distant.

We rely on the intuition that if a term develops a metaphoric sense, its contexts of occurrence will start to differ, at least partially, from those observed for the very same term at the time the metaphorical meaning had not emerged yet. This implies that detecting a distance in space across time could be indicative of a meaning shift. Hence, instead of comparing different terms synchronically, we focus on their *self-distance* across time, thus tracing their diachronic evolution of meaning.

Practically, we train vector representations of words in consecutive time spans, and compare such representations to one another, for a set of pilot terms. As a default, a term is expected to exhibit a vector representation roughly similar to itself across time. If we observe a drop in similarity between vectors in consecutive spaces, we can hypothesise the emergence of a new sense for this term, potentially metaphoric.

By using the information recorded for the selected terms in a reference dictionary for the Italian language, we observe whether there is some correspondence between the observed similarity drop, if present, and the time of inclusion of a figurative sense. Finally, for each year cluster, we manually inspect the occurrences of our target terms in order to see if changes of use can be observed.

We are aware of the fact that changes in distance of a word to itself across time might be triggered by phenomena other than the rise of a metaphoric shift. Indeed, especially for polysemous words, extra-linguistic factors could cause the dominance of one sense over the others at a given time. In a larger-scale, bottom-up approach to detect metaphorical shifts, this would need to be properly accounted for. In the context of this proof-of-concept, we control for this factor by choosing words that are not or are minimally polysemous (see Section 4.1).

3 Related Work

The automatic modelling of diachronic shift of meaning has been investigated employing several different techniques. Among these, most recently, Latent Semantic Analysis (Sagi et al., 2011; Jatowt and Duh, 2014), topic clustering (Wijaya and Yeniterzi, 2011) and dynamic topic modeling (Frermann and Lapata, 2016). Vector representations for diachronic shift of meaning have been used by Gulordava and Baroni (2011), with a simple co-occurence matrix of target words and context terms. Jatowt and Duh (2014) and Xu and Kemp (2015) experimented both with a bag-of-words approach and a more linguistically motivated representation that also captures the relative position of lexical items in relation to the target word.

Recently, Word Embeddings (Mikolov and Dean (2013), see also Section 4.3) have been used to investigate diachronic meaning shifts: vectors are usually created independently for each time span and then mapped from one year to another via a transformation matrix, thus leveraging the stability of the relative positions of vectors in different spaces (Kulkarni et al., 2015; Zhang et al., 2015; Hamilton et al., 2016).

An alternative approach, which we also adopt – with a slight change – in our work, is introduced by Kim et al. (2014), who propose a simple but effective methodology to make vectors trained on different corpora directly comparable: embeddings created for year y are used to initialise the vectors for year $y+1$. The process is progressively applied to all time spans.

4 Experiment

Following the approach described in Section 2, we selected a small set of pilot terms from a lexicographic reference, and observed their space development across time, on a diachronic corpus for Italian that we collected for this purpose. Due to the absence of datasets in which words are annotated for meaning change, a qualitative analysis of a set of hand-selected words like the one we propose has established itself as a common evaluation method in previous work on diachronic meaning change (Frermann and Lapata, 2016).

4.1 Lexicographic reference and term selection

The Zingarelli dictionary is a reference dictionary for the Italian language, updated and published every year, both in digital and paper version. The dictionary is traditionally dated one year ahead of the year it is published, hence the Zingarelli 2017

Table 1: Selected terms. **a-date** = first attested; **d-date** = decision date for extended meaning to be included in dictionary; **i-date** = actual inclusion date in Zingarelli for extended meaning.

term	literal	figurative	a-date	d-date	i-date
implosione	implosion	cedimento, tracollo improvviso (collapse)	1932	2013	2015
kamikaze	kamikaze	chi compie un'impresa rischiosa o destinata al fallimento (daredevil, reckless)	1944	2007	2009
rottamatore	dismantler	nel linguaggio giornalistico e della politica, chi si propone di allontanare e sostituire un gruppo dirigente considerato antiquato (new broom)	1990	2012	2014
talebano	Taliban	che (o chi) è dogmatico, integralista (hard-liner, extremist)	1995	2007	2009
tsunami	tsunami	evento che determina lo sconvolgimento di un assetto costituito (devastation, havoc)	1907	2008	2010

is published in June 2016, and it refers to decisions about new words and new meanings (including metaphorical ones) made up until December 2015.

We analysed the behaviour of a small set of terms extracted from the dictionary. We searched the 2017 edition to extract nouns that record a figurative meaning, limiting our search to words whose first occurrence is recorded in the 20th or 21st century. Newly born words (including borrowings) are more likely to show a meaning shift in the time span considered in our search (1984-2015) than older words (especially if derived directly form Latin, where the figurative meaning was also originally highly available, so probably arisen earlier). Out of a total of 447 hits, five target words were chosen for this pilot study. They are reported in Table 1 together with relevant information.

In order to minimise (at least in the context of this experiment) the influence of polysemy in the observable similarity distance across years, we verified that the selected terms are not polysemous, or minimally so. For the words "rottamatore", "talebano", and "tsunami", the Zingarelli records one sense only. For the word "implosione" three senses in total are recorded, two of which are however technical language, in the fields of linguistics (phonology) and psychology, and we assume will not be used much in newswire. For "kamikaze" the Zingarelli records one meaning only (Japanese pilot) to which is associated the extended sense of someone who kills himself in a terrorist attack; in our corpus the extended meaning is clearly the primary one, and the figurative sense that we consider is derived from it (see also Section 4.4).

4.2 Corpus

We created a diachronic corpus of approximately 60 millions tokens by collecting articles from the Italian newspaper *la Repubblica* from 1984 (the first year for which data is available digitally) to 2015. All texts were tokenised and lowercased. Because we are interested in how a term's context changes over time, we had to determine time-spans for our corpus, and we settled on two-year blocks, for a total of 16 time spans, the first one being 1984-1985 and the last 2014-2015. These subcorpora are used to train consecutive vector space models.

4.3 Model

We implemented vector representations using the skip-gram architecture introduced by Mikolov and Dean (2013). Such representations (Word Embeddings) are low dimensional, dense and real-valued vectors that have been proved to preserve syntactic and semantic information in several NLP tasks (Baroni et al., 2014).

Vectors created on different corpora cannot be directly compared, since every semantic space implements arbitrary orthogonal transformations and hence there is no direct correspondence between word vectors in different semantic spaces (Zhang et al., 2015). This would hold true also for our data, since we create a different corpus for each time span. Therefore, in order to create comparable vector representations for each word in any time span, we adopt the methodology introduced by Kim et al. (2014) (see Section 3), slightly modifying it. While Kim et al. (2014) use vectors of span y to initialise the vectors for year $y + 1$, we do the opposite, i.e.

we start with 2014-15, and use those vectors to initialise the 2012-13 time span, and thus backwards until 1984-85.

This methodological choice is due to the fact that the majority of the words in the set we considered for this experiment (included the selected target words, see 4.1) have few or no occurrences in the first time spans of the corpus: for example, "rottamatore" and "talebano" occur for the first time in 96/97. Indeed, using Kim et al. (2014)'s original approach, which we implemented in a preliminary experiment, the vectors for these words were correctly initialised, but were basically random vectors with no meaningful information. Conversely, our *reverse* setting, while still offering the same opportunity to trace shifts of meaning across time, allows to initialise all target words on a time span (14/15) in which they occur a number of times sufficient to create a more stable, meaningful representation.

Using the `gensim` library (Řehůřek and Sojka, 2010), we trained the models with the following parameters: window size of 5, learning rate of 0.01 and dimensionality of 200. We filtered out words with frequency lower than 5 occurrences. The vocabulary was initialised over the whole dataset.

4.4 Results and discussion

Figure 1 shows the similarity values for one time span to the next (dotted line), together with the average shift of meaning of a subset of 5000 nouns randomly selected (solid line). While we cannot draw any statistically significant conclusions from such little data, we aim at potentially observing patterns of shift of meaning through change of vector representations that could be used for developing predictive metrics of metaphorical shifts in time.

We interpret the results of our models according to (i) information in the Zingarelli dictionary and (ii) a manual inspection of the context of use of our target words in the corpus.

For (i), we verify if, for a given term, an observable correlation exists between changes in its vector representations and the insertion of a figurative sense in the dictionary. Results show that such a correlation exists for "talebano", "rottamatore", and "tsunami". For these words a drop in cosine similarity can be observed between three and five years before the insertion of the figurative meaning in the dictionary. This fits well with the timing for new meanings to be recorded in lexicographic resources (see Section 4.1). The nouns "kamikaze" and "implosione", instead, show a more stable evolution of meaning in time, with no clear drop in cosine similarity, and thus no evident correlation between changes in vector representations and insertion of a figurative meaning in dictionary.

For (ii), we manually inspected the contexts in which target terms occur in the the corpus as literal or metaphoric, in order to check if some relevant change in words usage could be observed in correspondence to drops in cosine similarity between time spans.

"Tsunami" occurs 27 times between 84/85 and 02/03: in 88.9% of the cases the word is used literally, with only 3 metaphorical uses in 98/99 (mirrored in a slight drop in cosine similarity). Of the 930 occurrences from 04/05 to 14/15, only 59.1% are literal. In Figure 1 we can observe a major drop in cosine similarity exactly between 04/05 and 06/06.

"Rottamatore" occurs 4 times between 84/85 and 08/09, always used literally. From 10/11 on, there are 156 occurrences, all metaphorical. Thus, the drop corresponds to change in usage here too.

"Talebano" occurs 12 times between 84/85 and 02/03, with 83.3% of literal usage. Once again, the drop in cosine coincides with the time span in which the term started to be used metaphorically: between 02/03 and 08/09 40% of the occurrences of "talebano" are metaphorical. Then, another relevant drop is observed between 08/09 and 10/11, and this is due to the sudden return of the literal usage of this word (86.1%), which continues also in the following years.

As already noticed, "kamikaze" and "implosione" do not seem to undergo a clear shift. As for the former, the analysis of its contexts of use reveals that indeed it is not possible to clearly identify, in our corpus, when exactly the term started to be used metaphorically: of the 25 occurrences of "kamikaze" in 84/85, 32% are metaphorical. This trend is fairly constant, and it explains why the vector representation of "kamikaze", which from the very beginning conflates literal and metaphorical usages, is stable in time. There is only a relevant change starting from 10/11: from this period onwards, the metaphorical use decreases, and almost all the occurrences are literal.[3] Accordingly, this

[3]Interestingly, this increase of literal usage is observed in the same years also for "talebano", a term that is semantically related to "kamikaze". This observation would require further investigation in connection with the socio-political events of those time spans.

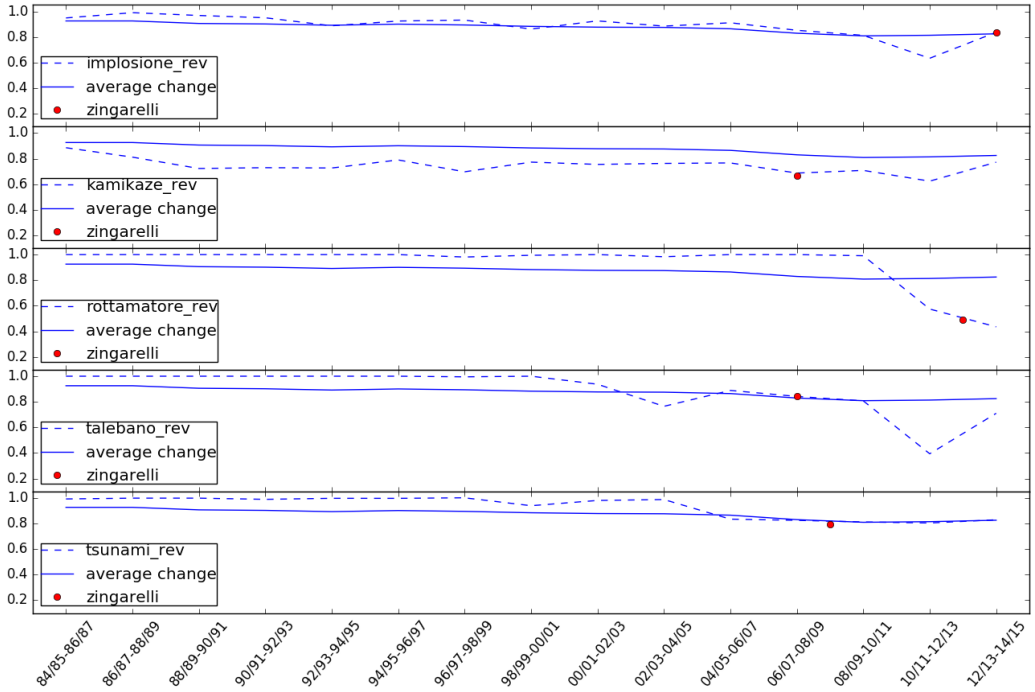


Figure 1: Cosine similarity values across time spans for target words (dotted line), average similarity of nouns (solid line) and date of insertion of metaphorical meaning in the Zingarelli dictionary (red dot).

almost exclusively return to the literal meaning corresponds to a slight increase in cosine similarity between the two last time spans.

"Implosione" occurs 433 times overall and in 92.4% of them is used metaphorically, but in few and specific contexts. A metaphorical, quite specific, sense of "implosione" is thus the main sense for this term in our corpus, and this is why we observe, on average, a high similarity across time spans. There is only a small drop between 10/11 and 12/13, when the word started to be used in the context of the economical crisis ("l'implosione dell'euro").

To sum up, both "kamikaze" and "implosione" show a similar stable behaviour in time, with only small drops. However, while for "kamikaze" such stability is due to a relatively constant ratio between literal and metaphorical meanings, in the case of "implosione" the observed stability is given by the constant predominance of the metaphorical sense across all the time spans.

5 Conclusion and future work

This work was meant as an exploration of the assumption that the emergence of the metaphorical use of a term might be mirrored in changes in cosine similarity of the term to itself across time. Such assumption has been partially confirmed by the comparison to the Zingarelli dictionary, while we found a more robust evidence when inspecting the terms' contexts of use manually.

Future work will stem from methodology and observations discussed here. Specifically, we plan to investigate further several aspects of this initial work, including the relation between changes in cosine similarity and frequency of use of a word: to which extent a change of the former relates to an increase of the latter? Mostly though, we plan to run experiments on larger sets of words with the aim to consolidate and then further exploit the mainly qualitative observations reported here towards the development of reliable predictive metrics which can serve to detect the emergence of shifts automatically, in a completely bottom-up fashion.

Acknowledgments

Malvina Nissim would like to thank the ILC-CNR ItaliaNLP Lab for their hospitality while working on this project. We are also grateful to the anonymous reviewers who provided insightful comments that doubtlessly contributed to improve this paper.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June. Association for Computational Linguistics.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71. Association for Computational Linguistics.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Ilana Heintz, Ryan Gabbard, Mahesh Srinivasan, David Barner, Donald S Black, Marjorie Freedman, and Ralph Weischedel. 2013. Automatic extraction of linguistic metaphor with LDA topic modeling. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 58–66.

Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huiying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 52–57.

Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 229–238. IEEE Press.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65. Association for Computational Linguistics.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. ACM.

T Mikolov and J Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2011. Tracing semantic change with latent semantic analysis. *Current methods in historical semantics*, pages 161–183.

Ekaterina Shutova and Lin Sun. 2013. Unsupervised metaphor identification using hierarchical graph factorization clustering. In *HLT-NAACL*, pages 978–988.

Ekaterina Shutova. 2015. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Peter D Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing*, pages 680–690.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, pages 35–40. ACM.

Y. Xu and C. Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Yating Zhang, Adam Jatowt, Sourav S Bhowmick, and Katsumi Tanaka. 2015. Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time. In *Proc. of ACL*, pages 645–655.

N. Zingarelli. 1993–2017. *Lo Zingarelli - Vocabolario della lingua italiana*. Zanichelli editore, Bologna.

KD Strikes Back: from Keyphrases to Labelled Domains Using External Knowledge Sources

Giovanni Moretti[1], Rachele Sprugnoli[1-2], Sara Tonelli[1]
[1]Fondazione Bruno Kessler, Trento
[2] Università di Trento
{moretti,sprugnoli,satonelli}@fbk.eu

Abstract

English. This paper presents L-KD, a tool that relies on available linguistic and knowledge resources to perform keyphrase clustering and labelling. The aim of L-KD is to help finding and tracing themes in English and Italian text data, represented by groups of keyphrases and associated domains. We perform an evaluation of the top-ranked domains using the 20 Newsgroup dataset, and we show that 8 domains out of 10 match with manually assigned labels. This confirms the good accuracy of this approach, which does not require supervision.

Italiano. *In questo lavoro descriviamo L-KD, un sistema che utilizza risorse linguistiche e basate su conoscenza per raggruppare concetti-chiave e categorizzarli. L'obiettivo di L-KD è quello di supportare gli utenti nel rilevare la presenza di specifici temi in documenti italiani e inglesi, rappresentandoli attraverso gruppi di concetti-chiave e relativi domini. Abbiamo valutato l'affidabilità del sistema analizzando i domini più rilevanti nel 20 Newsgroup dataset, e dimostrando che 8 su 10 domini nel gold standard sono assegnati correttamente anche dal sistema. Questa valutazione conferma le buone performance di L-KD, senza il bisogno di supervisione.*

1 Introduction

With the increasing availability of large document collections in digital format, companies, organizations but also non-expert users face everyday the need to efficiently extract and categorize relevant information from large corpora. The possibility to extract key-concepts and assign them to a domain without the need of supervision would allow them to systematically track the flow of information and retain only relevant content at two granularity levels: key-concepts, and domains to which these key-concepts can be ascribed. Although topic models (Blei et al., 2003) can be used to this purpose, they have two main drawbacks: the number of topics for a corpus is arbitrary and topics are often not labelled.

In this work, we present a solution to the aforementioned research problem by presenting L-KD (Labelled-KD), a tool to perform keyphrase clustering and labelling through the exploitation of external linguistic and knowledge resources. The tool takes advantage of the availability of Keyphrase Digger[1] (KD), a multilingual rule-based system that detects a weighted list of n-grams representing the most important concepts in a text (Moretti et al., 2015). These key-concepts are then linked to WordNet Domains (Magnini and Cavaglia, 2000) in order to create clusters of key-concepts labelled by domain. The problem of ambiguous concepts, i.e. possibly belonging to more than one WordNet domain, is tackled by using ConceptNet 5 (Speer and Havasi, 2013), a multilingual knowledge source containing single and multi-word concepts linked to each other by a broad set of relations covering different types of associations. The outcome of this study is the L-KD tool, supporting both English and Italian, which we make available to the research community[2]. L-KD takes in input a document in plain text format, and outputs the ranked list of semantic domains discussed in the documents, each associated with a set of keyphrases.

[1]http://dh.fbk.eu/technologies/kd
[2]https://dh.fbk.eu/technologies/l-kd

2 Related Works

In the last years, a number of works dealing with the unsupervised clustering of keyphrases has been presented (Hasan and Ng, 2014). Liu et al. (2009) use Wikipedia and co-occurrence-based statistics to semantically cluster similar keyphrases in a set of unweighted topics. In order to improve this approach by weighting topics, Liu et al. (2010) and Grineva et al. (2009) propose a topic-decomposed PageRank and a network analysis algorithm respectively to perform hierarchical clustering. Our method is simpler than the previously mentioned studies, and relies on available resources to label the clusters. Indeed, the lists of terms listed in the topics are not always easy to interpret (Aletras et al., 2015), and adding a label that captures the meaning of each cluster is a way to enhance its understanding. The problem of interpretation affects also the output of topic modelling algorithms, i.e. unsupervised statistical methods such as Latent Dirichlet Allocation (Blei et al., 2003). Many techniques have been developed to automatically label topics for example by using probabilistic approaches (Mei et al., 2007), Wikipedia links (Xu and Oard, 2011) and DBpedia structured data (Hulpus et al., 2013). As for the automatic labelling of keyphrase clusters, Carmel et al. (2009) adopt Wikipedia as an external resource to extract candidate labels. To the best of our knowledge, no available system performs this task by combining WordNet Domains and ConceptNet 5.

3 System Overview

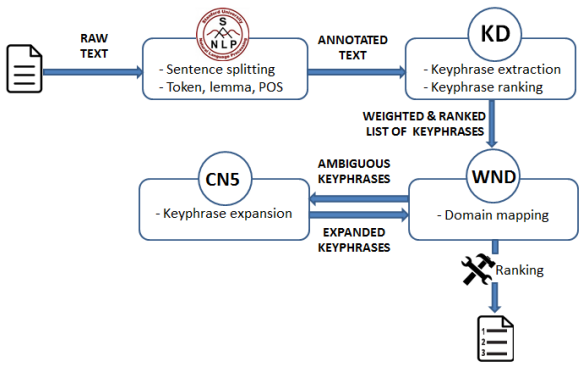


Figure 1: General workflow underlying L-KD for English documents with the steps involving the use of Stanford CoreNLP, KD, WordNet Domains (WND) and ConceptNet 5 (CN5).

```
nature -> [country_side, environment,
flower_lake, mother_goddess, science,
outdoor_world, tree_animal, many_wonder,
great_place, fauna, flora, ecology,
conservation, country, ecosystem,
everything_live, flower, plant …….]
```

```
{Biology=19, Geography=13, Plants=8,
Animals=5 …}
```

Figure 2: Excerpt of the expansion of an ambiguous keyphrase using ConceptNet 5 (top) and top domains assigned to this expansion (bottom).

L-KD performs several steps (see Fig. 1) to semantically cluster keyphrases and label each cluster:

1) Text preprocessing: Stanford CoreNLP (Manning et al., 2014) is used to split sentences, tokenize, lemmatize and tag the part-of-speech of the input English text. For Italian texts, we rely on Tint[3], a suite of NLP tools (Aprosio and Moretti, 2016) based on the Stanford CoreNLP pipeline.

2) Keyphrase extraction and ranking: L-KD integrates KD, a keyphrase extraction tool that combines statistical and linguistic knowledge, given by recurrent relevant PoS patterns, to extract single words and multi-token expressions encoding the main concepts of a document. A detailed description of KD functionalities is given in Moretti et al. (2015). The output of this step is a weighted and ranked list of keyphrases.

3) Domain mapping: L-KD maps the lemma forms of keyphrases with the lemmas in WND aligned to WordNet 3.0[4]. For Italian we rely on the data available through the Open Multilingual WordNet project (Bond and Paik, 2012) as a bridge between lemmas and WND. In case of multi-token expressions (e.g. "federal government"), the system looks for a perfect match. If no match is found, the tokens are splitted and only the nouns are searched in WND (e.g. "government"). A list of domain-keyphrases associations is created, as well as a list of ambiguous keyphrases. The latter comprises those that are assigned to the `Factotum` domain and those that could belong to several domains, if none of them contains > 3 keyphrases. This threshold was manually set in order to identify domains that are likely to be little relevant.

4) Expansion of ambiguous keyphrases: The lemmas of ambiguous keyphrases are aligned with

[3] `http://tint.fbk.eu/`

[4] Courtesy of Carlo Strapparava.

the lemmas in ConceptNet 5 and are expanded by retrieving all the connected concepts following ConcepNet 5 relations. L-KD relies on a subset of relations including hierarchical (*HasA*, *PartOf*, *MadeOf*, *IsA*, *DerivedFrom*) and synonymous (*Synonym*, *RelatedTo*) ones (Mukherjee and Joshi, 2013). Functional relations such as *CapableOf* and *UsedFor* are not taken into consideration because the concepts evoked by these relations may be too far from the original meaning of the key-concept. The upper part of Fig. 2 shows how "nature", an ambiguous keyphrase, is expanded following this procedure. Examples of the relations that lead to this expansion are the following:
- nature ⇒ RelatedTo ⇒ flora
- nature ⇒ IsA ⇒ great_place
- nature ⇒ HasA ⇒ many_wonder

5) Domain mapping of expanded keyphrases: All the lemmas included in the expansion created in the previous step are mapped to domains using WND. The lower part of Fig. 2 reports the top domains related to the expansion of "nature" together with the number of lemmas associated with them, e.g. 19 lemmas are mapped to the `Biology` domain. A relevance score (i.e. number of keyphrases associated with a domain) is computed for the domains retrieved for each expanded keyphrase. Domains are then compared with the ones found in Step (3) starting from the domain with the highest score. If it is already present in the domain-keyphrases list compiled in Step (3), then the keyphrase is associated with this domain, otherwise the other domains are checked. If the domain is not present in the list, it is added to the list with its associated keyphrase. The final relevance score of the domains is recalculated at the end of this step. Four sub-domains of `Factotum`, i.e. `Time_Period`, `Person`, `Metrology` and `Numbers`, which are very generic, usually have a high relevance because they tend to include many keywords. Therefore, we introduce a final reweighting step to deboost them.

6) Final ranking. L-KD creates a final ranked list of domains associated with clusters of keyphrases. The ranking is based on the relevance score of the domains as described in the previous step and on the rank of keyphrases as given by KD in step (2).

4 Evaluation

We evaluated L-KD using the 20 Newsgroup dataset (Joachims, 1996), a corpus of 20,000 documents extracted from UseNet discussion groups. This dataset is freely available online[5] and has been often employed to train and test text categorization algorithms (Moschitti and Basili, 2004). Specifically, each of its documents was manually assigned to one out of twenty different categories, which can be easily mapped to WND labels. Although L-KD can assign a ranked list of domains to one or more documents, thus providing a richer representation of the document(s) content, we did not find a suitable gold standard to evaluate the rank. Therefore, we limit our evaluation to the top-ranked domain extracted by the tool. We also decided to group Newsgroup categories that are strictly related to each other: e.g. documents in *talk.religion.misc*, *alt.atheism*, and *soc.religion.christian* all discuss religious issues and for this reason their texts are collapsed in a single category.
Table 1 reports the results of L-KD on the documents included in each category or group of categories. The second column shows the top two domains retrieved by the system and the third column presents some of the extracted keyphrases. Only in 2 cases out of 10, the first ranked domain does not perfectly match the original category: indeed `Law` is the top domain of *sci.eletronics* and of the documents related to political themes (*talk.politics.misc*, *talk.politics.guns*, *talk.politics.mideast*). We can notice that `Law` is a very frequent domain because it contains generic and recurring words such as "article", "opinion" and "information". In the rest of the cases (8 out of 10), the match between the first ranked domain and the original category is perfect: for example, the domain with the highest rank for documents discussing computer technologies is `Computer_Science`. In many cases also the second domain is extremely relevant. For instance, *misc.forsale* contains messages of people searching or selling goods with a focus on computer devices and components: the first retrieved domain is `Commerce` and the second one is `Computer_Science`. Each domain is associated with pertinent keyphrases such as "best offer" for the first domain and "floppy drive" for the second.

[5] `http://qwone.com/~jason/20Newsgroups/`

ORIGINAL CATEGORIES	TOP DOMAINS	KEYPHRASES
sci.med	**Medicine**	doctor, infectious disease, side effect
	School	course, science, study
sci.space	**Astronomy**	solar system, physical universe, satellite
	Transport	spacecraft, shuttle, high-speed collision
sci.crypt	**Computer_Science**	internet, e-mail, bit
	Law	security, second amendment, criminal
sci.electronics	Law	article, opinion, information
	Electricity	amateur radio, voltage, wire
talk.religion.misc - alt.atheism - soc.religion.christian	**Religion**	christian, atheist, objective morality
	Law	law, evidence, private activities
rec.sport.baseball - rec.sport.hockey	**Sport**	game, playoff, second period
	Play	player, baseball
rec.autos - rec.motorcycles	**Transport**	car, mph, front wheel
	Law	article, opinion
comp.graphics - comp.os.mswindows.misc - comp.sys.ibm.pc.hardware - comp.windows.x - comp.sys.mac.hardware	**Computer_Science**	software, hard drive, anonymous ftp
	Publishing	article, opinion
talk.politics.misc - talk.politics.guns - talk.politics.mideast	Law	opinion, second amendment
	Transport	road, ways of escape
misc.forsale	**Commerce**	best offer, price, excellent condition
	Computer_Science	hard drive, floppy drive, email

Table 1: Results of L-KD on the 20 Newsgroup dataset. The original categories are compared with the top domains extracted by the systems. Examples of keyphrases are provided for each domain. Perfect matches between the main theme of the original classification and L-KD top domains are in bold.

5 Use Case: the De Gasperi Project

L-KD has been recently applied to the analysis of the complete corpus of public writings by Alcide De Gasperi (De Gasperi, 2006) in the context of a research project, whose goal is to give insight into De Gasperi's communication strategy with the help of innovative tools for text analysis. We processed the 2,762 documents (around 3,000,000 tokens) in the corpus, published between 1901 and 1954, to analyse which domains appeared in the collection and how they changed over time. The advantage of L-KD is that it can provide both a distant view, by computing aggregated information on the domains, and a close reading of the documents, showing which key-concepts are mapped to which domain. As an example, we report in Fig. 3 the analysis related to two documents, entitled "Rene de la Tour du Pin" and "I cattolici nell'evoluzione sociale'. For each of them, the dendogram shows the three top domains and the associated key-concepts. The proposed analysis was validated at different granularities by two history scholars, who confirmed the consistency of L-KD analysis and found correspondences between the top domains and relevant events in De Gasperi's life.

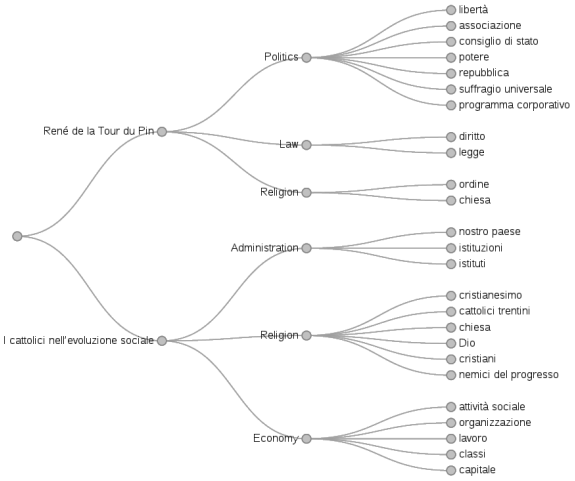


Figure 3: Dendogram related to two documents from De Gasperi's corpus

6 Tool Availability

L-KD is available as a web application[6] through which users can copy&paste a document and run the tool processing it on the fly. This application makes L-KD easily accessible also by users without a technical background.

[6] http://dhlab.fbk.eu:8080/L_KD/

In the application some parameters are given, while others can be changed by the user according to his/her needs. As for the fixed parameters, proper names are always discarded so to exclude them from the list of keyphrases: this setting is justified by the fact that WordNet, and consequently WND, contains few proper nouns[7] while we want to maximize the mapping. For the same reason, short keyphrases, i.e. single words and multi-token expressions with a maximum length of 4 words[8], are preferred. On the contrary, the minimum number of occurrences for a word or expression to be considered as a candidate keyphrase and the number of keyphrases to be extracted can be customized by the user. For example, in case of short documents, a low number of keyphrases (e.g. up to 20) can be set together with a minimum frequency of 1 or 2 (in a short text repetitions are less likely to occur). For long documents more keyphrases can be extracted: in this way it would be easy to find clusters covering multiple themes.

7 Conclusions and Future Works

This paper presents L-KD, a tool that extracts keyphrases from text data, clusters them according to the domain and assigns a label to each cluster. The process underlying L-KD is based on the exploitation of external linguistic and knowledge resources, i.e. WordNet Domains and ConceptNet 5. Our tool can process both English and Italian texts of different length and content, from a single news article to an entire book, from single-theme to multi-theme documents.

In the future we will explore different research directions. First of all we want to evaluate the tool on Italian data, even if we have not found a suitable gold standard so far. Resorting to crowd-sourcing may be a viable solution. We expect lower performances than the ones obtained for English, given that the current mapping between Open Multilingual WordNet and WordNet 3.0 covers only the 32.5% of the English synsets: this consequently affects the mapping on the domains of WND. Moreover, the coverage of Italian in ConcepNet 5 is limited. As for the availability of L-KD, we plan to release the tool as a standalone module. It will also be integrated in the ALCIDE platform (Moretti et al., 2016) that supports the analysis of large document collections for humanities studies.

[7] Only the 9.4% of synsets are tagged as being instances, i.e. proper nouns, in WordNet 3.0 (Abrate et al., 2012).

[8] In WordNet 3.0 only the 0.2% of noun synsets have a length greater than 4 words.

Acknowledgments

The research leading to this paper was partially supported by the EU Horizon 2020 Programme via the SIMPATICO Project (H2020-EURO-6-2015, n. 692819). We thanks Alessio Palmero Aprosio for his help in the evaluation process.

References

Matteo Abrate, Clara Bacciu, Andrea Marchetti, and Maurizio Tesconi. 2012. WordNet atlas: a web application for visualizing WordNet as a zoomable map. In *GWC 2012 6th International Global Wordnet Conference*, page 23.

Nikolaos Aletras, Timothy Baldwin, Jey Han Lau, and Mark Stevenson. 2015. Evaluating topic representations for exploring document collections. *Journal of the Association for Information Science and Technology*.

Alessio Palmero Aprosio and Giovanni Moretti. 2016. Italy goes to Stanford: a collection of CoreNLP modules for Italian. *arXiv preprint arXiv:1609.06204*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Francis Bond and Kyonghee Paik. 2012. A survey of WordNets and their licenses. *Small*, 8(4):5.

David Carmel, Haggai Roitman, and Naama Zwerdling. 2009. Enhancing cluster labeling using wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 139–146. ACM.

A. De Gasperi. 2006. Scritti e discorsi politici. In E. Tonezzer, M. Bigaran, and M. Guiotto, editors, *Scritti e discorsi politici*, volume 1. Il Mulino.

Maria Grineva, Maxim Grinev, and Dmitry Lizorkin. 2009. Extracting key terms from noisy and multitheme documents. In *Proceedings of the 18th international conference on World wide web*, pages 661–670. ACM.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1262–1273. Association for Computational Linguistics.

Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. 2013. Unsupervised graph-based topic labelling using dbpedia. In *Proceedings of the*

sixth ACM international conference on Web search and data mining, pages 465–474. ACM.

Thorsten Joachims. 1996. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. Technical report, DTIC Document.

Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun. 2009. Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 257–266. Association for Computational Linguistics.

Zhiyuan Liu, Wenyi Huang, Yabin Zheng, and Maosong Sun. 2010. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 366–376. Association for Computational Linguistics.

Bernardo Magnini and Gabriela Cavaglia. 2000. Integrating Subject Field Codes into WordNet. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000)*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. 2007. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499. ACM.

Giovanni Moretti, Rachele Sprugnoli, and Sara Tonelli. 2015. Digging in the dirt: Extracting keyphrases from texts with kd. In *Proceedings of CLiC-it 2016*, page 198.

Giovanni Moretti, Rachele Sprugnoli, Stefano Menini, and Sara Tonelli. 2016. ALCIDE: Extracting and visualising content from large document collections to support Humanities studies. *Knowledge-Based Systems*, 111:100–112.

Alessandro Moschitti and Roberto Basili. 2004. Complex linguistic features for text classification: A comprehensive study. In *European Conference on Information Retrieval*, pages 181–196. Springer.

Subhabrata Mukherjee and Sachindra Joshi. 2013. Sentiment Aggregation using ConceptNet Ontology. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCNLP)*, pages 570–578.

Robert Speer and Catherine Havasi. 2013. ConceptNet 5: A large semantic network for relational knowledge. In *The Peoples Web Meets NLP*, pages 161–176. Springer.

Tan Xu and Douglas W Oard. 2011. Wikipedia-based topic clustering for microblogs. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10.

Investigating the Application of Distributional Semantics to Stylometry

Giulia Benotto, Emiliano Giovannetti, Simone Marchi
Istituto di Linguistica Computazionale "A. Zampolli"
Consiglio Nazionale delle Ricerche
Via G. Moruzzi 1, 56124, Pisa - Italy
{name.surname}@ilc.cnr.it

Abstract

English. The inclusion of semantic features in the stylometric analysis of literary texts appears to be poorly investigated. In this work, we experiment with the application of Distributional Semantics to a corpus of Italian literature to test if words distribution can convey stylistic cues. To verify our hypothesis, we have set up an Authorship Attribution experiment. Indeed, the results we have obtained suggest that the style of an author can reveal itself through words distribution too.

Italiano. *L'inclusione di caratteristiche semantiche nell'analisi stilometrica di testi letterari appare poco studiata. In questo lavoro, sperimentiamo l'applicazione della Semantica Distribuzionale ad un corpus di letteratura italiana per verificare se la distribuzione delle parole possa fornire indizi stilistici. Per verificare la nostra ipotesi, abbiamo imbastito un esperimento di Authorship Attribution. I risultati ottenuti suggeriscono che, effettivamente, lo stile di un autore pu rivelarsi anche attraverso la distribuzione delle parole.*

1 Introduction

Stylometry, that is the application of the study of linguistic style, offers a means of capturing the elusive character of an author's style by quantifying some of its features. The basic stylometric assumption is that each writer has certain stylistic idiosyncrasies (a "human stylome" (Van Halteren et al., 2005)) that define their style. Analysis based on stylometry are often used for Authorship Attribution (AA) tasks, since the main idea behind computationally supported AA is that by measuring some textual features, we can distinguish between texts written by different authors (Stamatatos, 2009).

One of the less investigated stylistic feature is the way in which authors use words from a semantic point of view, e.g. if they tend to use more, when dealing with polysemous words, a certain sense over the others, or senses that differ (even slightly) from the one that's more commonly used (as it happens, typically, in poetry).

A possible approach to the analysis of this characteristic is to consider the textual contexts in which certain words appear. According to Distributional Semantics (DS), certain aspects of the meaning of lexical expressions depend on the distributional properties of such expressions, or better, on the contexts in which they are observed (Lenci, 2008; Miller and Charles, 1991). The semantic properties of a word can then be defined by inspecting a significant number of linguistic contexts, representative of the distributional behavior of such word.

In this work we would like to investigate if the analysis of the distribution of words in a text can be exploited to provide a stylistic cue. In order to inspect that, we have experimented with the application of DS to the stylometric analysis of literary texts belonging to a corpus constituted by texts pertaining to the work of six Italian writers of the late nineteenth century.

In the following, Section 2 gives a short insight on the state of the art of computational stylistic analysis, Section 3 describes the approach together with the corpus used to conduct our investigation and Section 4 discuss about results. Finally, Section 5 draws some conclusions and outlines some possible future works.

2 State of the Art

The very first attempts to analyze the style of an author were based on simple lexical features such

as sentence length counts and word length counts, since they can be applied to any language and any corpus with no additional requirements (Koppel and Schler, 2004; Stamatatos, 2006; Zhao and Zobel, 2005; Argamon et al., 2007). Similarly, character measures have been proven to be quite useful to quantify the writing style (Grieve, 2007; De Vel et al., 2001; Zheng et al., 2006). Basically, a text can be viewed as a mere sequence of characters, so that various measures can be defined (including alphabetic, digit, uppercase and lowercase characters count, etc.). A more elaborate text representation method is to employ syntactic information (Gamon, 2004; Stamatatos et al., 2000; Stamatatos et al., 2001; Hirst and Feiguina, 2007; Uzuner and Katz, 2005). The idea is that authors tend to use similar syntactic patterns unconsciously. Therefore, syntactic information is considered a more reliable authorial fingerprint in comparison to lexical information.

More complicated tasks such as full syntactic parsing, semantic analysis, or pragmatic analysis cannot yet be handled adequately by current NLP technologies for unrestricted text. As a result, very few attempts have been made to exploit high-level features for stylometric purposes. Perhaps the most important method of exploiting semantic information so far was described in (Argamon et al., 2007). This work was based on the theory of Systemic Functional Grammar (SFG) (Halliday, 1994) and consisted on the definition of a set of functional features that associate certain words or phrases with semantic information.

The previously described features are application independent since they can be extracted from any textual data. Beyond that, one can define application-specific measures in order to better represent the nuances of style in a given text domain (such as e-mail messages, or online forum messages) (Li et al., 2006; Teng et al., 2004).

To the best of our knowledge, the application of DS to the analysis of literary texts has been documented in a rather small number of works (Buitelaar et al., 2014; Herbelot, 2015). In both these works, DS is used as a theoretical basis in order to verify some hypotheses on specific semantic characteristics of poetic works. In more details, in (Buitelaar et al., 2014) the authors investigated through DS the influence of Lord Byron's work on Thomas Moore trying to find a shared vocabulary or specific formal textual characteristics. In (Herbelot, 2015) it is argued how distributionalism can support the notion that the meaning of poetry comes from the meaning of ordinary language and how distributional representations can model the link between ordinary and poetic language. However, the role of DS in the study of a style of an author was not the aim of these works.

3 Experimental Setup

First, we want to specify that it is not our purpose to propose new ways to improve state-of-the-art AA algorithms. Indeed, our aim is just to verify the hypothesis that the distribution of words can provide an indication of a distributional stylistic fingerprint of an author. To do this, we have set up a simple classification task. Subsection 3.1 briefly depicts the data set we used, while Section 3.2 describes the steps implemented in our experiment.

3.1 Data Set Construction

In order to build the reference and test corpora, we started from texts pertaining to the work of six Italian writers working at the turn of the 20th century, namely, Luigi Capuana, Federico De Roberto, Luigi Pirandello, Italo Svevo, Federigo Tozzi and Giovanni Verga. We chose contiguous authors in chronological sense, whose texts are available in digital format (in fact we could not do a similar survey on the narrative of the 90s because it is still under copyrights). Indeed, we used texts freely available for download from the digital library of the Manunzio project, via the LiberLiber website[1]. Since they were encoded in various formats, such as .epub, .odt and .txt, our pre-processing consisted in converting them all in .txt format and getting rid of all xml tags, together with footnotes and editors' notes and comments.

3.2 Experiment Description

According to Rudman (1997), a striking problem in stylometry is due to the lack of homogeneity of the examined corpora, in particular to the improper selection or fragmentation of the texts, that might cause alterations in the writers' style. In order to create balanced reference corpora, i.e. covering all the authors' different stylistic and thematic phases, for each author, as shown in Figure 1, we built a reference corpus as the composition of the 70% of each single work (usually a novel). The same technique was used to create the

[1] http://www.liberliber.it/

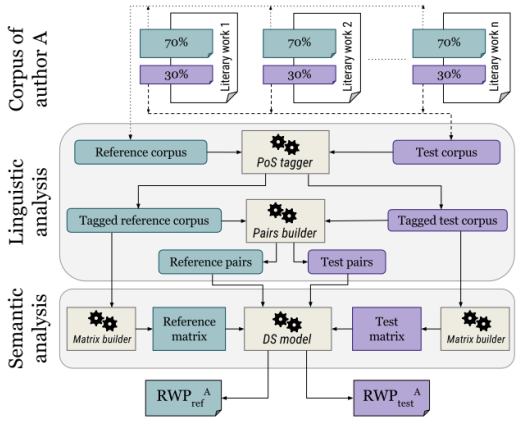


Figure 1: RWP$_{ref}$ and RWP$_{test}$ creation process for an author.

test corpus by using the remaining 30% of each work. Typical AA approaches consist in analyzing known authors and assigning authorship to previously unseen text on the basis of various features. Train and test sets should then contain different texts. Contrary to the classical AA task, our train and test sets contain different parts of the same texts. Indeed, with this experiment, we wanted to understand if the semantics that an author bestows to a word, is peculiar to his writing. To prove this, we wanted to cover all the different stylistic and thematic phases an author can go through during his activity, hence the partition of all his texts in a reference and a test portion.

We then analyzed each reference and test corpora with a Part-of-Speech (PoS) tagger and a lemmatizer for Italian (Dell'Orletta et al., 2014). For every author, we built two lists of word pairs (with their lemma and PoS), one relative to the tagged reference corpus (*reference pairs*) and the other to the tagged test set (*test pairs*), where each word was paired with all the other words with the same PoS. We also filtered the pairs to leave only nouns, adjectives and verbs. Starting from the tagged corpora, we built two words-by-words matrixes[2] of co-occurrence counts (co-occurrence matrixes) for each author, using a context window of 4[3]. The chosen DS model (Baroni and Lenci, 2010) was applied to each matrix to calculate the cosine between the vectors representing the two words of each pair. This allowed us to evaluate the semantic relatedness between the words by assessing their proximity in the distributional space as represented by the cosine value: the more this value tends to 1, the more the two words of the pair are considered to be related. We then obtained two related word pair (RWP) lists for each author *A*: RWP$_{\text{ref}}^{\text{A}}$ and RWP$_{\text{test}}^{\text{A}}$. Figure 1 shows the process described above.

Since we wanted to focus on the analysis of the semantic distribution of words, we decided to exclude any possible "lexical bias". For this reason, we restricted the analysis on a common vocabulary, i.e. a vocabulary constituted by the intersection of the six authors' vocabularies. In this way, we prevent our classifier to exploit, as a feature, the presence of words used by some (but not all) of the authors. Moreover, we removed from the RWP$_{\text{test}}$ lists all those pairs of words occurring frequently together in the same context, since they might constitute a multiword expression that, once again, could be pertaining with the signature lexicon of each author. To remove them, we computed the number of times (*#co-occ* in Table 1) they appeared together in the context window, as well as their total number of occurrences (*#occ$_a$* and *#occ$_b$*) and we excluded from the analysis those pairs for which the ratio between the number of co-occurrences and the total occurrences of the less frequent word was higher than the empirically set threshold of 0.5. The first two pairs of Table 1 would be removed as probable multiword (PM column in Table 1): "*scoppio*" (burst) and "*risa*" (laughter) could mostly co-occur in "*scoppio di risa*" (meaning "burst of laughter") and the words "*man*" and "*mano*" (both meaning "hand") could mostly co-occur in "*man mano*" (meaning "little by little", or "progressively").

W$_a$	W$_b$	#occ$_a$	#occ$_b$	#co-occ	ratio	PM
scoppio–s	risa–s	19	9	7	0.78	yes
man–n	mano–n	50	1325	47	0.94	yes
nausea–n	disgusto–n	27	26	0	0	no
piccolo–a	grande–a	248	237	14	0.06	no

Table 1: An example of co-occurring RWs from Pirandello's test list: the first two pairs would be removed.

Finally, we reduced the size of the six RWP$_{\text{ref}}$ and RWP$_{\text{test}}$ lists by sorting them in decreasing order of the cosine value and then by keeping the

[2]Being the corpus relatively small and not having particular computability issues, we chose not to apply decomposition techniques to reduce the size of the matrixes (and thus not losing any information).

[3]We performed different empiric setup of the window's size and chose the one that showed more suitable results, according to what is stated by Kruszewski and Baroni (2014).

pairs with the highest cosine, selected using a percentage parameter θ as threshold[4]. We chose to introduce the parameter θ for two reasons: i) to avoid the classification algorithm to be disturbed by noisy (i.e. not significative) pairs which would not hold any relevant stylistic cue, and ii) to ease a literary scholar in the interpretation of the results by having to analyze just a limited selection of (potentially) semantically related word pairs.

For the last phase of our experiment we defined a classification algorithm to test the effective presence of stylistic cues inside the obtained RWP_{test} lists. We defined a classifier using a nearest-cosine method to attribute each test list to an author. The method consisted in searching for a pair of words contained in the test list inside each reference list and incrementing by 1 the score of the author whose reference list included the pair with the more similar cosine value (i.e. having the minimum difference): the chosen author was the one with the highest score. Table 2 shows the classification results for θ = 5%.

	Capuana	De Roberto	Pirandello	Svevo	Tozzi	Verga
Capuana	1884	1269	1321	797	755	1054
De Roberto	729	1041	712	498	451	579
Pirandello	1387	1278	2114	937	747	1056
Svevo	353	371	341	593	372	356
Tozzi	199	219	183	242	281	244
Verga	650	671	656	473	430	851

Table 2: Classification results, obtained via the nearest-cosine method for θ = 5%.

4 Interpreting the Results

As summarized in Table 3, a correct classification of all RWPs in RWP_{test} lists has been obtained with a θ value of 5%.

To help in interpreting the failure of the algorithm in classifying Tozzi's test list for θ values lower than 5% (as shown in Table 3) we calculated the cardinality of the RWP_{test} lists for each author with the change in θ value (Tables 4).

It is possible to observe how the choice of θ influences the correct classification of Tozzi's test list. Indeed, the use of a θ value below 5% has the effect of remarkably reducing an already small test list ($\text{RWP}_{\text{text}}{}^{\text{Tozzi}}$) as shown in Table 4. It is apparent that increasing the value of θ and consequently the number of significant RW pairs that are analysed, the system is able to correctly classify $\text{RWP}_{\text{test}}{}^{\text{Tozzi}}$ (see the values in Tozzi's row of Table 3).

[4]At the following url we have uploaded an archive containing all the data we have used and processed for our experiment: https://goo.gl/nrTqWh

	0.5%	1%	2%	5%
Capuana	Capuana	Capuana	Capuana	Capuana
De Roberto	De Roberto	De Roberto	De Roberto	De Roberto
Pirandello	Pirandello	Pirandello	Pirandello	Pirandello
Svevo	Svevo	Svevo	Svevo	Svevo
Tozzi	Verga	Verga	Tozzi/Verga	Tozzi
Verga	Verga	Verga	Verga	Verga

Table 3: Results of the classification. Classification errors are highlighted.

	0.5%	1%	2%	5%
#$\text{RWP}_{\text{test}}{}^{\text{Capuana}}$	678	1357	2714	6785
#$\text{RWP}_{\text{test}}{}^{\text{De Roberto}}$	488	977	1954	4886
#$\text{RWP}_{\text{test}}{}^{\text{Pirandello}}$	692	1385	2770	6925
#$\text{RWP}_{\text{test}}{}^{\text{Svevo}}$	425	851	1702	4257
#$\text{RWP}_{\text{test}}{}^{\text{Tozzi}}$	246	493	986	2466
#$\text{RWP}_{\text{test}}{}^{\text{Verga}}$	526	1053	2106	5267

Table 4: Cardinality of RWP_{test} for each author and for each θ value.

5 Conclusion and Next Steps

In this paper we investigated the possibility that an analysis of the semantic distribution of words in a text can be potentially exploited to get cues about the style of an author. In order to validate our hypothesis, we conducted a first experiment on six different Italian authors. The results seem to suggest that the way words are distributed across a text, can provide a valid stylistic cue to distinguish an author's work. Of course, it is not our intent, with this paper, to define new methods for enhancing state-of-the-art authorship attribution algorithms. Our research will focus, in the next steps, in detecting and providing useful indications about the style of an author. This can be done by highlighting, for example, atypical distributions of words (e.g. with contrastive methods) or by analysing their distributional variability. Furthermore, it could be interesting to use a different distributional measure, than the cosine, to test our hypothesis.

References

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822, April.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Paul Buitelaar, Nitish Aggarwal, and Justin Tonra. 2014. Using distributional semantics to trace influence and imitation in romantic orientalist poetry. In *AHA!-Workshop 2014 on Information Discovery in Text*. ACL.

Olivier De Vel, Alison Anderson, Malcolm Corney, and George Mohay. 2001. Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64.

Felice Dell'Orletta, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. 2014. T2kˆ 2: a system for automatically extracting and organizing knowledge from texts. In *LREC*, pages 2062–2070.

Michael Gamon. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proceedings of the 20th international conference on Computational Linguistics*, page 611. Association for Computational Linguistics.

Jack Grieve. 2007. Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, 22(3):251–270, May.

Michael AK Halliday. 1994. Functional grammar. *London: Edward Arnold*.

Aurélie Herbelot. 2015. The semantics of poetry: A distributional reading. *Digital Scholarship in the Humanities*, 30(4):516–531.

Graeme Hirst and Olga Feiguina. 2007. Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts. *Literary and Linguistic Computing*, 22(4):405–417, September.

Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*, page 62. ACM.

Germán Kruszewski and Marco Baroni. 2014. Dead parrots make bad pets: Exploring modifier effects in noun phrases. *Lexical and Computational Semantics (* SEM 2014)*, page 171.

Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

Jiexun Li, Rong Zheng, and Hsinchun Chen. 2006. From fingerprint to writeprint. *Communications of the ACM*, 49(4):76–82.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Joseph Rudman. 1997. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31(4):351–365.

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2000. Automatic text categorization in terms of genre and author. *Computational linguistics*, 26(4):471–495.

Efstathios Stamatatos, Nikos Fakotakis, and Georgios Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214.

Efstathios Stamatatos. 2006. Authorship attribution based on feature set subspacing ensembles. *International Journal on Artificial Intelligence Tools*, 15(05):823–838.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.*, 60(3):538–556, March.

Gui-Fa Teng, Mao-Sheng Lai, Jian-Bin Ma, and Ying Li. 2004. E-mail authorship mining based on svm for computer forensic. In *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, volume 2, pages 1204–1207. IEEE.

Özlem Uzuner and Boris Katz. 2005. A comparative study of language models for book and author recognition. In *Natural Language Processing–IJCNLP 2005*, pages 969–980. Springer.

Hans Van Halteren, Harald Baayen, Fiona Tweedie, Marco Haverkort, and Anneke Neijt. 2005. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1):65–77.

Ying Zhao and Justin Zobel. 2005. Effective and scalable authorship attribution using function words. In *Information Retrieval Technology*, pages 174–189. Springer.

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393, February.

Diachronic Analysis of the Italian Language exploiting Google Ngram

Pierpaolo Basile[1] and **Annalina Caputo**[1] and **Roberta Luisi**[2] and **Giovanni Semeraro**[1]
Department of Computer Science
University of Bari Aldo Moro
Via, E. Orabona, 4 - 70125 Bari (Italy)
[1]{firstname.surname}@uniba.it
[2]{roby.luisi}@gmail.com

Abstract

English. In this paper, we propose several methods for the diachronic analysis of the Italian language. We build several models by exploiting Temporal Random Indexing and the Google Ngram dataset for the Italian language. Each proposed method is evaluated on the ability to automatically identify meaning shift over time. To this end, we introduce a new dataset built by looking at the etymological information reported in some dictionaries.

Italiano. *In questo lavoro proponiamo alcuni metodi per l'analisi diacronica della lingua italiana. Abbiamo costruito differenti modelli utilizzando la tecnica del Temporal Random Indexing e Google Ngram per l'italiano. Ciascun metodo proposto è stato valutato rispetto alla capacità di identificare automaticamente i cambi di significato nel tempo. A tale scopo introduciamo uno nuovo dataset costruito mediante le informazioni etimologiche presenti in alcuni dizionari.*

1 Motivation and Background

Languages can be studied from two different and complementary viewpoints: the *diachronic* perspective considers the evolution of a language over time, while the *synchronic* perspective describes the language rules at a specific point of time without taking its history into account (De Saussure, 1983). In this work, we focus on the diachronic approach, since language appears to be unquestionably immersed in the temporal dimension. Language is subject to a constant evolution driven by the need to reflect the continuous changes of the world. The evolution of word meanings has been studied for several centuries, but this kind of investigation has been limited by the low amount of data on which to perform the analysis. Moreover, in order to reveal structural changes in word meanings, this analysis has to explore long periods of time.

Nowadays, the large amount of digital content opens new perspectives for the diachronic analysis of language. This large amount of data needs efficient computational approaches. In this scenario, Distributional Semantic Models (DSMs) represent a promising solution. DSMs are able to represent words as points in a geometric space, generally called WordSpace (Schiitze, 1993; Sahlgren, 2006) simply analysing how words are used in a corpus. However, a WordSpace represents a snapshot of a specific corpus and it does not take into account temporal information.

Since its first release, the Google Ngram dataset (Michel et al., 2011) has inspired a lot of works on the analysis of cultural trends and linguistic variations. Moving away from mere frequentist approaches, Distributional Semantic Models have proved to be quite effective in measuring a meaning shift through the analysis of variation of word co-occurrences. One of the earlier attempts can be dated to Gulordava and Baroni (2011), where a co-occurrence matrix is used to model the semantics of terms. In this model, similarly to ours, the cosine similarity between vectors representing a term in two different periods is exploited as a predictor of the meaning shift: low values suggest a change in the words that co-occur with the target. The co-occurrence matrix is computed with local mutual information scores and the context elements are fixed with respect to the different time

periods, hence the spaces are directly comparable. However, this kind of direct comparison does not hold when the vector representation is manipulated, like in reduction methods (SVD), or learning approaches (word2vec). In these cases, each space has its own coordinate axis. Then, some kind of alignment between spaces is required. To this end, Hamilton et al. (2016) use orthogonal Procrustes, while Kulkarni et al. (2015a) learn a transformation matrix.

In this paper, we propose an evolution of our previous work (Basile et al., 2014; Basile et al., 2015) for analysing word meanings over time. This model, differently from those of Hamilton et al. (2016) and Kulkarni et al. (2015a), creates different WordSpaces for each time period in terms of the same common random vectors; then, the resulting word vectors are directly comparable with one another. In particular, we propose an efficient method for building a DSM model which takes into account temporal information relying on a very large corpus: the Google Ngram for the Italian language. Moreover, for the first time, we provide a dataset for the evaluation of word meaning change points detection specifically set up for the Italian language.

The paper is structured as follows: Section 2 provides details about our methodology, while Section 3 describes the dataset that we have developed and the results of a preliminary evaluation. Section 4 reports final remarks and future work.

2 Methodology

Our method has its roots in a previous model based on Temporal Random Indexing (TRI) (Basile et al., 2014; Basile et al., 2015). In particular, we evolve the TRI approach in two directions: 1) we improve the system in order to manage very large datasets, such as Google Ngram; 2) we introduce a new approach based on Reflective Random Indexing (RRI) (Cohen et al., 2010) with the aim of identifying indirect inferences that can lead to the discovery of implicit connections between word meanings.

The idea behind TRI is to build different WordSpaces for each time period that we want to analyse. The peculiarity of TRI is that word vectors over different time periods are directly comparable because they are built using the same random vectors. In particular TRI works as follows:

1. Given a corpus C of documents and a vocabulary V of terms[1] extracted from C, the method assigns a random vector r_i to each term $t_i \in V$. A random vector is a vector that has values only in {-1, 0, 1} and it is sparse with few non-zero elements distributed randomly along its dimensions. The set of random vectors assigned to all terms in V are near-orthogonal;

2. The corpus C is split in different time periods T_k using temporal information, for example the year of publication;

3. For each period T_k, a WordSpace WS_k is built. All the terms of V occurring in T_k are represented by a semantic vector. The semantic vector sv_i^k for the i-th term in T_k is built as the sum of all the random vectors of the terms co-occurring with t_i in T_k. When computing the sum, we weigh the random vector; in this case we adopt a formula based on inverse document frequency. Formally, the weight is computed as $w(r_i) = log\left(\frac{C_k}{\#t_i^k}\right)$, where C_k is the total number of occurrences in T_k and $\#t_i^k$ is the occurrences of the term t_i in T_k. The idea is to give less weight to the most frequent words.

In this way, the semantic vectors across all time periods are comparable since they are the sum of the same random vectors.

RRI can be implemented by repeating the steps 2 and 3 several times. Where at each iteration random vectors are replaced by the semantic vectors built in the previous step. The idea is to model implicit connections between terms that never co-occur together, but that could occur frequently with other shared terms.

The next two sub-sections provide details about the Google Ngram dataset and the method used to automatically detect word meanings shift.

2.1 Google Ngram

Google Ngram is a very large dataset containing all the n-grams (up to five) extracted from Google Books. It is built by analysing over five millions books spanning the years from 1500 to 2012, but the developers estimate that the most reliable period is from 1800 to 2012. The dataset covers several languages including Italian. For each

[1] The terms that we want to analyse. Usually, the most n frequent terms are extracted.

language, several compressed files are released. Each file contains for each line the following information: `Ngram <TAB> year <TAB> match_count <TAB> volume_count`. For example, the line "`analysis is often described as 1991 104 5`" means that the ngram "*analysis is often described as*" occurs 104 times in 5 books in the 1991 .

We modify TRI for building the WordSpaces directly from the Google Ngram dataset. In particular, we need a pre-processing step in which we split the n-grams in several files according to the time periods we want to analyse. For example, if we fix the dimension of a time period to ten years from 1850 to 2012, we build several files for each period: T_1 = 1850-1859, T_2 = 1860-1869, ..., T_{16} = 2000-2009, T_{17} = 2010-2012. Each file contains only the n-grams that occur in the specific time period. We remove information about the year and the book count since they are not useful in the subsequent steps. Considering the previous example, the line "`analysis is often described as 104`" will be stored in the file 1990-1999.

After this pre-processing step, we can easily run TRI and RRI, where RRI can be repeated multiple times.

2.2 Change point detection

To track the word meaning change over time, for each term t_i we build a time series $\Gamma(t_i)$ taking into account several methods. A time series is a sequence of values, one value for each time period, that indicates the semantic shift of that term in the specific period. We adopt several strategies for building time series. The first strategy is based on term log-frequency; each value in the series is defined as: $\Gamma_k(t_i) = log(\frac{\#t_i^k}{C_k})$.

In order to exploit the ability of our methods in computing vectors similarity over time periods, we define two strategies for building the time series:

point-wise: $\Gamma_k(t_i)$ is defined as the cosine similarity between sv_i^k and sv_i^{k-1}. In this way, we want to capture vector changes between two time periods;

cumulative: we build a cumulative vector $sv_i^{C_{k-1}} = \sum_{j=0}^{k-1} sv_i^j$ and compute the cosine similarity with respect to the vector sv_i^k. The idea is that the semantics at point $k-1$ depends on the semantic of all the previous time periods.

Given a time series we need a method for finding significant change points in the series. We adopt the strategy proposed in (Kulkarni et al., 2015b) based on the *Mean shift model* (Taylor, 2000). According to this model, we define a mean shift of a general time series Γ pivoted at time period j as:

$$K(\Gamma) = \frac{1}{l-j}\sum_{k=j+1}^{l}\Gamma_k - \frac{1}{j}\sum_{k=1}^{j}\Gamma_k \quad (1)$$

In order to understand if a mean shift is statistically significant at time j, a bootstrapping (Efron and Tibshirani, 1994) approach under the null hypothesis that there is no change in the mean is adopted. In particular, statistical significance is computed by first constructing B bootstrap samples by permuting $\Gamma(t_i)$. Second, for each bootstrap sample P, $K(P)$ is calculated to provide its corresponding bootstrap statistic and statistical significance (p-value) of observing the mean shift at time j compared to the null distribution. Finally, we estimate the change point by considering the time point j with the minimum p-value score. Since multiple words can have the same p-value, we sort them according to their frequency. The output of this process is a ranking of words that potentially have changed meaning.

3 Evaluation

The goal of the evaluation is twofold: 1) to build a standard benchmarking for meaning shift detection for the Italian language; 2) to evaluate the performance of the proposed methods and compare them with the baseline model based on the word frequency.

A list of meaning shifts for the Italian language is not available, then we build a new dataset using a pooling strategy. In particular, we retrieve the list of meaning shifts, as explained in Section 2.2, using the cumulative strategy for each of the following methods: word frequency, TRI, TRRI with one iteration and TRRI with two iterations.

Taking into account the first 50 words for each system, we manually check for each word if a meaning shift occurs by exploiting some dictionaries. We use two dictionaries: the "Sabatino Coletti" available on-line[2] and the "Dizionario Eti-

[2] `http://dizionari.corriere.it/dizionario_italiano/`

mologico Zanichelli" available on CD-ROM. Finally, we obtain a gold standard that consists of 40 words and their corresponding change points.

All the methods, with exception of word frequency, are built using co-occurrences information extracted from 5-grams in the Google Ngram dataset for the Italian. The vector dimension is set to 1,000 for all the approaches based on Random Indexing using two non-zero elements in the random vector.

We adopt accuracy as evaluation metric. Given a list of n change points returned by the system, we compute the ratio between the number of change points correctly identified in the gold standard[3] and n. In order to identify the correct change points, we consider not only the word[4], but also the year of the change point. In particular, the year predicted by the system must be equal or greater than one of the years reported in the gold standard. We compute the accuracy using different values of n (10, 100, ALL). Results of the evaluation are reported in Table 1. In particular, we evaluate 7 systems: $logfreq$ is the baseline based on word frequency; TRI is the Temporal Random Indexing method, $TRRI1$ is the Temporal Reflective Random Indexing with one iteration, while $TRRI2$ adopts two iterations. For the methods based on Random Indexing, we investigate both the point-wise and the cumulative strategy to compute the change points.

Table 1: Results of the evaluation.

Method	acc@10	acc@100	ALL
TRI_{point}	0.0247	0.1111	0.3086
TRI_{cum}	0.0123	0.0247	0.2963
$TRRI1_{point}$	0.0000	0.0247	0.2716
$logfreq$	0.0247	0.1111	0.2346
$TRRI2_{point}$	0.0000	0.0370	0.1728
$TRRI1_{cum}$	0.0000	0.0000	0.1605
$TRRI2_{cum}$	0.0000	0.0000	0.1235

The analysis of the results shows that TRI generally provides better results than $TRRI$. Moreover, the point-wise strategy always outperforms the cumulative one. With respect to the baseline, it has the same accuracy of TRI for both acc@10 and acc@100, while it performs worse than TRI and $TRRI1$ when the accuracy is computed over the whole list of terms (ALL). These results suggest that, while there are not too many differences between the two methods considering smaller lists of results, TRI is actually able to detect more meaning shifts on a larger set of terms. $TRRI2$ always provides the worst results; we speculate that two iterations introduce too much noise in the model. A closer scrutiny to the list of words provided by $TRRI2$ highlights the presence of many foreign words: a simplistic conclusion may suggest that this approach is able to identify foreign terms that are introduced in the Italian language. However, we think that the output of this method deserves more investigations carried out by designing an ad-hoc evaluation.

The evaluation is based on the predicted year, which has to be equal or greater than one of the years reported in the gold standard, we conduct a further analysis to measure how far the prediction is from the exact value. In particular, we compute the mean and the standard deviation taking into account the differences between the predicted and the exact year. The results of this analysis are reported in Table 2. We observe the both $TRRI1_{cum}$ and $TRRI2_{cum}$ produce the best results despite their low accuracy, while TRI_{cum} reports the best trade-off between accuracy and precision in detecting the correct year. It is important to underline that the size of the time interval influences this kind of analysis since if the algorithm predicts 1900, the change point could happen in the interval 1900-1909[5]. We plan to design a more accurate analysis by exploring a time interval set to one year as future work.

Table 2: Mean and standard deviation of the differences between the predicted and the exact year.

Method	Mean	Std.Deviation
TRI_{point}	38.04	34.90
TRI_{cum}	26.45	19.60
$TRRI1_{point}$	65.86	49.96
$logfreq$	24.15	16.19
$TRRI2_{point}$	54.50	52.70
$TRRI1_{cum}$	16.61	14.62
$TRRI2_{cum}$	19.40	19.85

[3]The gold standard adopted in this evaluation is available here: `https://dl.dropboxusercontent.com/u/16026979/data/TRI_CLIC_2016_change_word`.

[4]The word matching is performed taking into account also the inflected forms.

[5]In our experiment, the size of the time interval is set to ten years.

4 Conclusions

In this work we proposed several methods based on Random Indexing for the diachronic analysis of the Italian language. We built a dataset for the evaluation of meaning shift by exploiting etymological information taken from two Italian dictionaries. We compared our approaches with respect a baseline based on word frequency obtaining promising results. In particular, the TRI method showed its better capability in retrieving more meaning shifts on a longer list of terms. As future work, we plan to extend the dataset with further words and to investigate other methods based on word-embeddings.

Acknowledgement

This work is partially supported by the project "Multilingual Entity Liking" funded by the Apulia Region under the program FutureInResearch.

References

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. Analysing word meaning over time by exploiting temporal random indexing. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors, *First Italian Conference on Computational Linguistics CLiC-it 2014*. Pisa University Press.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2015. Temporal random indexing: A system for analysing word meaning over time. *Italian Journal of Computational Linguistics*, 1(1):55–68, 12.

Trevor Cohen, Roger Schvaneveldt, and Dominic Widdows. 2010. Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of biomedical informatics*, 43(2):240–256.

Ferdinand De Saussure. 1983. *Course in general linguistics*. La Salle, Illinois: Open Court.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. Chapman and Hall/CRC.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK, July. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *CoRR*, abs/1605.09096.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015a. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 625–635, New York, NY, USA. ACM.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015b. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. ACM.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Magnus Sahlgren. 2006. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces.

Hinrich Schiitze. 1993. Word space. *Advances in neural information processing systems*, 5:895–902.

Wayne A Taylor. 2000. *Change-point analysis: a powerful new tool for detecting changes*. Taylor Enterprises, Inc.

Hybrid Language Segmentation for Historical Documents

Alfter, David and **Bizzoni, Yuri**
University of Gothenburg
{firstname.lastname}@gu.se

Abstract

English. Language segmentation, i.e. the division of a multilingual text into monolingual fragments has been addressed in the past, but its application to historical documents has been largely unexplored. We propose a method for language segmentation for multilingual historical documents. For documents that contain a mix of high- and low-resource languages, we leverage the high availability of high-resource language material and use unsupervised methods for the low-resource parts. We show that our method outperforms previous efforts in this field.

Italiano. *La segmentazione del linguaggio, la divisione di un testo multilingue in frammenti monolingue, è stata affrontata nel passato, ma la sua applicazione a documenti storici è rimasta in gran parte inesplorata. Proponiamo un metodo per la segmentazione linguistica di documenti storici multilingue. Per documenti che contengono sia lingue ad alta disponibilità di risorse che lingue sottorappresentate, utilizziamo a nostro vantaggio l'elevata disponibilità delle lingue con un'ampia gamma di risorse e impieghiamo sistemi non supervisionati per le parti che dispongono di un minor numero di risorse. Mostriamo che il nostro metodo supera gli sforzi precedenti in questo settore.*

1 Introduction

The computational processing of historical documents presents challenges that modern documents do not; often there is no standard orthography, and the documents may interleave multiple languages (Garrette et al., 2015). Furthermore, the languages used in the documents may by now be considered *dead* languages.

This work will address the issue of language segmentation, i.e. segmenting a multilingual text into monolingual fragments for further processing. While this task has been addressed in the past using supervised and weakly supervised methods such as trained language models (Řehůřek and Kolkus, 2009; King and Abney, 2013), unsupervised methods (Biemann and Teresniak, 2005; Yamaguchi and Tanaka-Ishii, 2012; Alfter, 2015a), the application to short messages (Porta, 2014; Alfter, 2015b) and the application to historical documents with regard to OCR tasks (Garrette et al., 2015), there is still room for improvement, especially concerning historical documents.

Due to the scarcity of multilingual corpora (Lui et al., 2014), a popular approach is to use monolingual training data. However, in the case of historical documents, the number of available texts in a given historical language might be too low to yield representative language models.

We propose a method that works on texts containing at least one high resource language and at least one low resource language. The intuition is to use supervised and weakly supervised methods for the high resource languages and unsupervised methods for the low resource languages to arrive at a better language segmentation; supervised methods derived from high-resource languages single out these languages while unsupervised algorithms tackle the remaining unknown language(s) and cluster them by similarity.

The presented approach is extendable to more than one high-resource language, in which case a separate language model has to be trained for each language; the approach is also applicable to more than one low-resource language, where the unsupervised methods are expected to produce an accurate split of all languages present.

2 Hybrid language segmentation

Let $D = w_1...w_n$ be a document consisting of the words w_1 to w_n. Let L_h be a character-level *n-gram* language model trained on data for a high resource language which occurs in the document D. We first apply the language model L_h to the document D and assign each word w_i the probability given by L_h (1).

$$\forall w_i \in D : P(w_i) = L_h(w_i) \quad (1)$$

The language model L_h is implemented as a trigram language model with non-linear back-off. For testing purposes, we trained a language model on a dump of the English Wikipedia (3 GB of compressed data).

Under the assumption that the text contains at least two languages with at least one word from each language, we determine the minimum probability P_{min} for a split (2). This probability corresponds to the lowest probability assigned by the language model L_h to any word in the text.

$$P_{min} = min_{i=1..n} P(w_i) \quad (2)$$

Next, we determine the maximum probability distance P_a between adjacent words (3) and the global maximum probability distance P_g between any two words (4).

$$P_a = max_{i=2..n}(\big|P(w_{i-1}) - P(w_i)\big|) \quad (3)$$

$$P_g = max_{i=1..n,j=1..n}(\big|P(w_i) - P(w_j)\big|) \quad (4)$$

We also calculate the mean probability P_{mean} between the two adjacent words which maximize P_a (5).

$$P_{mean} = \frac{P(w_i) + P(w_j)}{2} \quad (5)$$

Finally, we calculate the sharpest drop in probabilities and define $P_{mindrop}$ as the probability at the lowest point of the drop (6).

$$\begin{aligned} P_{mindrop} = & max_{i=3..n}(\big|P(w_{i-2}) - P(w_{i-1})\big| \\ & + \big|P(w_{i-1}) - P(w_i)\big|) \end{aligned} \quad (6)$$

We then set a preliminary language split threshold P_{split} based on P_{min}, P_a, P_g, P_{mean} and $P_{mindrop}$ (7).

$$P_{split} = \frac{\frac{\frac{P_a+P_g}{3}+P_{mean}}{2} + P_{mindrop}}{2} \quad (7)$$

In a first step, every word w_i with a probability P above the split threshold P_{split} is considered to belong to the high resource language modeled by L_h and is tagged as such, while every word w_j with a probability P below the split threshold is considered as belonging to an unknown language and is left untagged.

In a second step, all untagged words are clustered by similarity. This is done by using language model induction (Alfter, 2015a). All words left untagged by the previous step are regarded as one text. From the first word w_1, an initial language model L_i is created. The next word w_2 is tested against the initial model. If the probability $P(w_2|L_i)$ exceeds a certain threshold value, the model is updated with w_2, otherwise a new model is created. In this way, we iterate through the text, creating language models as necessary. The same procedure is done starting from the last word and moving towards the beginning of the text. From the two sets of language model inductions (forward, backward), the most similar models according to their n-gram distribution are then merged. This process is repeated, keeping the previously merged models, until no more models are induced.

Each word is then tagged with the language model L_m ($\approx$ cluster) which maximizes $P(w|L_m)$.

Finally, all words are evaluated in a local context using variable-length Markov Models (VMM). This step aims at eliminating inconsistencies, detecting other-language inclusions and merging back together same-language fragments. Řehůřek and Kolkus (2009) use a similar technique, but they use a fixed-width sliding window while we use a variable window size based on context.

For each word w_i, we look at its tag t_i. We then consider all the words immediately to the left of w_i and all the words immediately to the right of w_i that have a tag different from t_i. From these words, we create local context language models left (L_l) and right (L_r). We calculate the similarity between L_l and L_r as well as the similarity of w_i to L_l and L_r. There are different possible scenarios:

1. L_l is similar to L_r
   (a) w_i is similar to L_l or L_r
   (b) w_i is dissimilar to L_l or L_r
2. L_l is dissimilar to L_r
   (a) w_i is similar to L_l
   (b) w_i is similar to L_r
   (c) w_i is dissimilar to L_l and L_r

In case 1a, we assimilate the tag of w_i to the tag of either L_l or L_r; in that case, the labels for L_l and L_r are the same. In case 1b, w_i is probably an other-language inclusion, since it is dissimilar to its context, while the left and right contexts are similar. In case 2a, we assimilate the tag of w_i to the tag of L_l, and similarly in case 2b, we assimilate the tag of w_i to L_r. In case 2c, w_i is dissimilar to its context and the left and right contexts are also dissimilar. In this case, we leave the tag unchanged.

The following sections describe the data used for evaluation as well as the results.

3 Data and Evaluation

> Pacati, [Ved. pacati, Idg. *peqŭō, Av. pac-; Obulg. peka to fry, roast, Lith, kepū bake, Gr. pέssw cook, pέpwn ripe] to cook, boil, roast Vin. IV, 264; fig. torment in purgatory (trs. and intrs.): Niraye pacitvā after roasting in N.S.II, 225, PvA. 10, 14. – ppr. pacanto tormenting, Gen. pacato (+Caus. pācayato) D. I, 52 (expld at DA. I, 159, where read pacato for paccato, by pare daṇḍena pīḷentassa). – pp. pakka (q.v.). <->Caus. pacāpeti & pāceti (q. v.). – Pass. paccati to be roasted or tormented (q. v.). (Page 382)

In the absence of better comparable data, we re-use the Pali dictionary data entries presented in Alfter (2015a) and compare our calculated language segmentation to the segmentation presented in Alfter (2015a).

The extract shown corresponds to the fifth Pali text used in the experiments. It shows among others some of the languages used, the unclear boundaries between languages, abbreviations, symbols and references. Monolingual stretches tend to be short with interspersed language inclusions.

Based on the findings in Alfter (2015a) that neither a high Rand Index nor a high F-score alone yield good segmentations, but a combination of high Rand Index and F-score yield good segmentations, we have adopted a new measure of goodness-of-segmentation G_s, which is the arithmetic mean of the Rand Index and F5 score (8).

$$G_s = \frac{RI + F5}{2} \tag{8}$$

Due to how precision and recall are calculated in the context of cluster evaluation, setting $\beta > 1$, and thus placing more emphasis on recall, penalizes the algorithm for clustering together data points that are separated in the gold standard and lowers the impact splitting of data points which are clustered together in the gold standard. Indeed, it is preferable to have multiple clusters of a certain language than to have clusters of mixed languages. Thus, we use F5 ($\beta = 5$) instead of F1 scores.

We have found left context assimilation to be working better than right context assimilation or both side context assimilation. We therefore use only left context assimilation and leave out the other two options.

4 Results

The following table shows our results (Hybrid Language Segmentation, HLS) compared to the results given in Alfter (2015a) (Language Model Induction, LMI). We converted the scores given in Alfter (2015a) to the new compound score G_s. The baselines from Alfter (2015a) are also indicated. AIO indicates the baseline where each word is thrown into the same cluster; there is only one cluster (all-in-one). AID indicates the baseline where each word is separated into its own cluster; there is one cluster per word (all-in-different).

Text	AIO	AID	LMI	HLS
Pali 1	0.3174	0.4643	0.5296	0.6665
Pali 2	0.3635	0.5188	0.7662	0.5916
Pali 3	0.4996	0.3071	0.4700	0.6056
Pali 4	0.4047	n/a	n/a	0.4730
Pali 5	0.5848	0.2833	0.4402	0.5863

Table 1: Results

As can be seen from the results, our approach outperforms the baselines as well as the

purely unsupervised language model induction approach except for one data point where the language model induction produced an almost perfect clustering whereas the hybrid language segmentation method did not.

5 Discussion

A big problem with the dictionary data is that it is transcribed in a noisy manner. This is not immediately clear from looking at the data, but on closer inspection, it can be seen that some symbols like commas and full stops are rendered with non-standard Unicode characters (Unicode codepoint U+FF0C (FULLWIDTH COMMA) and Unicode codepoint U+FF0E (FULLWIDTH FULL STOP)) which break the chosen whitespace tokenization method. This results in chunks that are bigger than they should be, often containing multiple languages. We can also see that the transcription of Greek characters were rendered as character that look alike but are not actually Greek characters (see the quote at the beginning of section 3).

If we look more closely at the results, we can see that our approach tends to be overly confident when assigning words to the high-resource language, which in this case is English. This includes words that clearly are not English, such as '°itar' and '°ātar'[1]. The following example (Pali 1) shows the full dictionary entry.

> [n. ag. fr. abhijjhita in med. function] one who covets M <smallcaps>i.</smallcaps> 287 (T. abhijjhātar, v. l. °itar) = A <smallcaps>v.</smallcaps> 265 (T. °itar, v. l. °ātar).

The poor discriminatory power of the model is probably related to the training data. While the English Wikipedia offers a huge amount of training data, it also includes many non-English words in explanations and on pages about non-English non-translatable terms for example. Thus, the resulting language model is noisy.

It might be possible to increase accuracy by changing the split threshold P_{split}, but while choosing a higher P_{split} will effectively reduce the amount of erroneous English tags, it will also decrease the amount of correctly tagged words. It is possible that the unsupervised approach followed by the local context smoothing might re-assign the English words to the English model or at least to a consistent, second model. However, this remains to be tested. We think that simply using more 'pure' English training data will improve the language model's accuracy.

As for local context smoothing, we have not reached conclusive results. While in some cases, it succeeds in re-assigning the correct tag to a previously incorrectly tagged word, it also induces errors by erroneously re-tagging previously correct tags. This is most probably due to the short monolingual fragments in our data; longer monolingual fragments would yield more reliable language models. In connection to this, calculating similarity based on small contexts seems problematic. Another problem are non-words and their treatment. We have chosen not to cross non-word boundaries when calculating local context, but doing so might improve the results.

Finally, we have only tested the approach with one high resource language and a multitude of low-resource languages. It would be interesting to test the method more extensively using more high resource language models (which in turn might interfere with each other).

6 Conclusion

We have introduced a hybrid language segmentation method which leverages the presence of high-resource language content in mixed language historical documents and the availability of the necessary resources to build language models, coupled with an unsupervised language model induction approach which covers the low-resource parts. We have shown that our method outperforms the previously introduced unsupervised language model induction approach.

We have also found that our method seems to work both on longer texts and on shorter texts, whereas the approach described in Alfter (2015a) seems to be working better on shorter texts such as Twitter messages.

The local context approach yields inconclusive results. This is most probably due to the similarity measure used and the small size of the context. We would need, if possible, a better similarity measure for small language models or another method of evaluating the word in respect to its context.

[1] Here, ° stands for the root of the head word of the entry, so °itar should be read 'abhijjhitar' and °ātar should be read 'abhijjhātar'

References

Alfter, D. (2015a). Language Segmentation. Master's thesis, Universität Trier.

Alfter, D. (2015b). Language segmentation of twitter tweets using weakly supervised language model induction. *TweetMT @ SEPLN*.

Biemann, C. and Teresniak, S. (2005). Disentangling from babylonian confusion–unsupervised language identification. In *Computational Linguistics and Intelligent Text Processing*, pages 773–784. Springer.

Garrette, D., Alpert-Abrams, H., Berg-Kirkpatrick, T., and Klein, D. (2015). Unsupervised code-switching for multilingual historical document transcription. In *Proceedings of NAACL*.

King, B. and Abney, S. P. (2013). Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technologies*, pages 1110–1119.

Lui, M., Lau, J. H., and Baldwin, T. (2014). Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2:27–40.

Porta, J. (2014). Twitter Language Identification using Rational Kernels and its potential application to Sociolinguistics. *TweetLID @ SEPLN*.

Řehůřek, R. and Kolkus, M. (2009). Language identification on the web: Extending the dictionary method. In *Computational Linguistics and Intelligent Text Processing*, pages 357–368. Springer.

Yamaguchi, H. and Tanaka-Ishii, K. (2012). Text segmentation by language using minimum description length. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 969–978. Association for Computational Linguistics.

La leggibilità dei testi di ambito medico rivolti al paziente: il caso dei bugiardini di farmaci senza obbligo di prescrizione medica

Franca Orletti
Dipartimento di Studi Umanistici Università di Roma Tre
franca.orletti@uniroma3.it

Felice Dell'Orletta
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
felice.dellorletta@ilc.cnr.it

Rossella Iovino
Dipartimento di Studi Umanistici Università di Roma Tre
rossella.iovino@uniroma3.it

Abstract

English. In this paper we present the first results of an exploratory analysis of simplification of the package leaflets of medicines, considered representative texts of doctor-patient communication. It will be shown how natural language processing tools can be used to reconstruct the linguistic profile of these texts and to guide their simplification.

Italiano. *In questo articolo presentiamo i primi risultati di un lavoro esplorativo di analisi e semplificazione dei foglietti illustrativi dei medicinali, considerati testi rappresentativi della comunicazione medico-paziente. Verrà mostrato come strumenti per il trattamento automatico del linguaggio naturale (TAL) possono essere utilizzati per ricostruire il profilo linguistico di questi testi e per guidarne la semplificazione.*

1 Introduzione

Questo contributo intende affrontare la questione della semplificazione linguistica dei testi di ambito medico rivolti direttamente ai pazienti, come è il caso dei bugiardini di farmaci senza obbligo di prescrizione medica. Nonostante le iniziative ufficiali prese dall'AIFA, come quella della realizzazione della Banca Dati Farmaci per aiutare gli utenti a usare correttamente i farmaci comprendendone i benefici e gli eventuali effetti indesiderati, è noto che la lingua utilizzata in questi e altri testi di ambito medico (e non solo) è complessa e spesso non adeguata alle competenze dei cittadini. Si spiegano così anche le numerose iniziative di semplificazione linguistica diffusesi negli Stati Uniti e in Europa (compresa l'Italia) volte alla semplificazione linguistica dei testi istituzionali (cfr., tra gli altri, De Mauro 1980, Fioritto 1997, Cortelazzo 1999, De Mauro e Vedovelli 1999, Piemontese 2003, Fortis 2003).

In questo lavoro, dopo aver illustrato brevemente i principali aspetti di complessità della lingua di ambito medico e aver presentato il *corpus* di bugiardini, ci si soffermerà sul contributo offerto alla semplificazione linguistica dagli strumenti di Trattamento Automatico del Linguaggio; in particolare, si farà riferimento al software READ-IT (sezione 2). Il software è stato impiegato per monitorare (misurandone il livello di leggibilità) alcune caratteristiche linguistiche del bugiardino di VIVIN C®, scelto perché vendibile senza obbligo di prescrizione medica (sezione 2.1). READ-IT può costituire un valido aiuto per chi scrive testi di pubblica utilità: la misurazione della complessità del testo consente interventi di riscrittura nella direzione della massima leggibilità (sezione 3.2). Saranno, infine delineate delle conclusioni (sezione 4).

2 La comunicazione scritta medico-paziente

Nella complessa relazione tra medico e paziente, la lingua non è soltanto funzionale alla comunicazione, ma è utile al medico per ribadire il suo prestigio e la sua distanza sociale ed epistemica rispetto al paziente (Freidson 1970, Heritage e Maynard 2006, Heritage 2012).

La lingua medica, e soprattutto la componente tecnica che la caratterizza e che la qualifica come una lingua settoriale (Beccaria 1973) o lingua specialistica (Berruto 1993, Sobrero 1993), è uno strumento che aiuta il medico a esercitare la propria "visione professionale", proiettata sulla real-

tà quando questa è osservata dalla prospettiva dello specialista (Goodwin 1994; Orletti 2000). Questo è evidente nelle difficoltà che i pazienti incontrano nella comprensione della comunicazione di argomento medico, che viene percepita come difficile da molti punti di vista.

Il maggiore aspetto di difficoltà della lingua medica è connaturato proprio al suo essere una varietà di lingua utilizzata in una specifica situazione comunicativa collegata a un preciso ambito tematico. Oltre a ciò sono stati messi in evidenza in letteratura altri aspetti di complessità della lingua medica. Sul piano lessicale si osserva una notevole stratificazione che caratterizza questa varietà settoriale (Serianni 2005): vi si trovano, infatti, grecismi, latinismi, arabismi, anglismi usati spesso come tecnicismi. Su quello morfo-sintattico sono significativi i processi di formazione delle parole che prevedono, ad esempio, una specializzazione dei suffissi (es. *-ite* per i processi infiammatori, *-osi* per le affezioni di carattere degenerativo), la presenza di acronimi (es. TAC = Tomografia Assiale Computerizzata) e di eponomi per indicare patologie (es. *morbo di Parkinson*) oppure parti del corpo (es. *tube di Falloppio*), nonché di composti con elementi neoclassici (es. *pediatria*) (Iacobini 2004, Thornton 2004). Interessanti sono poi le osservazioni di natura sintattica e testuale (Serianni 2001), che riguardano la preferenza per aggettivi di relazione (*stato febbrile*), l'anteposizione del soggetto nelle frasi passive (*cautela deve essere prestata*), il frequente ricorso a strutture frasali impersonali (*si deve prestare attenzione*), la preferenza per nomi astratti, nominalizzazioni (*l'uso della dose minima di farmaco per il trattamento più breve possibile*) e forme nominali del verbo (*antireumatico non steroideo appartenente alla classe dei derivati dell'acido propionico*). Tutto ciò risponde alle esigenze di sintesi e compattezza che accomunano lo stile dei testi di ambito medico a quello dei testi specialistici in generale (Dardano 1986), contribuendo a renderli difficili, in quanto brevità non sempre è sinonimo di semplicità (Orletti 2015). Questi presentano, infatti, una struttura codificata (Cortelazzo 2010) molto vincolata (Sabatini 1990).

Data la complessità della lingua medica, che si realizza a livelli linguistici diversi, per l'analisi dei bugiardini è stato utilizzato il software READ-IT, sviluppato dall'*Italian Natural Language Processing Laboratory* (ItaliaNLP Lab) dell'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del CNR di Pisa (Dell'Orletta, Montemagni, Venturi, 2011). Questo software è in grado di computare parametri linguistici molto più articolati rispetto a software tradizionali che restituiscono indici di leggibilità basati per lo più sulla lunghezza di frasi e parole. READ-IT implementa, infatti, un indice di leggibilità "avanzato" basato su una analisi linguistica multi-livello del testo. In aggiunta all'analisi a livello semantico-lessicale, READ-IT è in grado di fornire una descrizione linguistica anche a livello morfologico e sintattico. Oltre a ciò, ai fini della riscrittura semplificata del testo, READ-IT offre la possibilità di avere un indice della leggibilità del testo non solo a livello dell'intero documento, ma anche della singola frase, favorendo così interventi analitici.

2.1 Il *corpus*: alcuni dati quantitativi

Il *corpus* è costituito da 7335 bugiardini estratti dal portale di informazione sanitaria e farmaceutica www.torrinomedica.it. Dal *corpus* sono stati selezionati circa 100 bugiardini relativi ad alcuni tra i farmaci vendibili senza obbligo di prescrizione medica tra più diffusi in commercio, per un totale di 189315 *token* e 16937 frasi.

La scelta è ricaduta sui farmaci vendibili senza prescrizione medica perché, essendo accessibili direttamente ai cittadini, dovrebbero presentare bugiardini quanto più possibile comprensibili per garantire un corretto uso autonomo del farmaco.

Per dare un'idea della difficoltà dei testi presi in esame, i 100 bugiardini sono stati analizzati automaticamente e arricchiti con annotazione morfo-sintattica e sintattica. A tal fine è stata utilizzata una piattaforma di strumenti per il trattamento automatico dell'italiano: per l'annotazione morfo-sintattica, lo strumento utilizzato è descritto in Dell'Orletta (2009), mentre per l'analisi sintattica a dipendenze, abbiamo utilizzato DeSR (Attardi *et al*. 2009). Il testo linguisticamente annotato costituisce il punto di partenza per le analisi successive volte a definire il profilo linguistico sottostante al testo a partire dal quale, e attraverso il confronto con il profilo linguistico di altri *corpora* di riferimento, è possibile ricostruirne un quadro della complessità. I *corpora* utilizzati nel confronto appartengono a due generi testuali diversi: quello giornalistico (Piemontese 1996 e Marinelli *et al*. 2003) e quello contenente materiali didattici (Dell'Orletta *et al*. 2011). Per ognuno dei due generi si individuano due classi: testi semplici e difficili. Per i testi giornalistici, quelli facili sono i testi estratti dal mensile *2Parole* (2P) di "facile lettura", che si presta alla comprensione anche da parte di per-

sone con deficit cognitivi, e quelli difficili sono quelli estratti da *Repubblica* (R); mentre nei testi didattici sono inclusi quelli per le scuole elementari (MDE) e superiori (MDS).

I risultati quantitativi del confronto tra i *corpora* sono riassunti nelle Tabelle 1 e 2, in cui sono forniti risultati dei confronti quantitativi tra il *corpus* di riferimento e, rispettivamente, i *corpora* ritenuti facili e difficili:

		Facili	
	Corpus	**2P**	**MDE**
***Tokens* frase**	11.18	12.14	18.36
Altezza alberi	3.21	5.29	5.54
Lunghezza link	4.4	7.91	8.89
***Tokens* clausola**	17.3	9.81	8.09

Tabella 1: Confronto quantitativo con *corpora* facili

		Difficili	
	Corpus	**R**	**MDS**
***Tokens* frase**	11.18	28.94	22.76
Altezza alberi	3.21	6.51	7.36
Lunghezza link	4.4	10.28	12.5
***Tokens* clausola**	17.3	10.12	9.2

Tabella 2 Confronto quantitativo con corpora difficili

Le frasi del *corpus* di bugiardini analizzati risultano essere piuttosto brevi con un numero di *token* per frase pari a 11.18, valore ancora più basso di quelli riscontrati nei due *corpora* di controllo ritenuti facili (12.14 per 2P 18.36 per MDE). La brevità delle frasi, insieme al fatto che molte di esse sono prive di verbo, condiziona di conseguenza almeno altri due parametri quantitativi: la media dell'altezza degli alberi sintattici che nel *corpus* di riferimento è pari a 3.21, valore molto basso rispetto a *corpora* facili come 2P (5.29) e MDE (5.54); e la lunghezza media dei link sintattici più lunghi (calcolata in termini di *token* che intercorrono tra la *testa* ed il *dipendente*) di ogni frase che è di 4.4 nel *corpus* di riferimento rispetto a 7.91 per 2P, 10.28 per R, 8.89 per MDE e 12.5 per MDS. La situazione si capovolge completamente quando si valuta il numero medio di *token* per clausola verbale. Notiamo subito come in presenza di strutture verbali il testo diventi estremamente complesso mostrando un numero medio di 17.36 *token*, che è molto alto se viene confrontato con i valori estratti dai *corpora* di riferimento considerati difficili: 10.12 per R e 9.2 per MDS.

Interessante è poi la distribuzione del lessico, riassunta nella Tabella 3:

Parole del *corpus* presenti in VdB 41.12%		
Diz. fondamentale	**Diz. alto uso**	**Diz. alta dispon.**
66.48%	25.84%	7.67%
Parole del *corpus* R in VdB 67.09%		
Diz. fondamentale	**Diz. alto uso**	**Diz. alta dispon.**
75.5%	18.95%	5.8%

Tabella 3 Distribuzione del lessico

la media di parole trovate nel Vocabolario di Base (VdB, De Mauro 1980) è soltanto del 41.12%, di cui il 66.48% appartiene al Dizionario Fondamentale, il 25.84% al Dizionario di Alto Uso e il 7.67% al Dizionario di Alta Disponibilità. Il *corpus* R, pur ritenuto difficile, ha un valore più alto di parole appartenenti al VdB (67.09%), di cui il 75.5% appartiene al Dizionario fondamentale, il 18.95% al Dizionario di Alto Uso e il 5.28% al Dizionario di Alta Disponibilità. Questo qualifica il lessico dei bugiardini analizzati come molto difficile. Ciò è dovuto al fatto che molte delle parole non contenute nel VdB coincidono con termini tecnici, termini astratti, termini appartenenti al lessico specialistico, compresi i prestiti da altre lingue moderne e i latinismi, alcuni dei quali sono utilizzati come tecnicismi e altri come espressioni utili per innalzare il registro linguistico (Orletti e Iovino in stampa).

3 Il software READ-IT per la semplificazione dei testi: esempi di riscrittura

L'applicazione di READ-IT a supporto della semplificazione dei testi dei bugiardini è stata sperimentata sul bugiardino di VIVIN C®.

READ-IT è in grado di valutare la leggibilità su due livelli: il documento e la singola frase. Per quanto riguarda il documento, il software misura la difficoltà del testo, espressa in percentuale, rispetto a quattro configurazioni diverse: *base* (che tiene in considerazione la lunghezza delle frasi e delle parole), *lessicale* (composizione del vocabolario), *sintattico* (misura delle categorie

morfo-sintattiche e struttura sintattica) e *globale* (misura riassuntiva della leggibilità sulla base della combinazione di tutti i tratti).

L'analisi effettuata con READ-IT restituisce una situazione di forte complessità per il testo del bugiardino di VIVIN C®, come si evince dai dati nella Tabella 4:

Livello globale	100%
Livello base	43.8%
Livello lessicale	99.7%
Livello sintattico	97%

Tabella 4 Complessità linguistica bugiardino VIVIN C®

la leggibilità del testo mostra, infatti, una difficoltà globale del 100%. Più in dettaglio, è interessante notare come prendendo in considerazione solo le caratteristiche di base il livello di complessità del testo è solo del 43.8%, mentre è molto più alto a livello lessicale (99.7%) e sintattico (97%).

Allo scopo di semplificare il testo preso in esame, la capacità di READ-IT di analizzare la leggibilità rispetto alle singole frasi riveste un ruolo centrale. Attraverso l'identificazione dei luoghi di complessità che necessitano di revisione e semplificazione, accompagnata da una classificazione del tipo di difficoltà riscontrata (di naturale lessicale vs grammaticale), READ-IT può essere utilizzato come guida alla semplificazione del testo. Ad esempio la prima frase del testo del bugiardino preso in esame è la seguente:

1. *Mal di testa e di denti, nevralgie, dolori mestruali, dolori reumatici e muscolari.*

L'indice di difficoltà globale della frase della frase è pari a 89.4%. A livello base, lessicale e sintattico la complessità è confermata con percentuali, rispettivamente, del 71.8%, 64.8%, 71.2%. Per quanto riguarda più in dettaglio il livello lessicale, delle 10 parole di cui è composta, se non si contano le ripetizioni, soltanto 5 si trovano nel VdB (*di*, *testa*, *e*, *mal(e)*, *dente*, *dolore*) di cui due sono parole funzionali (la congiunzione *e* e la preposizione *di*).

È possibile fare una serie di osservazioni sulle caratteristiche della frase selezionata. Innanzitutto, come si è già accennato, si nota l'assenza del verbo. Se da un lato questo si spiega con il titolo del paragrafo in cui la frase è inserita (*Indicazioni terapeutiche*), dall'altro lato sarebbe auspicabile evitare frasi nominali soprattutto in presenza di un titolo a sua volta poco chiaro contenente un aggettivo relazionale che si qualifica come un tecnicismo. La presenza di aggettivi relazionali, tipici di molti linguaggi scientifici, è notevole nella frase in esame, dal momento che se ne trovano ben 3: *mestruali*, *reumatici*, *muscolari*. Se il concetto di *dolori mestruali* e di *dolori muscolari* può risultare comprensibile perché utilizzato anche nella lingua d'uso, lo stesso non vale per *dolori reumatici*, laddove non è così noto che i *reumatismi* o *malattie reumatiche* causino scompensi a livello dell'apparato locomotore e dei tessuti di sostegno del corpo. Interessante è anche l'alternanza tra espressioni della lingua d'uso come *mal di testa* (in luogo del tecnicismo *emicrania*), e *mal di denti* (piuttosto che *odontalgia*) e tecnicismi come *nevralgie*. Alla luce di quanto osservato, una possibile proposta di riscrittura semplificata della frase comprensiva del titolo può essere la seguente:

2. *A cosa serve questa medicina.*
   Potete prendere la medicina in caso di mal di testa e mal di denti, infiammazioni dei nervi, dolori da ciclo mestruale, dolori alle articolazioni e dolore ai muscoli.

Nella proposta di riscrittura si è fatto ricorso a diverse strategie di semplificazione linguistica sebbene non tutte siano esenti da problemi. Ad esempio, si è adottata una varietà meno formale della lingua, evitando oltre ai tecnicismi specifici (es. *nevralgie*), anche quelli collaterali (Serianni, 2005) (es. *assumere* vs. *prendere*). Problematica risulta essere talvolta la sostituzione dei tecnicismi specifici, spesso acclimatati nella lingua corrente, con termini di più largo uso la cui verbosità (*infiammazione ai nervi* vs *nevralgie*) può costituire un ulteriore aspetto di difficoltà.

Inoltre, sono state evitate le frasi nominali, le frasi impersonali e quelle passive, rendendo esplicito l'agente. Infine, una maggiore informalità è stata raggiunta attraverso una allocuzione diretta al destinatario del medicinale mediante l'uso dell'allocutivo *voi* (*potete prendere*). Va da sé che la proposta di riscrittura prevede l'impiego di un maggior numero di parole, in quanto gli aggettivi relazionali sono sostituiti da sintagmi e le ellissi sintattiche sono evitate. La semplificazione del testo in seguito alla riscrittura risulta comunque evidente dal riscontro di READ-IT che stima la difficoltà del testo a livello globale a solo il 10% rispetto al 89.4% del testo originale.

La seconda e ultima frase della sezione "Indicazioni terapeutiche" è la seguente:

3. *Terapia sintomatica degli stati febbrili e delle sindromi influenzali e da raffreddamento.*

Anche in questo caso, delle 11 parole contenute nella frase, soltanto 3 (*degli*, *e*, *stato* ammesso che l'accezione del termine *stato* nel VdB non coincida con il participio passato del verbo *essere* né con Stato) appartengono al VdB e questo è un fattore significativo per la stima della difficoltà del testo pari al 96.3%. Come si è visto nell'esempio precedente anche in questo caso si segnala il ricorso ad aggettivi relazionali: se ne trovano 3 (*sintomatica*, *febbrili*, *influenzali*), che esprimono significati diversi (Pustejovsky 1995), contribuendo ad accentuare la difficoltà di comprensione. Se da una parte la *terapia sintomatica* è finalizzata a guarire i sintomi, dall'altra, gli *stati febbrili* sono causati dalla febbre così come le *sindromi influenzali* dall'influenza. In questo caso, il valore causale dell'aggettivo relazionale è segnalato dalla congiunzione con il sintagma ellittico di testa *(sindromi) da raffreddamento*. Alla luce di ciò, si può proporre la riscrittura seguente, che vede l'uso esclusivo di parole del VdB:

4. *Terapia per curare i sintomi della febbre, dell'influenza e del raffreddore.*

La difficoltà del testo riscritto ammonta al 76.3% rispetto al 96.3% del testo originale.

Dopo aver preso in considerazione due frasi risultate difficili in seguito al monitoraggio linguistico, ne viene descritta ora una ritenuta facile. La sezione successiva del bugiardino è dedicata a "Posologia e modo di somministrazione". Il termine posologia è un calco dal greco *posos* "quanto" + *logia* e sta ad indicare le dosi e le modalità di assunzione del farmaco. La frase, riportata in 5, in cui sono fornite le istruzioni per assumere il farmaco viene valutata da READ-IT con un basso indice di difficoltà (17.7%):

5. *Sciogliere in mezzo bicchiere d'acqua non gassata una o due compresse di VIVIN C. L'assunzione del prodotto deve avvenire a stomaco pieno.*

In effetti, l'unico intervento di riscrittura di questo testo potrebbe riguardare la sostituzione del nome deverbale *assunzione* con l'infinito del verbo corrispondente o meglio del verbo *prendere* oppure con la seconda persona plurale (*assumete/prendete*) come già proposto in (2), anche se l'infinito si giustificherebbe con la presenza dell'analogo *sciogliere* nella frase precedente. In luogo di *L'assunzione del prodotto deve avvenire a stomaco pieno* si può optare, quindi, per *prendere/prendete il prodotto* o *il farmaco a stomaco pieno*, con una riduzione del numero delle parole (da 9 a 6).

4 Conclusione

Questo articolo presenta un lavoro esplorativo di analisi di un *corpus* di foglietti illustrativi dei medicinali senza obbligo di prescrizione medica.

L'obiettivo era valutare l'uso di strumenti di TAL per il monitoraggio di questi testi a supporto della loro semplificazione manuale che prevede interventi linguistici volti alla semplificazione della lingua a livello lessicale, con l'eliminazione di tecnicismi, nomi astratti e deverbali, e morfosintattico, con la preferenza per l'allocutivo *voi* e l'eliminazione di frasi nominali, impersonali e passive a vantaggio di strutture transitive in cui tanto il recupero dell'agente quanto l'esplicitazione delle relazioni sintattiche tra gli elementi costitutivi della frase risultino facili e accessibili a tutte le categorie di utenti.

Bibliografia

Attardi, G., Dell'Orletta, F., Simi, M. and Turian, J. 2009. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. In Proceedings of EVALITA'09, 1-8, Reggio Emilia (Italia)

Beccaria, S.L. (1973). I linguaggi settoriali in Italia, Milano, Bompiani.

Berruto, G. (1993). "Varietà diamesiche, diastratiche, diafasiche", in Sobrero A. (ed.), Introduzione all'italiano contemporaneo. La variazione e gli usi, 2 voll, Roma-Bari, Laterza, 37-92.

Cortelazzo M., con la collaborazione di Federica Pellegrino e Matteo Viale (a cura di), 1999, Semplificazione del linguaggio amministrativo. Esempi di scrittura per le comunicazioni ai cittadini, Padova, Comune di Padova.

Cortelazzo, M. (2010), "Linguaggio giuridico-amministrativo", Enciclopedia dell'italiano Treccani.

Dardano, M. (1986), *Il linguaggio dei giornali italiani*, Roma-Bari, Laterza.

De Mauro, T. (1987), *Guida all'uso delle parole*, Roma-Bari, Laterza.

De Mauro T., Vedovelli M. (a cura di), 1999, *Dante, il gendarme e la bolletta. La comunicazione pubblica in Italia e la nuova bolletta Enel*, Bari-Roma, Laterza.

Dell'Orletta, F. 2009. *Ensemble system for Part-of-Speech tagging*. In Proceedings of EVALITA'09, pp. 1-8, Reggio Emilia (Italia).

Dell'Orletta, F. Montemagni, S., Vecchi, E., M., and Giulia Venturi. 2011. Tecnologie linguistico–computazionali per il monitoraggio della competenza linguistica italiana degli alunni stranieri nella scuola primaria e secondaria. In G. C. Bruno, I. Caruso, M. Sanna, I. Vellecco (eds.), Percorsi migranti: uomini, diritto, lavoro, linguaggi, McGraw–Hill, 319–336.

Fioritto, A. 1997, Manuale di stile. Strumenti per semplificare il linguaggio delle amministrazioni pubbliche, Bologna, il Mulino.

Fortis, D. 2003, Il plain language, quando le istituzioni si fanno capire, I quaderni del MdS; Raso, T. 2003. La scrittura burocratica. Carocci. Roma;

Freidson, E. (1970) Professional Dominance: The Social Structure of Medical Care, New York, Atherton Press.

Goodwin, C. (1994), "Professional vision", American Anthropologist, 96 (3) 606-33.

Heritage, J. (2012), "Epistemics in action: Action formation and territories of knowledge", Research on Language and Social Interaction, 45, 1-25.

Heritage, J. e D. Maynard (eds) (2006), Communication in medical care: interaction between primary care physicians and patients, Cambridge, CUP.

Iacobini, C. (2004), "Composizione con elementi neoclassici", in M. Grossmann e F. Rainer (eds), La formazione delle parole in italiano, Tübingen, Niemeyer, pp. 69-95.

Marinelli, R., Biagini, L., Bindi, R., Goggi, S., Monachini, M., Orsolini, P., Picchi, E., Rossi, S., Calzolari, N., Zampolli, A. (2003). The Italian PAROLE corpus: an overview. In Zampolli A. et al. (eds.), Computational Linguistics in Pisa, Special Issue, XVI–XVII, Pisa-Roma, IEPI. Tomo I, 401–421.

Orletti, F. (2000), La conversazione diseguale. Potere e interazione, Roma, Carocci.

Orletti, F. (2015), "Quando breve non è semplice: sul rapporto tra scritture brevi e semplicità", Convegno interannuale PRIN SCRIBE S*critture brevi, forme, modelli e applicazioni per l'analisi e per il dizionario*, Macerata 28-30 maggio 2015.

Orletti, F. e Iovino, R. (in stampa), Il parlar chiaro nella comunicazione medica. Fra etica e linguistica, Roma, Carocci.

Piemontese, M.E. (1996), Capire e farsi capire. Teorie e tecniche della scrittura controllata, Tecnodid, 1996.

Pustejovsky, J. (1995) The Generative Lexicon, Cambridge, Mass, The MIT Press.

Sabatini, F. 1990. Analisi del linguaggio giuridico. Il testo normativo in una tipologia generale dei testi , in M. D'Antonio (cur.), Corso di studi superiori legislativi (1988-1989) , Padova, Cedam, pp. 675-724.

Serianni, L. (2001), Italiani scritti, Bologna, il Mulino.

Serianni, L. (2005), Un treno di sintomi. I medici e le parole: percorsi linguistici nel passato e nel presente. Milano, Garzanti.

Sobrero, A.A. (1993), "Lingue speciali", in Sobrero A. (ed.), Introduzione all'italiano contemporaneo. La variazione e gli usi, 2 voll, Roma-Bari, Laterza, 237-278.

Thornton, A. (2004), "Conversione", in M. Grossmann e F. Rainer (eds), La formazione delle parole in italiano, Tübingen, Niemeyer, pp. 499-533.

La Modellazione Diacronica di Risorse Termino-Ontologiche nell'Ambito delle Digital Humanities: Esperimenti su Clavius

Silvia Piccini, Andrea Bellandi, Giulia Benotto, Emiliano Giovannetti
Istituto di Linguistica Computazionale A. Zampolli
Via G. Moruzzi, 1 - CNR Pisa, Italy
name.surname@ilc.cnr.it

Abstract

English. In this work, we present an experiment in the modeling of a diachronic termino-ontological resource named CLAVIUS through both the N-ary relations model and the 4D-fluents approach. Some of the salient differences of these two models are discussed. The overall objective of this research is to illustrate the main advantages and disadvantages in the adoption of a given model to build diachronic resources.

Italiano. *In questo lavoro, si illustra un esperimento di modellazione di una risorsa termino-ontologica diacronica (CLAVIUS) secondo due approcci, quello N-ario e quello dei 4D-fluents. Le differenze salienti dei due approcci verranno presentate e discusse. L'obiettivo generale della ricerca qui introdotta è quello di mostrare i principali vantaggi e svantaggi che l'adozione di un determinato modello può comportare nella modellazione di risorse diacroniche.*

1 Introduzione

Pànta rei è la celebre espressione attribuita da Platone ad Eraclito. Tutto è sottoposto alla inesorabile legge del mutamento: la realtà, le categorie attraverso le quali la organizziamo e le parole che usiamo per parlare di essa.

Quali sono gli strumenti a disposizione dell'umanista digitale di oggi, che si trovi a dover rappresentare in modo esplicito e formale tale evoluzione diacronica dei concetti e dei termini in un determinato ambito, in modo che tale formalizzazione sia computabile ad un calcolatore?

In questi ultimi anni, ed in particolar modo nell'ambito delle *Digital Humanities*, si è sottolineata l'importanza di operare con tecnologie che siano alla base del *Semantic Web* e dei *Linked Open Data* per garantire interoperabilità e riuso delle risorse all'interno della comunità scientifica (Ciotti, 2014).

In questa ottica, le ontologie - e l'OWL, il loro linguaggio di rappresentazione standard - giocano un ruolo fondamentale. Tuttavia, il carattere fondamentalmente statico di questi ultimi e la necessità di modellare aspetti di evoluzione temporale sembrano a prima vista inconciliabili.

Le riflessioni che presentiamo in questo articolo nascono dalle esperienze condotte in seno al Progetto *Clavius on the Web*[1]. Tra gli obiettivi del Progetto, infatti, vi è anche quello di creare una risorsa termino-ontologica (RTO) che rappresenti l'evoluzione delle teorie matematico-astronomiche dall'antichità al XVI - XVII secolo, così come viene descritta da Clavius nei suoi *Euclidis Elementorum Libri XV. Accessit XVI* e *In sphaeram Ioannis de Sacro Bosco Commentarius*.

2 Il Contesto

Come sottolineato nell'Introduzione, il linguaggio OWL (e la sua estensione OWL2) è lo standard W3C per la creazione e condivisione di ontologie nel Semantic Web.

In particolare, OWL DL implementa la logica descrittiva $SHOIN(D_n)$, che garantisce una maggiore espressività rispetto a RDF e RDFS, senza compromettere la decidibilità e il meccanismo inferenziale.

Tuttavia, OWL è un linguaggio statico; in esso le proprietà e le relazioni tra entità sono fondamentalmente binarie, espresse sotto forma di triple <Subject predicate Object>. Tale restrizione sintattica rende più complessa la rappresentazione

[1] http://claviusontheweb.it (ultimo accesso: 13/10/2016)

di entità temporali, caratterizzate da relazioni ternarie.

Diversi approcci sono stati proposti in letteratura per superare tale limite e rappresentare l'evoluzione diacronica dei concetti (Krieger, 2014; Welty et al., 2006). Tra questi vi sono: i) il versioning, che consiste nel creare una versione differente dell'ontologia, una per ciascun istante temporale (Grandi and Scalas, 2009), ii) l'estensione delle triple RDF con argomenti supplementari temporali (Krieger, 2012), iii) la reficazione, che trasforma le proprietà in classi (Manola and Miller, 2004), iv) il modello N-ario (Noy and Rector, 2006), v) il modello perdurantista o dei *4D-fluents* (Welty et al., 2006), vi) la codifica dell'estensione temporale attraverso nuove proprietà sintetiche (Gangemi, 2011). Allo stato dell'arte, i due principali approcci adottati nella modellazione di risorse diacroniche (solitamente ontologie) sono il modello delle relazioni N-arie e il modello dei perduranti. Nel primo, per ogni relazione temporale viene introdotto un nuovo oggetto, istanza della classe EVENT. La durata della relazione coincide con la durata dell'evento. In questo modello, pertanto, una proprietà n-aria è rappresentata come una classe, le cui istanze corrispondono alle istanze della proprietà. Nel modello *4D-fluents*, invece, le entità coinvolte in relazioni temporali sono rappresentate da oggetti 4-dimensionali detti *timeslices*. Ogni *timeslice* rappresenta una sorta di "fotogramma", in cui l'entità ad esso riferita ha natura immutabile (perdurante) in quello specifico istante o intervallo di tempo.

3 La risorsa CLAVIUS

Clavius costituisce una RTO. In accordo con le più recenti acquisizioni teoriche (Roche, 2007; Temmerman et al., 2005), in essa infatti sono formalizzate come indipendenti due componenti strettamente legate tra di loro: la componente terminologica nella quale vengono strutturati i termini presenti in un testo e la componente ontologica attraverso la quale vengono formalizzati i concetti evocati da quei termini. Il livello più alto dell'ontologia è rappresentato così dalle due classi OWL disgiunte CONCEPT e TERM. La superclasse CONCEPT sussume due sottoclassi, ASTRONOMICAL CONCEPT e MATHEMATICAL CONCEPT, che a loro volta sussumono rispettivamente tutti i concetti astronomici e matematici. La superclasse TERM si articola, invece, in due sottoclassi ASTRONOMICAL TERM e MATHEMATICAL TERM. Quest'ultima sussume le classi OWL disgiunte LATIN TERM e GREEK TERM, di cui i termini latini e greci costituiscono le istanze. Nella risorsa la parte concettuale è espressa in inglese mentre la parte terminologica è costituita dai termini latini, legati ai concetti che essi evocano attraverso la relazione *denotes* e la sua inversa *isDenotedBy*. L'ontologia è attualmente composta da 106 classi organizzate in quattro livelli gerarchici, un insieme di 10 *Data Properties* e 18 *Object properties*, che permettono di dare una rappresentazione precisa dei concetti e termini. Le relazioni sono suddivise in: relazioni lessicali, paradigmatiche tra i termini (iperonimia, iponimia, meronimia, olonimia, sinonimia e antonimia); relazioni interlivello tra termini e concetti; relazioni concettuali, molte delle quali sono state introdotte per dare una definizione più precisa del dominio matematico-astronomico (Piccini et al., 2016). Per quanto concerne la modellazione degli aspetti temporali è stata importata l'ontologia OWL-Time, che si basa sulle relazioni binarie tra intervalli introdotte in (Allen and Ferguson, 1997) ed è stata integrata una serie di regole *Semantic Web Rule Language* (SWRL), elaborate in (Batsakis et al., 2016). In questo modo possono essere rappresentate e trattate dai motori inferenziali sia informazioni temporali quantitative (informazioni temporali precise), sia informazioni temporali qualitative (delle quali, cioè, non si possono specificare con esattezza l'istante iniziale e finale). Queste ultime, in particolare, sono molto frequenti nel testo di Clavius, e più in generale, nella letteratura, dove non è sempre possibile definire con esattezza la data di inizio e di fine nella quale si è verificato o sussiste un determinato fatto[2]. Le relazioni tra le entità ontologiche sono state temporalizzate in base alla loro validità nel tempo, seguendo i due approcci, quello N-ario e quello perdurantista, come descritto nella Sezione 2.

4 Comparazione della risorsa nei due modelli

In questa Sezione sono definiti i criteri di comparazione dei due modelli e vengono presentati

[2]La rappresentazione di tali informazioni temporali è indipendente dai modelli ivi trattati e non verrà pertanto discussa.

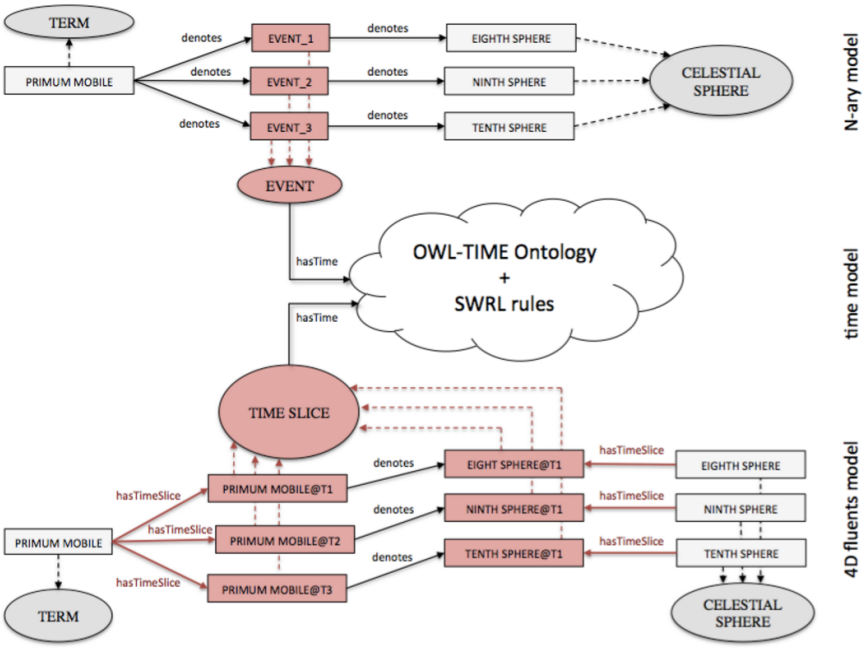


Figure 1: I due modelli a confronto. In rosso i nuovi oggetti da creare nei diversi modelli. Le linee tratteggiate rappresentano la relazione di *instance of*.

i vantaggi e le criticità di entrambi nella loro applicazione alla risorsa Clavius in OWL. Un primo criterio di comparazione riguarda il numero di assiomi logici necessari alla rappresentazione delle relazioni temporali. Come illustrato in precedenza, infatti, entrambi i modelli richiedono l'introduzione di nuove entità temporali (eventi o *timeslice*). A titolo d'esempio, presentiamo il mutamento nel tempo del concetto denotato dal termine *primum mobile*. Il numero delle sfere ed il loro relativo ordine costituisce un problema ampiamente dibattuto nell'astronomia del Seicento. Secondo la visione aristotelica del *kosmos* il movimento dei sette cieli attorno alla terra era dovuto ad una sfera, la più esterna, detta *primum mobile*. Mano a mano che nei secoli si vennero a scoprire maggiori dettagli sul moto dei pianeti, divenne necessario aggiungere altre sfere: così con Tolomeo il loro numero salì a 9, nelle tavole alfonsine a 10, quindi a 11 ed infine a 12 con Magini. Alla introduzione di una ulteriore sfera sul piano concettuale corrisponde un mutamento nel codominio della relazione interlivello *denotes* il cui dominio è rappresentato dal termine *primum mobile*: i) primum_mobile *denotes* eighth_sphere (dal 300 a.C. al 152 d.C.), ii) primum_mobile *denotes* ninth_sphere (dal 152 d.C. al 1252 d.C.), iii) primum_mobile *denotes* tenth_sphere (dal 1252 d.C. al 1589 d.C.). La Figura 1 mostra un esempio di rappresentazione di questi *statements*.

Le classi, le proprietà e le istanze in rosso, rappresentano le entità ontologiche che debbono essere introdotte per tradurre in linguaggio formale tali *statements*. Si nota come il modello dei perduranti sia soggetto ad una maggiore proliferazione delle entità: accanto alla classe TIMESLICE sono stati introdotti una nuova *Object property* (*hasTimeSlice*), istanziata sei volte, e sei nuovi individui della classe TIMESLICE (le parti temporali degli individui coinvolti nella temporalizzazione).

Il modello N-ario, d'altro canto, risulta più snello, richiedendo solo la creazione di tre eventi, uno per ogni *statement*. Tuttavia, tale modellazione modifica, da un punto di vista intuitivo, la semantica della relazione *denotes*, spezzando e duplicando la proprietà per legare l'evento ai due individui coinvolti nella temporalizzazione. In entrambi i modelli il dominio e il codominio di una proprietà temporale vengono modificati: nel modello N-ario si aggiunge ad essi la classe EVENT, nel modello dei perduranti si sostituiscono le classi originarie con la classe TIMESLICE. Precisiamo che la nostra analisi è limitata ai modelli delle relazioni da temporalizzare, ed è quindi indipendente sia rispetto al modo in cui gli intervalli e gli istanti temporali sono rappresentati (OWL-TIME in Figura 1), sia rispetto a come i costrutti e le regole di OWL ("SWRL rules" in Figura 1) vengono

Table 1: Confronto tra i modelli

	modello n-ario	**modello 4D-fluents**
Metriche dell'ontologia		
numero di classi	106	106
numero di individui	244	265
numero di istanze di object property	117	159
Temporalizzazione del predicato		
$pred : X \rightarrow Y$	$pred_{n-ary}$: $X \sqcup Event \rightarrow Y \sqcup Event$	$pred_{perd}$: $TimeSlice \rightarrow TimeSlice$
Interrogazione		
query SPARQL: lunghezza del path	2	3
query SPARQL: numero di variabili	2	3
Tools		
Availability	yes	no^*

combinati per fornire una procedura di reasoning solida e completa. Per una trattazione accurata di queste problematiche rimandiamo a (Batsakis et al., 2016). In relazione ad una maggiore proliferazione di entità, inoltre, l'interrogazione della risorsa può diventare più complessa e quindi maggiormente prona ad errori. Per quanto riguarda l'interrogazione, abbiamo considerato una semplice query di interrogazione della risorsa per conoscere quali sono i concetti che il termine *primum_mobile* ha denotato nel tempo. Le query sono state implementate tramite il linguaggio di interrogazione SPARQL nei due modelli. Per il modello N-ario abbiamo:

```
SELECT ?concept
WHERE { n-ary:primum_mobile a n-ary:TERM .
n-ary:primum_mobile n-ary:denotes ?event .
?event a n-ary:EVENT .
?event n-ary:denotes ?concept }
```

Per il modello dei perduranti abbiamo:

```
SELECT ?concept
WHERE { ?ts@t1 4d:denotes ?ts@t2 .
?ts@t1 a 4d:TIMESLICE .
?ts@t2 a 4d:TIMESLICE .
4d:primum_mobile a 4d:TERM .
4d:primum_mobile 4d:hasTimeSlice ?ts@t1 .
?concept 4d:hasTimeSlice ?ts@t2 }
```

Da una analisi dei due codici si evince che per interrogare il secondo modello occorre un numero maggiore di variabili nella query e che il path dell'interrogazione del grafo è più lungo e coinvolge un maggior numero di entità ontologiche. Questo implica che, in termini di scalabilità ed efficienza nel processo di interrogazione, il modello N-ario potrebbe risultare migliore. Come ultimo criterio abbiamo considerato la disponibilità di strumenti che supportino l'umanista nella creazione di RTO diacroniche senza richiedere la conoscenza dei modelli matematici sottostanti. Per quanto riguarda il modello N-ario è disponibile il plug-in di *Protégé CHRONOS-ED* (Preventis et al., 2012). Il plug-in per il modello *4D-fluents*, al contrario, risulta ad oggi non funzionante nelle ultime versioni di *Protégé* (vedi asterisco in Tabella 1). I risulati dell'analisi comparativa sono riassunti nella Tabella 1.

5 Discussione e lavori futuri

La modellazione della risorsa diacronica termino-ontologica CLAVIUS attraverso il modello a relazioni N-arie e quello *4D-fluents* ha consentito di valutare, empiricamente, quelle che sono le differenze salienti tra i due paradigmi temporali. Per quanto riguarda l'aspetto di *authoring*, di primario interesse in questo lavoro in quanto funzionale alla necessità di un umanista digitale nella costruzione (tipicamente manuale) di una risorsa diacronica, i risultati hanno evidenziato come, ad oggi, l'approccio a relazioni N-arie risulti il più vantaggioso, in termini sia di numero di entità ontologiche coinvolte, sia di complessità delle query SPARQL sia di disponibilità di strumenti di supporto. Si intende approfondire questa ricerca prendendo in considerazione gli aspetti che, naturalmente, seguono quelli più essenziali di formalizzazione della terminologia e della conoscenza. In primis, verrà studiato l'impatto determinato dall'adozione di un certo modello sui meccanismi di *reasoning*, e quindi, sulla possibilità da parte di un motore inferenziale di derivare nuova conoscenza a partire da quella rappresentata esplicitamente nella risorsa.

References

James F. Allen and George Ferguson. 1997 *Actions and Events in Interval Temporal Logic*. In: Spatial and Temporal Reasoning. O. Stock, ed., Kluwer, Dordrecht, Netherlands, pp. 205-245.

Sotiris Batsakis, Euripides Petrakis, Ilias Tachmazidis and Grigoris Antoniou 2016. Temporal Representation and Reasoning in OWL 2. *Semantic Web journal*, ISSN 1570-0844.

Fabio Ciotti. 2014 *Digital Literary and Cultural Studies: State of the Art and Perspectives*. Between, vol. 4, no. 8.

Noy N. Fridman, William Grosso, and Mark A. Musen 2000. *Knowledge-Acquisition Interfaces for Domain Experts: An Empirical Evaluation of Protege-2000*. SEKE2000 Proceedings. Chicago.

Aldo Gangemi. 2011 *SuperDuper schema: an OWL2+RIF DnS pattern*. KCAP 2011 Deep Knowledge Representation Challenge Workshop.

Fabio Grandi, and Maria Rita Scalas. 2009 *The Valid Ontology: A simple OWL temporal versioning framework*. Advances in Semantic Processing, SEMAPRO'09. Third International Conference IEEE, 2009. pp. 98-102.

Hans U. Krieger. 2014 *A detailed comparison of seven approaches for the annotation of time-dependent factual knowledge in RDF and OWL*. Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation, p. 1.

Hans U. Krieger. 2012 *A temporal extension of the Hayester Horst entailment rules and an alternative to W3C's n-ary relations*. Proceedings of the 7th International Conference on Formal Ontology in Information Systems (FOIS), pp. 323–336.

Frank Manola and Eric Miller. 2004 *RDF primer*. Technical report, W3C.

Natasha Noy and Alan Rector. 2006 *Defining N-ary Relations on the Semantic Web*. W3C Working Group Note 12, https://www.w3.org/TR/swbp-n-aryRelations/

Piccini, S., Bellandi, A., and Benotto, G. 2016 *Formalizing and Querying a Diachronic Termino-Ontological Resource: the CLAVIUS Case Study*. Proceedings of From Digitization to Knowledge 2016 workshop (D2K), July 11, 2016, Krakow (Poland).

Preventis A., Marki P., Petrakis E.G.M., Batsakis S. 2016 *Chronos: A Tool for Handling Temporal Ontologies in Protégé*. Proceedings of the 24th International Conference on Tools with Artificial Intelligence, Athens, Greece.

Christophe Roche. 2007 *Le terme et le concept: fondements d'une ontoterminologie*. Actes de la première conférence TOTh, pp.1–22, Annecy.

Evren Sirin, Bijan Parsia, Bernardo C. Grau, Aditya Kalyanpur, Yarden Katz 2007. *Pellet: A practical owl-dl reasoner*. Web Semantics: science, services and agents on the World Wide Web, 5.2 51-53.

Rob Shearer, Boris Motik, and Ian Horrocks. 2008 *HermiT: A Highly-Efficient OWL Reasoner*. OWLED, Vol. 432.

Rita Temmerman, Koen Kerremans, and Veerle Vandervoort 2005 *La termontographie en contexte(s)*. In Blampain, D., Thoiron, P., Van Campenhoudt, M. (eds). Mots, Termes et Contextes. Actes des septièmes Journées scientifiques du réseau de chercheurs Lexicologie Terminologie Traduction. Bruxelles, 429- 439.

Chris Welty, Richard Fikes, and Selene Makarios 2006 *A reusable ontology for fluents in OWL*. Proceedings of the 4th International Conference on Formal Ontology in Information Systems (FOIS), Vol. 150. pp. 226–336

Panta rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek

Martina A. Rodda
Scuola Normale Superiore
Piazza dei Cavalieri, 7
56126 Pisa – ITALY
martina.rodda@sns.it

Marco S.G. Senaldi
Scuola Normale Superiore
Piazza dei Cavalieri, 7
56126 Pisa – ITALY
marco.senaldi@sns.it

Alessandro Lenci
CoLing Lab
Università di Pisa
via S. Maria 36
alessandro.lenci@unipi.it

Abstract

English. We present a method to explore semantic change as a function of variation in distributional semantic spaces. In this paper we apply this approach to automatically identify the areas of semantic change in the lexicon of Ancient Greek between the pre-Christian and Christian era. Distributional Semantic Models are used to identify meaningful clusters and patterns of semantic shift within a set of target words, defined through a purely data-driven approach. The results emphasize the role played by the diffusion of Christianity and by technical languages in determining semantic change in Ancient Greek and show the potentialities of distributional models in diachronic semantics.

Italiano. *Si presenta un metodo per indagare il cambiamento semantico come funzione della variazione all'interno di spazi semantici. Questo approccio è applicato per identificare automaticamente aree di cambiamento semantico nel lessico greco antico tra età pre-cristiana e cristiana. Modelli della Semantica Distribuzionale sono usati per identificare cluster e pattern di cambiamento semantico in una lista di parole target, definita con un approccio puramente data-driven. I risultati mostrano il ruolo della diffusione del Cristianesimo e dei linguaggi tecnici nel determinare cambiamenti semantici in greco antico, nonché le potenzialità dei modelli distribuzionali nella semantica diacronica.*

1 Introduction and Related Work

Distributional Semantics is grounded on the assumption that the meaning of a word can be described as a function of its collocates in a corpus. This suggests that diachronic meaning shifts can be traced through changes in the distribution of these collocates over time (Sagi et al., 2011). While some studies focused on testing the explanatory power of this method over frequency- and syntax-based approaches (Wijaya and Yeniterzi, 2011; Kulkarni et al., 2015), more advanced contributions to the field explored how distributional models can be used to test competing hypotheses about semantic change (Xu and Kemp, 2015), or to investigate the productivity of constructions in diachrony (Perek, 2016). The results attest the explanatory power of distributional methods in modeling diachronic shifts in meaning.

In this paper, we propose a method to identify semantic change through the **Representational Similarity Analysis** (RSA; Kriegeskorte and Kievit, 2013) of distributional vector spaces built from diachronic corpora. RSA is a method extensively used in neuroscience to test cognitive and computational models by comparing the geometry of their representation spaces (Edelman, 1998). Stimuli are represented with a representational dissimilarity matrix that contains a measure of the dissimilarity relations of the stimuli with each other. Different matrices are compared to evaluate the correspondence of the representational spaces built from different sources (e.g., behavioral and neuroimaging data). We argue that this method can be applied to compare distributional representations of the lexicon at different temporal stages. The hypothesis is that the elements in the lexical spaces showing larger geometrical variations in time correspond to the lexical areas that have undergone major semantic

changes. To the best of our knowledge, this is the first time RSA is used in diachronic distributional semantics.

Here we present a case study that applies RSA to track patterns of semantic change within the lexicon of Ancient Greek. We focus on the first few centuries AD, when the rise of Christianity caused a deep and widespread cultural shift within the Hellenic world. We predict that this shift will be reflected in the Greek lexicon of the time. In addition to past studies (Boschetti, 2009; O'Donnell, 2005 is a general introduction), we apply a bottom-up approach to the detection of semantic change, with no prior definition of a list of lemmas to be analyzed. The goal is to develop a quantitative "discovery procedure" to detect lexical semantic changes.

From a methodological standpoint, this study aims to show how Distributional Semantics can be applied fruitfully to such a small and literary corpus as the collection of Ancient Greek texts. The results will also highlight the ways in which Distributional Semantics can complement the intuition of the researcher in analyzing semantic change in Ancient Greek, providing a useful tool for future studies in Classics.

2 Materials and Methods

The corpus used for this study is based on the TLG-E (*Thesaurus Linguae Graecae*) collection of Ancient Greek literary texts. The database was divided into two sub-corpora, the first of which contains texts from the 7th to the 1st century BC (pre-Christian era), while the second one spans from the 1st to the 5th century AD (early Christian era). The pre-Christian sub-corpus contains 6,795,253 tokens, while the Christian sub-corpus totalizes 29,051,269 tokens.

The texts were lemmatized using *Morpheus* (Crane, 1991). Any issues with the lemmatization should not have a significant impact on the results unless otherwise stated (cf. Boschetti, 2009, page 60 for a discussion). After filtering for stopwords (mainly particles, pronouns and connectives) and lemmas occurring with a frequency below 100 tokens, the pre-Christian and Christian sub-corpus contain, respectively, 4,109 and 10,052 lemmas, which were used both as targets and dimensions in our vector spaces.

A vector space model was then built for each sub-corpus using the DISSECT toolkit (Dinu et al., 2013). Henceforth, we refer to the pre-Christian era model as the **BC-Space**, and to the Cristian era model as the **AD-Space**. Co-occurrences were computed within a window of 11 words (5 content words to the right and to the left of each target word). Association scores were weighted using positive point-wise mutual information (PPMI) (Evert, 2008); the resulting matrices were reduced to 300 latent dimensions using Singular Value Decomposition (SVD).

2.1 RSA of the distributional vector spaces

We have adapted the RSA method to discover semantic changes between the two vector spaces:
1. we identified the words occurring in both sub-corpora with a frequency higher than 100 tokens, obtaining 3,977 lemmas;
2. we built a representational similarity matrix (RSM) from the BC-Space (RSM_{BC}) and one from the AD-Space (RSM_{AD}). Each RSM is a square matrix indexed horizontally and vertically by the 3,977 lemmas and containing in each cell the cosine similarity of a lemma with the other lemmas in a vector space (this is a minor variation with respect to the original RSA method, which instead uses dissimilarity matrices). A RSM is a global representation of the semantic space geometry in a given period: vectors represent lemmas in terms of their position relative to the other lemmas in the semantic space;
3. for each lemma, we computed the Pearson correlation coefficient between its vector in RSM_{BC} and the corresponding vector in RSM_{AD}.
The Pearson coefficient measures the degree of semantic shift across the two temporal slices. The lower the correlation, the more a word changed its meaning.

3 Discussion of Results

The following section focuses on the words that underwent the biggest changes, i.e. those for which the correlation scores are lower. The primary goal will be to establish whether these words can be clustered into meaningful groups. This would allow us to pinpoint the areas within the lexicon of Ancient Greek that have undergone a significant semantic shift during the early centuries of Christianity.

3.1 Qualitative Analysis

The 50 lemmas with the lowest correlation coefficients were scrutinized in order to establish whether meaningful subgroups emerge. (This list of words is not reproduced here due to space constraints. They are a subset of the 200 words used to build the plot in section 4.3.) The findings in this section, while inevitably limited by

the intuition of the researcher, will provide the starting point for a more sophisticated analysis to be performed in the following sections.

The lemmas under consideration form a somewhat heterogeneous collection, including concrete nouns and relatively common verbs such as ζυγόν (zygón "yoke") and ἕπομαι (hépomai "follow"), as well as some proper nouns. This notwithstanding, a promising subset of words emerges even at this preliminary stage. These are a number of nouns designating eminently Christian concepts, such as παραβολή (parabolé "parable", previously "comparison"), λαός (laós, used for the Christian community as opposed to non-Christians, previously "people"), κτίσις (ktísis "creation", previously "founding, settling").

These findings are in line with the idea that the diffusion of Christianity played a substantial role in semantic change in the first centuries AD (cf. Boschetti, 2009). Other Christian terms, such as θεός (theós "God"), ἄγγελος (ángelos "angel", previously "messenger"), πατήρ (patér "father"), υἱός (hyiós "son"), also occur among the 100 words with the lowest correlation coefficients.

Another group of lemmas comprises technical terms whose usage seems to have undergone a specialization or a shift from one domain of knowledge to another. These include words such as ὑπόστασις (hypóstasis "substance", previously "sediment, foundation"), δύναμις (dýnamis "property (of beings)", previously "power"), or ῥητός (rhetós "literal" as opposed to "allegorical", previously "stated").

3.2 Analysis of Nearest Neighbors

To corroborate the intuitions detailed above, the 10 nearest neighbors for each of the last 50 words according to the correlation coefficient were retrieved using DISSECT. The process was repeated for each sub-corpus and the results compared in order to look for visible shifts, especially those involving different semantic domains. A few examples of the results should suffice to confirm the findings in the last section.

For instance, among the nearest neighbors for πνεῦμα (pnêuma "spirit", previously "breath") in the AD-Space we find such words as θεάομαι (theáomai "contemplate"), ἀληθινός (alethinós "true"), κτίσις, υἱός, θεός and so forth, while in the BC-Space the strongest similarity is with terms pertaining to the domain of physics, such as ἀήρ (aér "air"), ὑγρός (hygrós "moist"), θερμός (thermós "hot"). Another clear-cut example is that of δύναμις, whose neighbors change from military terms such as πολιορκία (poliorkía "siege") and στρατόπεδον (stratópedon "encampment, army") to the physical and philosophical domain, with the closest term being ἐνέργεια (enérgeia "activity, actuality", an antonym of δύναμις in its philosophical sense of "potentiality"). The case of δύναμις also shows how nearest neighbor analysis can reveal shifts in the usage of heavily polysemous words.

Not all changes observed through the analysis of nearest neighbors, however, are so easily predictable. Thus, for instance, the neighbors for μοῖρα (môira, another highly polysemous word with meanings spanning from "part" to "destiny") in the AD-Space come exclusively from the domain of astronomy, showing a strong specialization towards a technical usage ("degree" or "division" of the Zodiac). Another remarkable result comes from a geographical adjective, Ποντικός (Pontikós "coming from Pontus"), whose nearest neighbors shift from proper names and philosophical terms in the pre-Christian age (an association due, without doubt, to the usage of "Ponticus" as an epithet for authors, e.g. Heraclides) to names of currency and trade wares, probably as a reflection of the integration of Pontus as a Roman province (with the obvious repercussions on trade) in the 1st century AD.

3.3 t-SNE Plot

As a final analysis, we embedded the RSM_{AD} vectors for the 200 words with the lowest correlation coefficient with the corresponding RSM_{BC} vectors in a two-dimensional space with t-SNE (Figure 1), a technique for dimensionality reduction and data visualization that overcomes some of the limitations of standard multidimensional scaling (van der Maaten and Hinton, 2008). This procedure allows for easy identification of clusters, thus revealing the semantic relation between the most recent meanings of the words that underwent the greatest semantic change.

A number of small clusters can be observed in the plot. Near the left periphery, the most relevant group is composed of terms pertaining to Christian theology (from κύριος kýrios "Lord", λαός and θεός, to παρουσία parousía "Advent" and ποιμήν poimén "shepherd"). The position of ψῦχος (psŷkhos "cold") nearby is due to the mislemmatization of some inflected forms of ψυχή (psyché "soul") under this lemma, as revealed by nearest neighbor analysis. To the left of this group, a small cluster of terms pertaining to Christian exegesis (ῥητός, παραβολή, διασαφέω diasaphéo "illustrate") can be recognized.

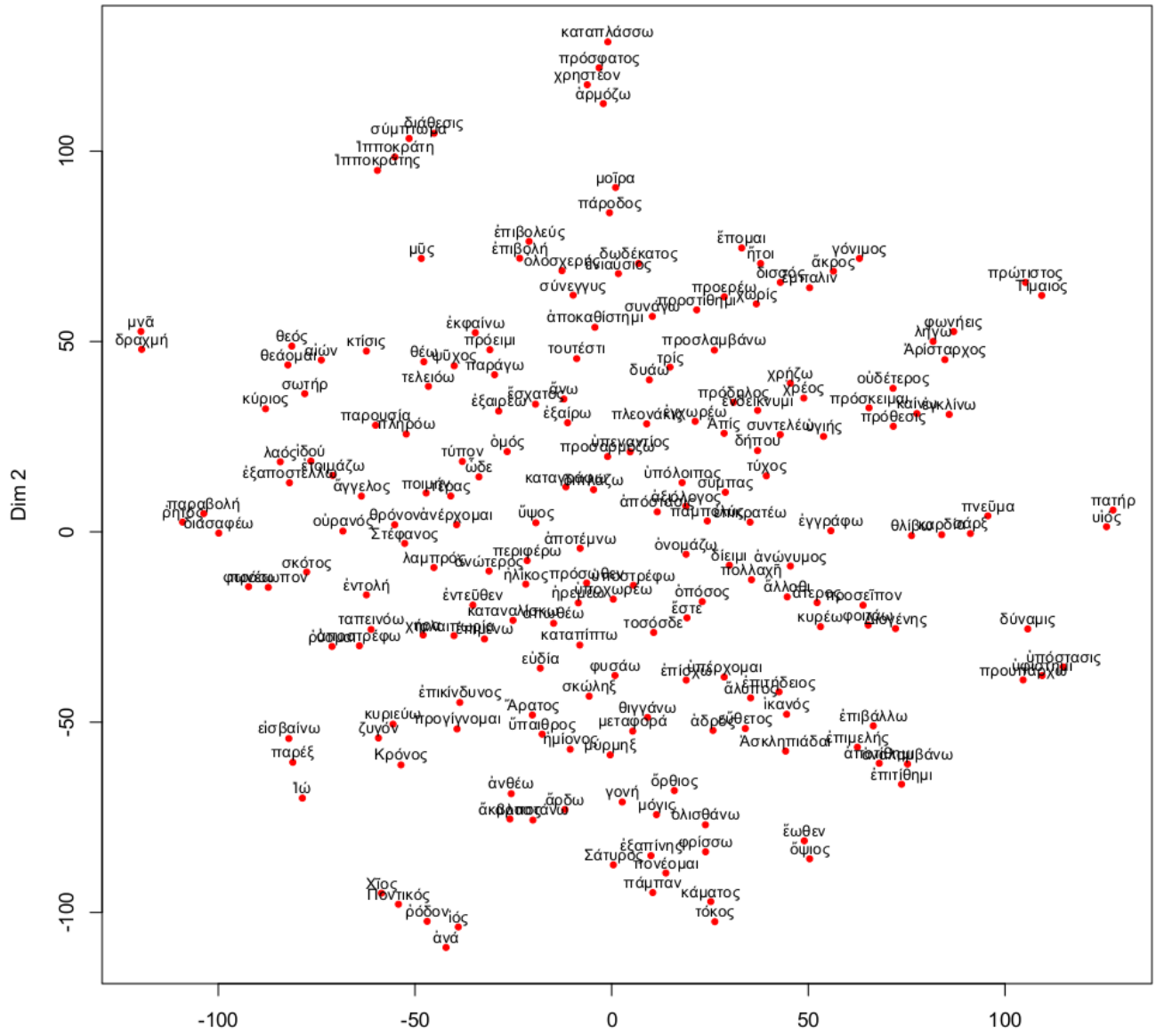


Figure 1. Relative positions within the AD-Space of the 200 words with the lowest correlation scores. Dimensionality reduction was performed using t-SNE (van der Maaten and Hinton, 2008).

The upper portion of the plot houses technical terms from the domains of medicine (the uppermost groups), astronomy and geometry, while philosophical terminology is found in the outer right area. Some smaller groups are also noticeable, such as μνᾶ (mnâ "mina") and δραχμή (drakhmé "drachma"), both units of currency, on the left, and πρώτιστος (prótistos "the very first") and Τίμαιος (the proper name Tímaios, Latin Timaeus), both connected to (Neo-)Platonic philosophy, on the right.

All in all, despite a certain amount of noise, the plot in Figure 1 supports the findings detailed so far. We can see how the main semantic changes in the Greek lexicon between the pre-Christian and Christian era affected the domains of religion (in a broader sense) and/or technical language. Within these domains, some more fine-grained relations between words that underwent significant semantic shifts can be observed.

4 Conclusion

This paper shows how Distributional Semantics can be used as an exploratory tool to detect semantic change. In this case study on Ancient Greek, the proposed method based on distributional RSA not only confirms the hypothesis that the diffusion of Christianity was a crucial cause of semantic change in the Greek lexicon, but also allows for the identification of unexpected patterns of evolution, such as the apparent specialization in the usage of technical terms. This last phenomenon could also be influenced by the fact

that the AD-corpus is richer in philosophical and technical treatises; however, a documented change in the proportion of different possible usages of a word is in itself a very informative result, especially in a field such as Classics, where the analysis of (literary) texts is paramount. Further research should undoubtedly highlight the effect of corpus composition. A focus on shorter periods of time might be of interest, since, for instance, the rise of technical prose writing is a characteristic of the Hellenistic Age (cf. e.g. Gutzwiller 2007, pages 154-167).

From a methodological standpoint, the fact that the results obtained from such a small corpus of purely literary texts are both meaningful and informative is of great relevance. Furthermore, the choice to adopt a data-driven approach proved fruitful, in that it brought to light directions of change that were not expected *a priori*. For traditional research in Classics, a computational approach to the lexicon of Ancient Greek is compelling because it provides new information about a language for which the judgments of native speakers are unavailable (cf. Perek, 2016). The results of this study show how Distributional Semantics can complement the assertions of the philologist, as well as help discover patterns of lexical change that would otherwise be impossible to grasp beyond an intuitive level.

References

Boschetti, Federico. 2009. A Corpus-based Approach to Philological Issues. PhD Thesis, University of Trento, Trento.

Crane, Gregory. 1991. Generating and parsing Classical Greek. *Literary and Linguistic Computing*, 6(4):243–245.

Dinu, Georgiana, Nghia The Pham and Marco Baroni. 2013. DISSECT – DIStributional SEmantics Composition Toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 31–36, Sofia.

Edelman, Shimon. 1998. Representation is representation of similarities. *Behavioral and Brain Sciences*, 21:449–467.

Evert, Stefan. 2008. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, pages 1212–1248, Berlin.

Gutzwiller, Kathryn J. 2007. *A guide to Hellenistic literature* (Blackwell guides to Classical literature). Blackwell Publishing, Oxford.

Kriegeskorte, Nikolaus and Roger A. Kievit. 2013. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412.

Kulkarni, Vivek, Rami Al-Rfou, Bryan Perozzi and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web* (WWW '15), pages 625–635, Firenze.

Van der Maaten, Laurens and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

O'Donnell, Matthew Brook. 2005. *Corpus Linguistics and the Greek of the New Testament* (New Testament Monographs, 6). Sheffield Phoenix Press, Sheffield.

Perek, Florent. 2016. Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics*, 54(1):149–188.

Sagi, Eyal, Stefan Kaufmann and Brady Clark. 2011. Tracing semantic change with Latent Semantic Analysis. In Kathryin Allan and Justyna A. Robinson, editors, *Current Methods in Historical Semantics*, pages 161–183, Boston, MA.

Wijaya, Derry Tanti and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 International Workshop on DETecting and Exploiting Cultural diversiTy on the Social Web* (DETECT '11), pages 35–40, Glasgow.

Xu, Yang and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (CogSci 2015), Pasadena, CA.

Reassessing inflectional regularity in Modern Greek conjugation

Stavros Bompolas
University of Patras, Greece
stavros.bompolas@gmail.com

Marcello Ferro
ILC-CNR Pisa, Italy
marcello.ferro@ilc.cnr.it

Claudia Marzi
ILC-CNR Pisa, Italy
claudia.marzi@ilc.cnr.it

Franco Alberto Cardillo
ILC-CNR Pisa, Italy
francoalberto.cardilllo@ilc.cnr.it

Vito Pirrelli
ILC-CNR Pisa, Italy
vito.pirrelli@ilc.cnr.it

Abstract

Paradigm-based approaches to word processing/learning assume that word forms are not acquired in isolation, but through associative relations linking members of the same word family (e.g. a paradigm, or a set of forms filling the same paradigm cell). Principles of correlative learning offer a set of dynamic equations that are key to modelling this complex dynamic at a considerable level of detail. We use these dynamic equations to simulate acquisition of Modern Greek conjugation, and we compare the results with evidence from German and Italian. Simulations show that different Greek verb classes are processed and acquired differentially, depending on their degrees of formal transparency and predictability. We relate these results to psycholinguistic evidence on Modern Greek word processing, and interpret our findings as supporting a view of the mental lexicon as an emergent integrative system.

Secondo l'approccio paradigmatico allo studio dell'elaborazione e dell'apprendimento lessicali, le parole di una lingua non sono acquisite in isolamento, ma attraverso legami associativi tra membri della stessa famiglia morfologica, la cui dinamica è modellata dalle equazioni dell'apprendimento correlativo. Il presente contributo offre una serie di esperimenti nei quali l'apprendimento del sistema verbale del greco moderno è simulato come un processo di auto-organizzazione dinamica di parole memorizzate in modo concorrente. I risultati mostrano chiari effetti di interazione dinamica tra trasparenza e regolarità morfologica nell'acquisizione di classi di forme del verbo greco.

1 Introduction

Issues of morphological (ir)regularity have traditionally been investigated through the prism of morphological competence, with particular emphasis on aspects of the internal structure of complex words (Bloomfield, 1933; Bloch, 1947; Chomsky and Halle, 1968; Lieber, 1980; Selkirk, 1984; among others). Within this framework, one of the most influential theoretical positions is that morphologically, phonologically, or/and semantically transparent words are always processed on-line through their constituent elements, whereas irregular, idiosyncratic (non-transparent) forms are stored (and retrieved) as wholes in the lexicon (Pinker and Prince, 1994). Likewise, Ullman and colleagues (1997) assume that the past tense formation of regular verbs in English requires on-line application of an affixation rule (e.g. *walk* > *walk+ed*), while irregular past tense forms, involving stem allomorphy (e.g. *drink* > *drank*), are retrieved from the lexicon.

Modern Greek introduces an interesting variation in this picture. First, stem allomorphy and suffixation are not necessarily mutually exclusive processes, but coexist in the same inflected forms (e.g. *lin-o* 'I untie' > *e-li-s-a* 'I untied', *aɣap(a)-o* 'I love' > *aɣapi-s-a* 'I loved'). Secondly, affixation rules may select unpredictable stem allomorphs: *aɣap(a)-o* 'I love' > *aɣapi-s-a* 'I loved', *for(a)-o* 'I wear' > *fore-s-a* 'I wore', *xal(a)-o* 'I demolish' > *xala-s-a* 'I demolished'. These cases suggest that inflectional (ir)regularity is not an all-or-nothing notion in Greek. Different inflectional processes may compound in the same words to provide a challenging word processing scenario (Tsapkini et al., 2004). From this perspective, Modern Greek offers the opportunity to test traditional hypotheses of grammar and lexicon interaction in word processing and learning, to explore the potential of single, distributed mechanisms in addressing word processing challenges (Alegre and Gordon, 1999; Baayen, 2007).

1.1 The evidence

Modern Greek conjugation is stem-based, each fully inflected verb form requiring obligatory suffixation of person, number and tense markers that attach to either a bare or a complex stem in both regular (*aɣap-o* 'I love' ~ *aɣapis-a* 'I loved') and irregular verbs (*pern-o* 'I take' ~ *pir-a* 'I took'). Unlike English speakers, Greek speakers must always resort to an inflectional process to understand or produce a fully inflected form, no matter how regular the form is (Terzi et al., 2005: 301).

Classifying a Greek verb as either regular or irregular thus requires observation of the stem formation processes on whose basis agreement and tense suffixes are selected. Accordingly, it is assumed that the presence or absence of the aspectual marker is a criterion for assessing the degree of regularity of a Greek verb. In particular, so-called "sigmatic" past-tense forms (e.g. *aɣap-o* ~ *aɣapis-a*) are traditionally considered to be regular, in that they involve a segmentable marker (*-s-*) combined with phonologically predictable or morphologically systematic stem-allomorphs. Asigmatic past-tense forms (e.g. *pern-o* ~ *pir-a*), in contrast, exhibit typical properties of irregular inflection, since they involve unsystematic stem allomorphs (in some cases suppletive stems), and no segmentable affixes marking perfective aspect. This distinction has also been supported by psycholinguistic evidence (Stamouli, 2000; Tsapkini et al., 2001, 2002a,b,c, 2004; Mastropavlou, 2006; Varlokosta et al., 2008; Stavrakaki and Clahsen, 2009a,b; Stathopoulou and Clahsen, 2010; Stavrakaki et al., 2012; Konstantinopoulou et al., 2013; among others), suggesting that sigmatic past-tense forms are typically produced on-line by rules, and asigmatic forms are stored and accessed from the mental lexicon.

However, careful analysis of the Greek verb system appears to question such a sharp processing-storage divide. In particular, Greek data provide the case of a mixed inflectional system where both stored allomorphy and rule-based affixation are simultaneously present in the formation of past tense forms. Ralli (2005) provided a classification of verb paradigms which is based on two criteria; firstly, the presence vs. absence of the sigmatic affix and, secondly, the presence vs. absence of (systematic) stem allomorphy. As a result, we can define the following three classes (see also Tsapkini et al., 2001, 2002a,b,c, 2004):

(i) an affix-based class, requiring the presence of the aspectual marker *-s-*, and including verbs with a predictable phonological stem-allomorph (e.g., *lin-o* 'I untie' ~ *e-li-s-a* 'I untied', *ɣraf-o* 'I write' ~ *e-ɣrap-s-a* 'I wrote');

(ii) a mixed class where active perfective past tense forms are produced by affixation of the aspectual marker *-s-* to a systematic morphological stem-allomorph (e.g., *mil-o* 'I speak' ~ *mili-s-a* 'I spoke');

(iii) an idiosyncratic verb class whose forms are based on non-systematic stem-allomorphy (requiring stem-internal alternation or suppletion) or no stem-allomorphy at all, and no (sigmatic) aspectual marker (e.g., *pern-o* 'I take' ~ *pir-a* 'I took', *tro-o* 'I eat' – *e-fag-a* 'I ate', *krin-o* 'I judge' – *e-krin-a* 'I judged').

It should be noted that, in regular Greek verbs, transparency/systematicity and predictability are not mutually implied. The morphologically-conditioned allomorphy of class-(ii) verbs requires a systematic pattern of perfective stem formation, namely $X_{(a)}$ ~ X + V (e.g. *aɣap(a)-* > *aɣapi-*), where 'X' is a variable standing for the bare stem, 'V' stands for a vowel, and the subscripted '(a)' indicates an optional 'a', forming a Modern Greek free variant of the imperfective stem (e.g. *aɣapo* ~ *aɣapao*, see Ralli, 2005, 2007). The variable *V* in the perfective stem can be instantiated as an *i*, *e* or *a*, and cannot be predicted from the bare stem. On the other hand, the phonologically-conditioned allomorphs of class-(i) verbs (e.g. *lin-* > *e-li-s-*) are the outcome of exception-less phonological rules, which nonetheless obfuscate a full formal correspondence (transparency) between the imperfective stem and the perfective stem.

Evidence from language acquisition and experimental psycholinguistics shows that perception of formal transparency between imperfective and perfective Greek stems plays a prominent role in human word processing strategies (Tsapkini et al., 2002c: 116, 2004: 616; Stavrakaki and Clahsen, 2009a: 117; Stathopoulou and Clahsen, 2010: 872). More specifically, lack of full formal nesting between imperfective and perfective stems (compare *aɣap-o* 'I love' ~ *aɣapi-s-a* 'I loved' *vs.* *ðulev-o* 'I work' ~ *ðulep-s-a* 'I worked') appears to have an extra processing cost (Tsapkini et al., 2002c: 116).

To sum up, analysis of Greek data offers evidence of graded levels of morphological regulari-

ty, based on the interaction between formal transparency (degrees of stem similarity) and (un)predictability of stem allomorphs. The evidence appears to question a dichotomous view of storage vs. rule-based processing mechanisms. In fact, no sharp distinction between affix processing and allomorph retrieval can account for the interaction of formal transparency and predictability in Greek word processing. On the one hand, rule-based mechanisms are called for to account for transparency effects of stem allomorphy on word processing. On the other hand, storage is required if the *same* allomorphs cannot be predicted. In the remainder of this paper, we test the hypothesis that this evidence is compatible with a parallel processing architecture (a Temporal Self-organising Map) where processing and storage are in fact mutually implied.

2 TSOMs

Temporal Self-organising Maps (TSOMs, Ferro et al., 2011; Marzi and Pirrelli, 2015; Pirrelli et al., 2015) are unsupervised artificial neural networks that learn to dynamically memorise input strings as chains of maximally-responding processing nodes (*Best Matching Units* or *BMU*s), whose level of sensitivity to input symbols in specific contexts is a continuous function of the distributional regularities of the input symbols during training. In a TSOM, each processing node has two layers of synaptic connectivity: an input layer, connecting the node to the current input stimulus (e.g. the letter of a written word), and a (re-entrant) temporal layer, connecting the node to all other nodes.

Given the *BMU* at time *t*, the temporal layer encodes the expectation of the current *BMU* for the node to be activated at time *t+1*. The strength of the connection between consecutively activated *BMU*s is trained through the following principles of correlative learning (compatible with Rescorla-Wagner (1972) equations):

Given the input bigram *ab*, the connection strength between *BMU* of *a* at time *t* and *BMU* of *b* at time *t+1* will

- increase if *a* often precedes *b* in training (entrenchment)
- decrease if *b* is often preceded by a symbol other than *a* (competition).

The interaction between entrenchment and competition in a TSOM accounts for important dynamic effects of self-organisation of stored words (Marzi et al., 2014, 2016). In particular, high-frequency words tend to recruit specialised (and stronger) chains of *BMU*s, while low-frequency words are responded to by more "blended" (and weaker) *BMU* chains. In what follows, we report how well a TSOM can accommodate the complexity of the Greek verb system, by controlling factors such as word frequency distribution, degrees of inflectional regularity and word length.

2.1 The experiment

To allow pairwise comparison with existing experimental evidence on German and Italian (Marzi et al., 2016), the Greek training dataset was designed to contain 50 top-ranked paradigms by cumulative token frequency, for a total of 750 verb forms, whose frequency distributions were sampled from the FREQcount section of the Greek SUBTLEX-GR corpus (BCBL, 2016). From each paradigm, 15 inflected forms were extracted: the full set of present indicative (6) and simple past tense (6) forms, and the singular forms of simple future (3). As we were mainly interested in effects of global paradigm-based organisation of active voice indicative forms, we excluded paradigms with systematic gaps, impersonal verbs, and deponent verbs. We included high-frequency paradigms with suppletive forms or/and non-systematic allomorphy (Ralli, 2007, 2014) as attested in the training set.

The dataset was administered to a 42x42 node map for 100 learning epochs. Word frequencies in the training data were a function of the real word frequency distribution in the reference corpus, fitted in the 1-1000 range. To control for random variability, we repeated the experiment 5 times.

For each repetition, we then assessed how well the map could acquire the 750 input forms, using the task of Word Recall as a probe. Word recall is defined as the process of retrieving a word form from its chain of *BMU*s. Successful recall is possible if inter-node connections on the temporal layer are finely tuned to the distribution of symbols in the training data. The more accurate the re-entrant temporal coding is, the easier for the map to retrieve the symbols of a word in their appropriate order. We make the further reasonable assumption that a word is acquired by a TSOM when the map is in a position to recall the word accurately and consistently from its *BMU* chain. Average recall accuracy at epoch 100 turned out to be considerably high: 99.6 % (std = 0.1%).

3 Data analysis

Results were analysed using Linear Mixed Effects (LME) models with experiment repetitions and training items as random variables.

Figure 1 shows the marginal plot of the interaction between word length and regular vs. irregular verb classes for German, Italian and Greek, using an LME model fitting word learning epochs, with (log) word frequency, inflectional class and word length as fixed effects. In German and Italian, the distinction between regular and irregular paradigms is based on the criterion of absence vs. presence of stem allomorphy across all forms of a paradigm (Marzi et al., 2016). In Greek, we consider regular all paradigms showing a sigmatic perfective stem, and irregular those with an asigmatic perfective stem.

Unlike German and Italian (Figure 1, top and middle panels), where irregulars tend to be acquired systematically later than length-matched regulars are, and no significant interaction is found, Greek data (Figure 1, bottom panel) show an interesting crossing pattern: shorter irregulars are acquired earlier than length-matched regulars of comparable frequency, but long irregulars are acquired later than long regulars.

Marzi and colleagues (2016) account for earlier learning epochs of both German and Italian regulars as an effect of stem transparency on cumulative input frequencies. With German and Italian regular verbs, stems are shown to the map consistently more often, since they are transparently nested in all forms of their own paradigm. This makes their acquisition quicker, due to specialised chains of stem-sensitive *BMUs* getting more quickly entrenched. Once a stem is acquired, it can easily be combined with a common pool of inflectional endings for tense and agreement, simulating an effect of (nearly) instantaneous (or paradigm-based, as opposed to item-based) acquisition. In contrast, Greek verb classes always present stem allomorphy throughout their paradigms, no matter whether allomorphy is systematic, phonologically motivated or unsystematic. In regular verbs, where perfective stem formation requires *-s-* affixation, perfective stems are systematically longer than their imperfective counterparts, and are acquired after them. Nonetheless, since imperfective stems are redundantly embedded in perfective stems, learning a long regular perfective form is easier (i.e. it takes a comparatively shorter time) than learning an irregular perfective form of comparable length. This is, again, a regularity-by-transparency effect, and explains why long regular forms tend to be acquired (on average) more easily than long irregular forms.

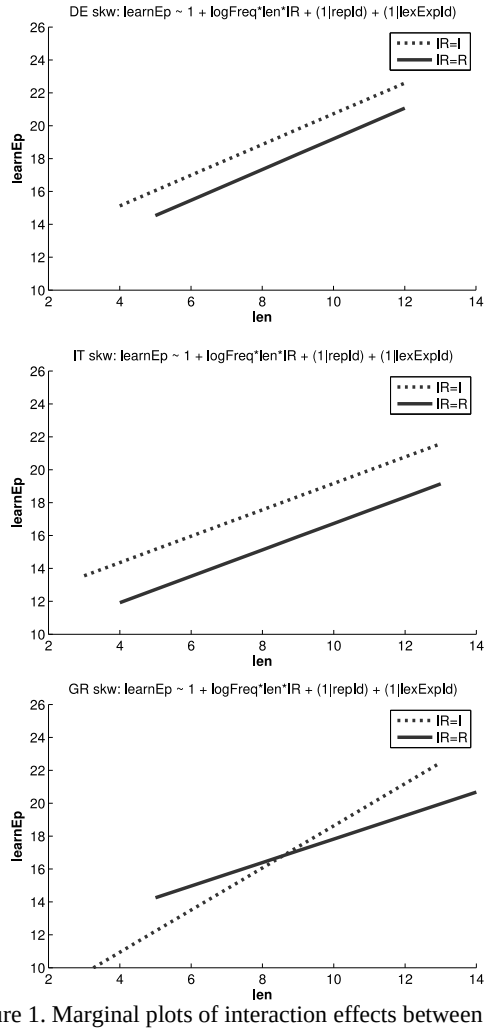


Figure 1. Marginal plots of interaction effects between word length and inflectional regularity in an LME model fitting word learning epochs in German (top), Italian (middle) and Greek (bottom). Solid lines = regulars, dotted lines = irregulars.

To further investigate the impact of degrees of formal transparency on the processing of Greek verb forms, we conducted an LME analysis of the interaction between word length and classes of (ir)regularity in word recall (Figure 2). When we control for length, regular verbs with systematic morphological allomorphs (e.g. *aɣap(a)-o ~ aɣapi-s-a*, solid line in the plot) are recalled more easily than regular verbs with phonological allomorphs (e.g. *ðulev-o ~ ðulep-s-a*, dashed line in the plot). Notably, both classes are easier to recall than asigmatic (irregular) verbs (dotted line in the plot), which show, in most cases, formally more opaque allomorphs (e.g. *pern-o ~ pir-a*). As shown by the difference in slope between the solid line and the other two lines of Figure 2,

facilitation increases with word length, supporting our interpretation of the crossing pattern in the bottom panel of Figure 1.

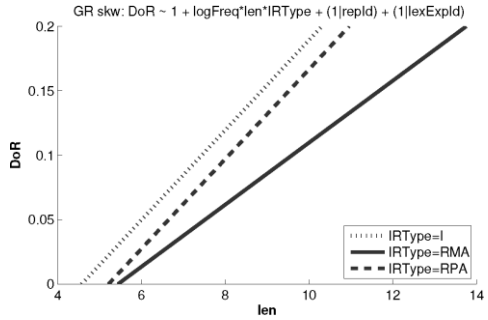


Figure 2. Marginal plot of interaction effects between length (x axis), and degrees of stem regularity in an LME model fitting Difficulty of Recall (y axis) by TSOMs trained on Greek verb forms

4 Conclusions

Data analysis highlighted a hierarchy of regularity-by-transparency effects that appear to have consequences on morphological processing. In particular, the evidence offered here emphasises the role of formal preservation of the stem (or stem transparency) in the paradigm as a key facilitation factor for morphological processing.

Our case study is focused on a distinguishing characteristic of Greek conjugation: all verb paradigms, both regulars and irregulars, involve (unpredictable) stem allomorphy in past-tense formation. Hence, the difference between regular and irregular verbs could not be attributed to the presence or absence of stem allomorphy as is the case with other languages, such as English and Italian (and, to a lesser extent, German), but rather to the type of stem allomorphy itself. The finding (Figure 2) that stem-final systematic change, as in the case of regulars, led to significantly easier recall than stem-internal vowel changes and non-systematic/non-predictable stem-final change, as is the case of irregulars, lends support to the conclusion that the type of stem allomorphy is what determines the different levels of morphological regularity in Greek. Crucially, this seems to involve a regularity-by-transparency interaction, with predictability playing second fiddle.

Our data meet growing psycholinguistic evidence pointing in the same direction, to emphasise the importance of formal redundancy for speakers' perception of morphological structure. Furthermore, it paves the way to the definition of a performance-oriented notion of inflectional regularity that may ultimately cut across traditional competence-based classifications, which presuppose a hardly tenable subdivision of work between storage and processing.

References

Alegre, M. and P. Gordon. 1999. Frequency effects, and the representational status of regular inflections. *Journal of Memory, and Language*, 40: 41-61.

Baayen, H. R. 2007. Storage, and computation in the mental lexicon. In: G. Jarema and G. Libben (eds.), *The Mental Lexicon: Core Perspectives*. Amsterdam: Elsevier, 81-104.

BCBL. 2016. *SUBTLEX-GR: The corpus*. Donostia: Basque Center on Cognition, Brain, and Language. [Retrieved from: http://www.bcbl.eu/subtlex-gr/, 12-04-2016].

Bloch, B. 1947. English verb inflection. *Language*, 23: 399-418.

Bloomfield, L. 1933. *Language*. New York: Henry Holt, and Co.

Chomsky, N. and M. Halle. 1968. *The sound pattern of English*. New York: Harper and Row.

Ferro, M., Marzi, C. and V. Pirrelli. 2011. A Self-Organizing model of word storage and processing: implications for morphology learning. *Lingue e Linguaggio*, X(2): 209-226.

Konstantinopoulou, P., Stavrakaki, S., Manouilidou, C. and D. Zafeiriou. 2013. Past tense in children with focal brain lesions. *Stem-, Spraak- en Taalpathologie*, 18(1): 90-94.

Lieber, R. 1980. *On the organization of the lexicon*. PhD thesis. Cambridge: MIT.

Marzi, C., Ferro, M. and V. Pirrelli. 2014. Morphological structure through lexical parsability. *Lingue e Linguaggio*, XIII(2): 263-290.

Marzi, C. and V. Pirrelli. 2015. A neuro-computational approach to understanding the Mental Lexicon. *Journal of Cognitive Science*, 16(4): 493-534.

Marzi, C., Ferro, M., Cardillo, F. A. and V. Pirrelli. 2016. Effects of frequency, and regularity in an integrative model of word storage, and processing. *Italian Journal of Linguistics*, 28(1): 79-114.

Mastropavlou, M. 2006. *The effect of phonological saliency, and LF-interpretability in the grammar of Greek normally developing, and language impaired children*. Ph.D. thesis. Thessaloniki: Aristotle University of Thessaloniki.

Pinker, S. and A. Prince. 1994. Regular, and irregular morphology, and the psychological status of rules of grammar. In: S. D. Lima, R. L. Corrigan and G. K. Iverson (eds.), *The reality of linguistic rules*. Amsterdam: Benjamins, 321-351.

Pirrelli, V., Ferro, M. and C. Marzi. 2015. Computational complexity of abstractive morphology, In: M. Baerman, D. Brown and G. Corbet (eds.), *Understanding and Measuring Morphological Complexity*. Oxford: Oxford University Press, 141-166.

Ralli, A. 2005. *Morfologia* [Morphology]. Athens: Patakis.

Ralli, A. 2007. On the role of allomorphy in inflectional morphology: evidence from dialectal variation. In: G. Sica (ed.), *Open problems in linguistics, and lexicography*. Milano: Polimetrica, 123-152.

Ralli, A. 2014. Suppletion. In: G. K. Giannakis (ed.), *Encyclopedia of Ancient Greek language, and linguistics*, Vol. 3. Leiden, Boston: Brill, 341-344.

Selkirk, E. 1984. *Phonology, and Syntax*. The MIT Press.

Stamouli, S. 2000. Simfonia, xronos ke opsi stin eliniki idiki glosiki diataraxi [Agreement, tense, and aspect in specific language impairment in Greek]. In: *Proceedings of the 8th symposium of the Panhellenic Association of Logopedists*. Athens: Ellinika Grammata.

Stathopoulou, N. and H. Clahsen. 2010. The perfective past tense in Greek adolescents with Down Syndrome. *Clinical Linguistics & Phonetics*, 24(11): 870-882.

Stavrakaki, S. and H. Clahsen. 2009a. The perfective past tense in Greek child language. *Journal of Child Language*, 36: 113-142.

Stavrakaki, S. and H. Clahsen. 2009b. Inflection in Williams Syndrome. The perfective past tense in Greek. *The Mental Lexicon*, 4: 215-238.

Stavrakaki, S., Koutsandreas, K. and H. Clahsen. 2012. The perfective past tense in Greek children with specific language impairment. *Morphology*, 22: 143-171.

Terzi, A., Papapetropoulos, S. and E. D. Kouvelas. 2005. Past tense formation, and comprehension of passive sentences in Parkinson's disease: Evidence from Greek. *Brain, and Language*, 94: 297-303.

Tsapkini, K., Jarema, G. and E. Kehayia. 2001. Manifestations of morphological impairments in Greek aphasia: A case study. *Journal of Neurolinguistics*, 14: 281-296.

Tsapkini, K., Jarema, G. and E. Kehayia. 2002a. A morphological processing deficit in verbs but not in nouns: A case study in a highly inflected language. *Journal of Neurolinguistics*, 15: 265-288.

Tsapkini, K., Jarema, G. and E. Kehayia. 2002b. The role of verbal morphology in aphasia during lexical access: Evidence from Greek. In: E. Fava (ed.), *Clinical linguistics, and phonetics: Theory, and applications in speech pathology, and therapy*. Amsterdam/Philadephia: John Benjamins, 315-335.

Tsapkini, K., Jarema, G. and E. Kehayia. 2002c. Regularity revisited: Evidence from lexical access of verbs, and nouns in Greek. *Brain, and Language*, 81: 103-119.

Tsapkini, K., Jarema, G. and E. Kehayia. 2004. Regularity re-revisited: Modality matters. *Brain, and Language*, 89: 611-616.

Ullman, M., Corkin, S., Coppola, M., Hickok, G., Growdon, J. H., Koroshetz, W. and S. Pinker. 1997. A neural dissociation within language: evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience*, 9(2): 266-276.

Varlokosta, S., Arhonti, A., Thomaidis, L. and V. Joffe. 2008. Past tense formation in Williams syndrome: evidence from Greek. In: A. Gavarro and M. Freitas (eds.), *Proceedings of GALA 2007*. Cambridge: Cambridge Scholars Publishing, 483-491.

Rescorla R. and A. Wagner. 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In: A. H. Black and W. F. Prokasy (eds.), *Classical conditioning ii*. New York: Appleton-Century-Crofts, 64-99.

Written word production and lexical self-organisation: evidence from English (pseudo)compounds

Marcello Ferro
ILC-CNR Pisa, Italy
marcello.ferro@ilc.crn.it

Franco Alberto Cardillo
ILC-CNR Pisa, Italy
francoalberto.cardillo@ilc.cnr.it

Vito Pirrelli
ILC-CNR Pisa, Italy
vito.pirrelli@ilc.cnr.it

Christina L. Gagné
Dept. of Psychology, University of Alberta, Canada
cgagne@ualberta.ca

Thomas L. Spalding
Dept. of Psychology, University of Alberta, Canada
spalding@ualberta.ca

Abstract

Elevation in typing latency for the initial letter of the second constituent of an English compound, relative to the latency for the final letter of the first constituent of the same compound, provides evidence that implementation of a motor plan for written compound production involves smaller constituents, in both semantically transparent and semantically opaque compounds. We investigate here the implications of this evidence for algorithmic models of lexical organisation, to show that effects of differential perception of the internal structure of compounds and pseudo-compounds can also be simulated as peripheral stages of lexical access by a self-organising connectionist architecture, even in the absence of morpho-semantic information. This complementary evidence supports a maximization-of-opportunity approach to lexical modelling, accounting for the integration of effects of pre-lexical and lexical access.

Il rallentamento nel tempo di battitura del primo carattere del secondo costituente di un composto inglese, rispetto al tempo dell'ultimo carattere del primo costituente, dimostra che l'implementazione del programma motorio per la scrittura di un composto è influenzata dai costituenti del composto stesso, siano essi semanticamente trasparenti o opachi. Il presente contributo offre un modello computazionale di questa evidenza, e ne valuta l'impatto sull'organizzazione del lessico mentale: la percezione del confine di morfema tra i due costituenti è analizzata come il risultato dell'interazione dinamica tra processi di accesso pre- e post-lessicale.

1 The evidence

A key question concerning the representation and processing of compound words has focused on whether (and, if so, how) morphological structure plays a role. The bulk of the research on this issue has come from recognition or comprehension tasks such as lexical decision or reading. However, written production provides a useful counterpart and allows researchers to examine whether morphemes are used even after a word has been accessed. One advantage of a typing task (in which the time to type each letter of a word is recorded) is that researchers can examine differences in processing difficulty at various points in the word. Previous research found an elevation in typing latency for the initial letter of the second constituent relative to the latency for the final letter of the first constituent for English (Gagné & Spalding 2014; Libben et al. 2012; Libben & Weber 2014) and German compounds (Sahel et al. 2008; Will et al. 2006). This elevation in typing latency at the morpheme boundary suggests that the system plans the output morpheme by morpheme, rather than as a whole unit, and that morphological programming is not complete when the motor system begins the output of the word (Kandel et al. 2008).

Gagné and Spalding (2016) examined the role of morphemic structure and semantic transparency on typing latency. The stimuli consisted in 200 compounds, 50 pseudo-compounds, and 250 monomorphemic words matched pairwise with the compounds and pseudo-compounds in the number of syllables and letters. The pseudo-compounds contain two words that do not function as morphemes (e.g., *carpet* contains *car* and *pet*). The compounds varied in whether the first and second constituent were semantically transparent. The items were displayed individually using a progressive demasking procedure and participants typed the word as the computer recorded the time required to type each letter.

The time to initiate the first letter was equivalent for monomorphemic and compound words. Typing times got faster across the word for both word types, but the rate of change was faster for monomorphemic words than for compound words. This difference was not observed when comparing monomorphemic words and pseudo-compounds.

For compounds, the rate of speed-up was slower when the first constituent was transparent than when it was opaque, but was unaffected by the transparency of the second constituent. The elevation in typing latency at the morpheme boundary was larger when the first constituent was transparent than when it was opaque, but was unaffected by the transparency of the second constituent. This difference is due to the final letter of the first constituent when the first constituent requiring less time to type when it was transparent than when it was opaque.

The data for the pseudo-compounds indicated that embedded morphemes influence production, even when they do not function as morphemes. Typing latency increased one letter prior to the end of the first constituent of a pseudo-compound and remained elevated through the boundary (e.g., both *r* and *c* in *scarcity* were elevated relative to the *a*).

1.1 Implications for lexical architectures

The reported evidence clearly indicates that morphemic structure is involved in written word production. The production of compounds differs from that of monomorphemic words and the semantic transparency of the two constituents leads to different effects. Furthermore, embedded pseudo-morphemes appear to influence the production of pseudo-compounds, but not in the same way that the embedded morphemes affect the production of compounds.

This appears to lend only partial support to models of lexical architecture where both compounds and their constituents are represented and processed as independent access units (Figure 1). In panel *A*, following Taft & Forster (1975), access and output of compounds are mediated by their constituents (Cs), but extra procedures would be needed to account for the role of semantic transparency in modulating the size of elevation in typing latency at the morpheme boundary. A supralexical account (panel *B*: Giraudo & Grainger 2000, Grainger et al. 1991), where constituents are activated upon compositional interpretation of compounds, cannot capture the persistence of typing effects in semantically opaque compounds (and, to an extent, in pseudo-compounds). Race models (panel *C*: Schreuder & Baayen 1995) posit parallel pathways for compound processing (both holistic and compositional), depending on variables such as whole word vs. constituent frequency, but it is not clear how they can account for effects of interaction between the two paths. Connectionist models (panel *D*: Rumelhart & McClelland 1986, Plaut & Gonnerman 2000), on the other hand, tend to dispense with specialized representational levels and access procedures, and make room for distributed effects of sublexical co-activation through overlaying patterns of processing units. A defining feature of these models is that they blur the traditional distinction between representations and processing units. We suggest that blurring this distinction can go a long way in addressing some of the issues that appear to elude models *A*, *B* and *C*.

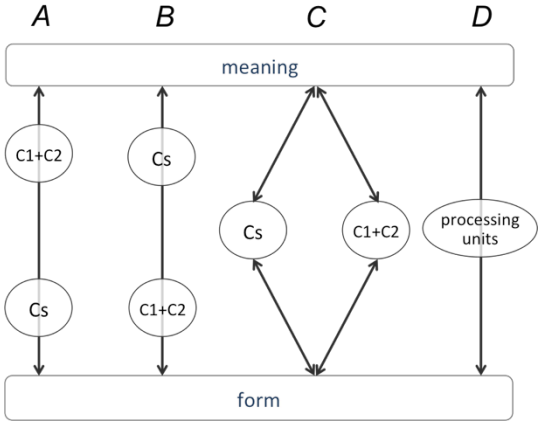


Figure 1 – Four architectures of form-meaning mapping in the mental lexicon: *C1+C2* designates two-word compounds and *Cs* mono-morphemic constituents (adapted from Diependaele et al. 2012)

Temporal Self-Organising Maps (TSOMs: Ferro et al. 2011; Marzi et al. 2014; Pirrelli et al. 2015), are a time-sensitive variant of Kohonen's SOMs (Kohonen, 2002), where words are stored through routinized, time-bound patterns of repeatedly successful processing units. Since all input words are stored concurrently on the same layer of fully connected nodes, TSOMs account for effects of co-activation of competing representations in terms of a continuous function of distributional regularities in the input data. In what follows, starting from Gagné & Spalding's evidence, we will focus on peripheral stages of lexical access/output, to verify if mechanisms of parallel, distributed pattern activation can account for differential processing effects between compounds and pseudo-compounds even in the

absence of morpho-semantic information. Although computational testing is carried out on TSOMs only, our discussion and concluding remarks address issues that go beyond a specific computational framework.

2 TSOMs

A TSOM consists of a grid of memory nodes with two layers of connectivity. The first layer (or I-layer) fully connects each node to the input vector, where symbols are sampled at discrete time ticks as patterns of activation ranging in the [0, 1] interval. Weights on the I-layer are adjusted in training for individual nodes to develop specialised sensitivity to particular input symbols. Each node is also connected to all other nodes through a layer of re-entrant connections (or T-layer), whose weight strength determines the amount of influence that activation of one node has on other nodes at a one-tick delay.

When an input symbol is presented at time *t*, the level of activation $y_i(t)$ of node *i* is a function of: (a) the node's sensitivity to the current input symbol ($y_{I\_layer,i}(t)$), and (b) the re-entrant support the node receives from the map activation state at *t-1* ($y_{T\_layer,i}(t) = f(y_j(t-1))$, where *f* is a linear function and *j* ranges over all map nodes). More formally:

$$y_i(t) = \alpha \cdot y_{I\_layer,i}(t) + (1-\alpha) \cdot y_{T\_layer,i}(t)$$

The node responding most strongly to the input symbol *S* at time tick *t* is called Best Matching Unit (hereafter *BMU*(*S*, *t*) or *BMU*(*t*) for short).

The map's response to a sequence of input symbols like *carpet* is a chain of consecutively firing *BMU*s, each responding to a letter in *carpet*. During training, connection weights between consecutive *BMU*s are adjusted to the frequency distribution of input symbols in the training set, according to Hebbian principles of correlative learning. Given the bigram *ab*, the connection strength between *BMU*(*a*, *t*-1) and *BMU*(*b*, *t*) increases if *a* often precedes *b* (entrenchment) and decreases if *b* is often preceded by a symbol other than *a* (competition) (Figure 2, left). Combination of entrenchment and competition yields selective specialisation of chains of *BMU*s (Figure 2, right). If the same input symbol follows different contexts, it will tend to be responded to by more *BMU*s, one for each context. The stronger the probabilistic support that the input symbol receives from its preceding context, the more likely the recruitment of a dedicated *BMU*, and the stronger its re-entrant connection. As a result of this dynamic, high-frequency words recruit specialised node chains, low-frequency words are responded to by weaker, "blended" node chains.

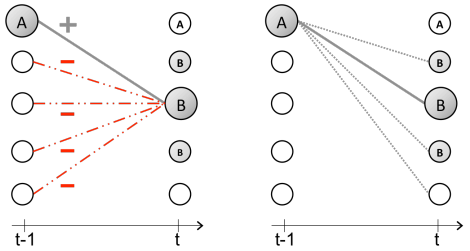


Figure 2 - Left: operation of Hebbian rules on potentiated ('+') and inhibited ('-') connections. Right: forward one-tick-delay connections leaving 'A' at time *t*-1. Larger nodes represent *BMU*s. Shades of grey indicate levels of node activation.

2.1 The Experiment

The 200 compounds and 50 pseudo-compounds used by Gagné & Spalding were used to train a 40x40 node TSOM for 100 learning epochs. Besides compounds and pseudo-compounds, the training set included 500 (pseudo)constituents as individual words (e.g. *car* and *wash* in *carwash*, *car* and *pet* in *carpet*), for a total amount of 750 items. At each training epoch, monomorphemic words were shown 10 times as often as compounds. We ran 5 repetitions of the experiment, and results were analysed using linear mixed effects models (LME), with experiment repetitions and training items as random variables.

To analyse differential processing effects for pseudo-compounds and compounds, we focused on two types of evidence: (i) per-letter performance of a trained TSOMs in incrementally anticipating compounds and pseudo-compounds; (ii) structural connectivity of BMUs responding to letter bigrams at the C1-C2 boundary.

To anticipate a progressively presented input word, a TSOM propagates the activation of the current *BMU*(*t*) through its forward temporal connections, and outputs, at each time tick, the symbol $S_{BMU(t+1)}$ encoded on the *I*_layer of the most strongly (pre)activated node:

$$BMU(t+1) = \underset{i=1,\ldots,N}{\operatorname{argmax}}\{m_{i,h}\} \qquad h = BMU(t)$$

where $m_{i,h}$ is the weight value on the forward temporal connection from node *h* to node *i*. Each correctly predicted symbol in the input word is assigned the prediction score of the preceding symbol incremented by 1. Otherwise, the symbol receives a 0-point score.

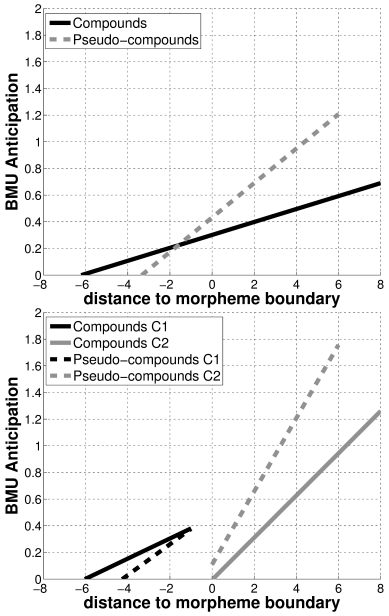


Figure 3 – Marginal plots of interaction effects between compounds vs. pseudo-compounds and letter distance to morpheme boundary in an LME model fitting anticipation of up-coming *BMU*s by a TSOM. Negative and positive *x* values indicate letter positions located, respectively, in the first and second constituent. Anticipation is plotted across whole (pseudo)compounds (top panel), and by individual constituents (bottom panel).

Figure 3 (top panel) illustrates the rate of letter anticipation across the word for both compounds and pseudo-compounds, plotted by distance to the morpheme boundary. The steeper rate for pseudo-compounds than for compounds shows that pseudo-compounds are easier to predict/anticipate than compounds. We take this evidence to be in line with evidence of a faster speedup rate in the typing of monomorphemic vs. compound words. A closer look at anticipation rates for individual constituents (Figure 3 bottom panel) shows a drop of anticipation at the C1-C2 boundary (more prominent for compounds than pseudo-compounds) with a steeper increase in C1 and C2 for pseudo-compounds, which happen to be, on average, shorter than C1 and C2 in real compounds.

To look for structural correlates of anticipation rates in the map, we conducted, for each item, a letter-by-letter analysis of values of pointwise entropy (PWH) for the connections between consecutive *BMU*s, namely *h=BMU(t-1)* and *i=BMU(t)*:

$$PWH(m_{i,h}) = -log\frac{m_{i,h}}{\sum_k m_{k,h}}$$

The value of PWH for the connection between end-C1 and start-C2 ($x = 0$) has a local peak in compounds only (Figure 4). Since PWH provides a measure of how unexpected the activation of *BMU*(*t*) is, this structural evidence can account for a delay in processing and a drop in anticipation at the morpheme boundary of compounds, but not of pseudo-compounds.

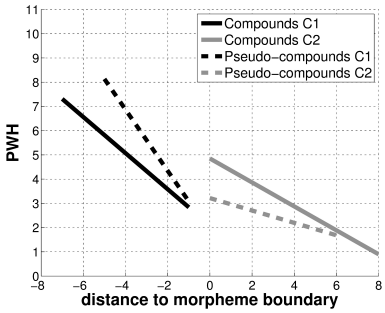


Figure 4 – Marginal plots of interaction effects between compound vs. pseudo-compound constituents and letter distance to morpheme boundary in an LME model fitting pointwise entropy of forward BMU connections. Negative and positive *x* values indicate letter positions located, respectively, in the first and second constituent.

3 Discussion and conclusions

Trained on both compounds and pseudo-compounds, TSOMs develop a growing sensitivity to surface distributional properties of input data, turning chains of randomly connected, general-purpose nodes into specialised sub-chains of *BMU*s that respond to specific letter strings at specific positions. Compounds not only tend to occur, on average, less frequently than their C1/C2 constituents do as independent words (Ji et al. 2011), but they tend to present lower-frequency bigrams at the C1-C2 boundary than do pseudo-compounds. Principles of Hebbian learning allow TSOMs to capitalise on both effects. Entrenchment makes expectations for high-frequency bigrams stronger and expectations for low-frequency bigrams weaker. At the same time, the competition between C1 as an independent word and C1 as the first constituent in a C1-C2 compound biases the map's expectation towards the most frequent event (C1 in isolation). Compound families, i.e. sets of compounds sharing C1 (*windmill*, *windshield* etc.) or C2 (*snowball*, *basketball* etc.), magnify these effects, making the map more sensitive to formal discontinuity at morpheme boundaries. When more C2s can follow the same C1 in complementary distribution, the left-to-right expectation for a particular C2 to occur, given C1, decreases. Likewise, when more C1s competitively select the same C2, the individual contribution of each C1 to the prediction of C2 decreases. We conjecture that more global effects of lexical organisation like these may eventually blur local memory

effects based on position-independent bigram frequencies.

Our simulations with TSOMs can model the correlation between continuously varying distributional regularities in the input data and peripheral levels of routinized recognition and production patterns. These patterns are in line with Gagné & Spalding's evidence of (a) the influence of embedded pseudo-morphemes on cascaded models of written word production, and (b) faster anticipation rates for monomorphemic vs. compound words.

Further experimental results (not reported here), obtained by including compound families in the training data, confirm slower anticipation rates for true compound constituents, due to the combined effect of word frequency distributions and word compositionality in compound families. The size of a compound family can arguably be a function of the degree of productivity and semantic transparency of its members (Baroni et al. 2007). The influence of the compound family size on anticipation rates can shed light on the influence of levels of semantic transparency on compound processing. Simulation evidence suggests that the bigger the family, the stronger its influence will be. Finally, we also monitored the influence of increasing token frequencies of monomorphemic words in the training data on the map perception of constituent boundaries within compounds. As expected, for constant frequency values of compounds in the training set, the higher the token frequency of monomorphemic words, the higher the pointwise entropy of connections at the C1-C2 boundary.

A full account of Gagné & Spalding's evidence of a graded influence of semantic transparency on compound processing is beyond the reach of the computational architecture presented here. Surface effects of discontinuity in the internal structure of compounds (as opposed to pseudo-compounds) appear to provide a purely formal, pre-lexical scaffolding for truly morpho-semantic effects to emerge at later processing stages. To model these effects, we appear to be in need of a parallel processing architecture able to effectively integrate several representational levels (orthographic, phonological, morphological, and conceptual) and different processing steps within a single distributed system (Smolka et al. 2009). Nonetheless, our simulations show that by letting compounds, pseudo-compounds and (pseudo)constituents compete for the same level of memory resources on a topological map, it is possible to account for apparently contradictory effects of a) graded perception of constituent boundary in both compounds and pseudo-compounds, apparently requiring prelexical decomposition, and b) higher anticipation rates for pseudo-compounds than compounds, supporting full form representations for lexical access.

References

Baroni M., Guevara E., & Pirrelli V. (2007) NN Compounds in Italian: Modelling Category Induction and Analogical Extension, in V. Pirrelli (Ed.) *Psycho-computational issues in morphology learning and processing*. *Lingue e Linguaggio VI* (2), 263 - 290.

Diependaele, K., Grainger, J., & Sandra, D. (2012). Derivational Morphology and Skilled Reading: An Empirical Overview. In M. J. Spivey, J. McRae, & M. F. Joanisse (Eds.) *The Cambridge Handbook of Psycholinguistics*, 311-332. Cambridge University Press.

Ferro M., Marzi, C., & Pirrelli, V. (2011). A Self-Organizing Model of Word Storage and Processing: Implications for Morphology Learning. *Lingue e Linguaggio X (2)*, 209 - 226.

Gagné, C. L., & Spalding, T. L. (2014). Typing time as an index of morphological and semantic effects during English compound processing. *Lingue e Linguaggio*, *XIII*(2), 241-262.

Gagné, C. L., & Spalding, T. L. (2016). Effects of morphology and semantic transparency on typing latencies in English compound and pseudo-compound words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42,* 1489-1495.

Giraudo, H., & Grainger, J. (2000). Effects of prime word frequency and cumulative root frequency in masked morphological priming. *Language and Cognitive Processes*, 15(4/5), 421–44.

Grainger, J., Colé, P., & Segui, J. (1991). Masked morphological priming in visual word recognition. *Journal of Memory and Language*, 30, 370–84.

Kandel, S., Álvarez, C. J., & Vallée, N. (2008). Morphemes also serve as processing units in handwriting production. In M. Baciu (Ed.), *Neuropsychology and cognition of language. Behavioural, neuropsychological and neuroimaging studies of spoken and written language*, 87–100. Kerala, India: Research Signpost.

Libben, G. (2010). Compound words, semantic transparency, and morphological transcendence. *Linguistische Berichte*, *Sonderheft 17*, 317-330.

Libben, G., & Weber, S. (2014). Semantic transparency, compounding, and the nature of independent variables. In F. Rainer, W. Dressler, F. Gardani, &

H. C. Luschutzky (Eds.), *Morphology and meaning*. Amsterdam: Benjamins.

Libben, G., Weber, S., & Miwa, K. (2012). P3: A technique for the study of perception, production, and participant properties. *The Mental Lexicon*, *7*(2), 237-248. doi:10.1075/ml.7.2.05lib

Marzi, C., Ferro, M., & Pirrelli, V. (2014) Morphological structure through lexical parsability. *Lingue e Linguaggio XIII (2)*, 263-290.

Pirrelli, V., Ferro, M., & Marzi, C. (2015). Computational complexity of abstractive morphology, In Baerman M., Brown D., Corbett G. (Eds.) *Understanding and Measuring Morphological Complexity*. 141-166, Oxford, United Kingdom: Oxford University Press.

Plaut, D. C., & Gonnerman, L. M. (2000). Are nonsemantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, 15(4/5), 445–85.

Sahel, S., Nottbusch, G., Grimm, A., & Weingarten, R. (2008). Written production of German compounds: Effects of lexical frequency and semantic transparency. *Written Language & Literacy*, *11*(2), 211-227. doi:10.1075/wll.11.2.06sah

Schreuder, R., & Baayen, H.R. (1995) Modeling morphological processing. In Feldman, L. B. (Ed.) *Morphological aspects of language processing*, 131–56, Hillsdale, NJ: Erlbaum.

Smolka, E., Komlosi, S., & Rösler, F. (2009). When semantics means less than morphology: The processing of German prefixed verbs. *Language and Cognitive Processes*, *24*(3), 337-375.

Rumelhart, D., & McClelland, J. 1986. On learning the past tense of English verbs. In Rumelhart, D.E, McClelland J. (Eds.), *Parallel distributed processing: Explanations in the microstructure of cognition*, 216-271, The MIT Press.

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, 14, 638–47.

Will, U., Nottbusch, G., & Weingarten, R. (2006). Linguistic units in word typing: Effects of word presentation modes and typing delay. *Written Language & Literacy*, *9*(1), 153-176.

Lexical categories or frequency effects? A feedback from quantitative methods applied to psycholinguistic models in two studies on Italian.

Francesca Franzon[*], Giorgio Arcara[°], Chiara Zanini[*]
[*]Dipartimento di Neuroscienze, Università degli Studi di Padova
[°]IRCSS Ospedale san Camillo, Lido di Venezia
{francescafranzon7;giorgio.arcara}@gmail.com;
chiara.zanini.2@unipd.it

Abstract

English. We examined two issues concerning Italian Number morphology: the phenomena related to mass and count nouns and to the plural dominance. By taking into account quantitative data from corpora and subjective frequency ratings in three mixed effect models, we found that differences in participants' performance in two lexical decision tasks could be better captured as differences in frequency rather than in terms of effects of lexical categories.

Italiano. *In questo studio sono stati posti a confronto due fenomeni pertinenti alla morfologia nominale di Numero in italiano: la contabilità dei nomi e la dominanza plurale. Integrando i dati quantitativi provenienti dai corpora e da due studi di rating in un'analisi statistica condotta tramite modelli a effetti misti, risulta che le differenze nella prestazione dei partecipanti in due studi di decisione lessicale sono riconducibili a effetti di frequenza piuttosto che alla presenza di tratti lessicali categoriali.*

Introduction

The role of frequency in lexical retrieval is well known for what concerns psycholinguistic studies (since, at least, Forster & Chambers, 1973): the higher the frequency of a word, the faster its retrieval. Generally, the singular form of a noun is more frequent than the corresponding plural, and thus retrieved faster. However, some nouns (e.g. *stelle*, 'stars') do occur more frequently in the plural than in the singular: the phenomenon is known as plural dominance. Plural dominant nouns are in fact accessed faster in the plural form than in the singular. Since such nouns are not identifiable as a homogeneous group by means of some semantic features, the phenomenon has been explained as a mere effect of the frequency of occurrence of the forms (Baayen et al., 1996; 1997; 2007; Biedermann et al., 2013).

While the plural dominance seems to be unrelated to grammatical constraints, another phenomenon involving Number morphology seems to be grammatically grounded instead, namely the mass-count issue (Borer, 2005; Cheng, 1973; Chierchia, 2010; Jackendoff, 1991). Nouns referring to countable entities are called 'count nouns' (*anello*, 'ring'), nouns referring to uncountable entities are called 'mass nouns' (*burro*, 'butter'). Some constraints rule the possibility for the two types of nouns to occur in some morphosyntactic contexts, for example count nouns cannot occur in the singular after a quantifier (**molto anello*, 'much ring'), while mass nouns cannot occur with numerals or the indeterminate article (**un burro*, 'a butter'). For what concerns Number morphology, mass nouns should occur only inflected in the singular (but for a deeper discussion, see i.a. Acquaviva, 2013; Marcantonio & Pretto, 2001; Pelletier, 2012).

Previous lexical decision tasks have pointed out to some differences in the processing of count nouns with respect to mass nouns, which would require longer response times (RTs) (i.a. Mondini et al. 2009; Gillon et al. 1999). In the light of these results, it has been proposed that an additional lexical feature has to be computed for mass nouns as compared to count nouns.

While psycholinguistic studies on plural dominance have relied on relative frequency of singular and plural forms in the selection of stimuli and in results analysis, even the most recent experimental studies on the mass-count issue have not quantified the actual occurrence of the exper-

imental stimuli in mass context and in count context: nouns have rather been assigned to a mass or to a count category on the basis of the experimenters' judgments. Quantitative data on syntactic contexts can instead provide a better estimate of the frequency of use of nouns as countable or uncountable: in the present study we relied on the actual occurrence of nouns in the different syntactic contexts in assigning them to the "mass" or to the "count" experimental list.

We will describe and put into comparison two lexical decision tasks, concerning the phenomena of mass-count and of plural dominance respectively. We will explore the possibility that the mass-count effects described in psycholinguistic literature could be better explained in terms of frequency of occurrence, as it is recognized by most literature with respect to plural dominance. We hypothesize that the frequency of occurrence of the word form (inflected in the singular or in the plural) will predict the RTs in lexical decision tasks contrasting mass and count nouns, as well as in the ones concerning the plural dominance issue. The frequency of occurrence will be measured by means of two subjective frequency rating studies and in the corpus ItWaC (Baroni et al. 2009). We will rely on quantitative measures to categorize experimental stimuli. Measures of plural dominance of nouns will be based on the ratio between their occurrence in the plural and in the singular; the mass and count experimental nouns will be categorized considering their distribution with respect to mass and count morpho-syntactic contexts.

1 First study: mass and count nouns

1.1 Rating and corpus analysis

448 concrete nouns, namely 224 nouns inflected both in the singular and in the plural, were selected following the theoretical definitions given in traditional grammars. The list included the plural of 45 nouns for which only singular occurrences would be expected on a normative basis (pure "mass" nouns such as *burro* 'butter' - **burri* 'butters').

A questionnaire was designed in order to evaluate the subjective frequency of the 448 nouns following the methods used in previous literature (Ferrand et al., 2008). The questionnaire was administered online by means of the SurveyMonkey platform. 126 informants participated in this study (age: range = 22 - 76 years, mean = 36.2, SD = 12.46; years of education: range = 8-21). Participants were instructed not to express normative judgments, but to focus on the frequency they had heard or read the words; they had to assign a score to the frequency of the nouns on a 7-point Likert scale, ranging from 0 = "never heard or seen" to 6 = "more than once a day". The nouns in the questionnaires were presented to each participant in a different random order.

Score mean	Singular	Plural
$n = 0$	0	0
$0 < n \leq 1$	0	7
$1 \leq n \leq 2$	3	47
$2 \leq n \leq 3$	45	60
$3 \leq n \leq 4$	88	63
$4 \leq n \leq 5$	70	36
$n > 5$	14	7

Table 1: Distribution of the subjective frequency scores.

Absolute frequency of the aforementioned nouns was collected on the ItWaC corpus (Baroni et al., 2009). A positive correlation was found between corpus frequency and subjective frequency: $r(446) = 0.75$, $p < .001$. In order to disambiguate the mass use from the count use of the nouns presented in the rating questionnaire, we designed queries in CQP syntax following the methods described by Katz & Zamparelli (2012). The occurrence of nouns with determiners such as the indeterminate article and quantifiers were used to trace the occurrence in unambiguous count or mass context.

1.2 Lexical decision task

From the initial list of 224 nouns, 80 nouns were selected and presented both in the singular and in the plural (totally 160 experimental stimuli). These stimuli were selected to span as uniformly as possible across the range of possible values of subjective frequency in order to use the subjective frequency as a continuous variable in the analysis. From the 80 nouns we classified as "mass" the 18 top mass-used nouns with the highest mass frequencies and values of count frequencies that were not among the top 18; we classified as "count" the 18 top count-used nouns with the highest count frequencies and values of mass frequencies that were not among the top 18. The nouns were presented both in the singular and in the plural (totally 72). The remaining stimuli were not categorised in such terms. Experimental stimuli are displayed in table 2. The

final list included 240 filler words, consisting in 80 adjectives and 160 phonotactically plausible non-words.

	N. of items	Corpus Frequency	Subjective Frequency	Length
All stimuli	160	11850.32 (27239.65)	3.29 (1.18)	6.41 (1.66)
"Mass" nouns: singular	18	26204.88 (28831.43)	4.36 (0.57)	6.22 (1.89)
"Mass" nouns: plural	18	824 (1187.38)	1.95 (0.72)	6.28 (1.96)
"Count" nouns: singular	18	38570.05 (54194.95)	4.09 (0.84)	5.78 (1.31)
"Count"nouns: plural	18	24365 (36455)	4.07 (0.80)	5.89 (1.27)

Table 2: Psycholinguistic properties of experimental stimuli.

60 Italian native speakers participated in the experiment (mean age = 23.5, SD = 2.37; years of education: mean = 15.16, SD = 1.64). Participants saw a series of letter strings presented at the center of the screen one at a time. They had to press a key if they thought the string was an Italian word, another key in the converse case.

1.3 Results

Results were analyzed by means of mixed effect models (Baayen, Davidson & Bates, 2008). In the model 1, summarized in table 3, we included the 72 stimuli classified as mass and count nouns. We considered as predictors: *category (mass/count)*, *Number (singular/plural)*, *corpus frequency*, *subjective frequency* and *orthographic length*. Results show significant effects of length (longer RTs for longer items), of Number (longer RTs for plurals) and of subjective frequency (longer RTs for low subjective frequency).

Fixed effect	Coefficient	Standard Error	df	t	p-value
Intercept	6.56	0.05	95.18	130.53	<0.001
Number= plural	0.37	0.02	64.33	2.04	0.04
Subjective frequency	-0.04	0.007	74.09	-4.27	<0.001
Orthographic length	0.009	0.004	65.86	2.077	0.04

Table 3: Results of model 1.

In model 2, summarized in table 4, we included all the 160 stimuli. We considered as predictors: *Number* (singular/plural), *corpus frequency*, *subjective frequency* and *orthographic length*. Results show significant effects of length (longer RTs for longer items), of corpus frequency (longer RTs for low corpus frequency) and of subjective frequency (longer RTs for low subjective frequency).

Notably, the predictor c*ategory* is not significant (p = 0.85); *corpus frequency* is a significant predictor in model 2 (p = 0.03), but it only approached significance in model 1 (p = 0.05). Possibly, in model 1 *Number* is a significant predictor because the categorised items represent a subset that differ for frequency of occurrence in the plural. In fact, in model 2, in which both categorised and not categorised items were considered, no effect of Number was found.

Fixed effect	Coefficient	Standard Error	df	t	p-value
Intercept	6.73	0.04	219.42	172.38	<0.001
Corpus frequency	-0.009	0.004	155.55	-2.16	0.03
Subjective frequency	-0.05	0.008	152.19	-5.37	< 0.001
Orthographic length	0.008	0.004	2.47	2.11	0.04

Table 4: Results of model 2.

2 Second study: plural dominance

2.1 Rating and corpus analysis

The ItWaC corpus was queried to obtain the frequency of occurrence of the singular and the plurals of nouns displaying the most common inflectional patterns (-o/-i; -a/-e). We discarded from testing material compounds, derived nouns and the nouns that differ for orthographic length or phonological form between singular and plural (e.g. *occhio* - *occhi* 'eye –eyes'). The remaining nouns were then ordered on the base of their plural dominance defined as the ratio plural frequency/singular frequency. We calculated stem frequency of nouns and selected 284 nouns uniformly span across the range of possible values of frequency.

A questionnaire was created in order to test the subjective frequency of the 284 selected nouns, both in the singular ad the plural (568 experimental items). The questionnaire was administered following the same methods described previously (§2.1). 150 Italian native speakers participated in the study (age: range = 18 – 69, mean = 29; years of education: range = 8-21). The distribution of the subjective frequency is plotted in Table 5. A positive correlation was found between the singular and plural forms

of nouns within the corpus ($r(282) = 0.70$, $p < .001$) and within the rating ($r(282) = 0.91$, $p < .001$).

Score mean	Singular	Plural
n = 0	0	0
0 < n ≤ 1	0	1
1 ≤ n ≤ 2	19	20
2 ≤ n ≤ 3	88	100
3 ≤ n ≤ 4	139	131
4 ≤ n ≤ 5	31	27
n > 5	7	5

Table 5: Distribution of the subjective frequency scores.

2.2 Lexical decision task

A lexical decision study was carried out, following the same methods described in §2.2. From the 284 nouns mentioned in §3.1, we chose: the 30 nouns with the highest ratio of plural dominance, the 30 nouns with the lowest ratio of plural dominance, the 30 nouns whose ratio between singular ad plural was the closest to 1 (see table 6). Each noun was presented in the singular and in the plural (totally 180 experimental stimuli). The final list included 364 filler words, consisting in 184 adjectives and 180 phonotactically plausible non-words.

43 Italian native speakers participated in the experiment.

Dominance (mean Pl/Sg)	Morphological Number	N. of items	Corpus Frequency	Subjective Frequency	Orthographic Length
Plural (3.61)	Singular	30	5260.3 (7547.43)	3.31 (0.77)	6.33 (1.09)
	Plural	30	19026.46 (25558.41)	3.48 (0.79)	
Singular (0.16)	Singular	30	25596.9 (44944.15)	3.44 (0.91)	6.13 (1.13)
	Plural	30	4276.3 (7186.03)	3.23 (0.79)	
Equal (0.9)	Singular	30	35430.33 (99471.4)	3.13 (0.57)	6.16 (1.17)
	Plural	30	31921.7 (93584.35)	3.1 (0.59)	

Table 6: Psycholinguistic properties of experimental stimuli.

2.3 Results

Results were analysed by means of mixed effect models (Baayen, Davidson & Bates 2008). In model 3, summarized in table 7, we considered as predictors: *category (plural/singular/equal dominant), Number (singular/plural), corpus frequency, subjective frequency* and *orthographic length*. Results show significant effects of length (longer RTs for longer items), of corpus frequency (longer RTs for low corpus frequency) and of subjective frequency (longer RTs for low subjective frequency).

Fixed effect	Coefficient	Standard Error	df	t	p-value
Intercept	6.79	0.04	211.64	137.79	<0.001
Corpus frequency	-0.02	0.003	171.55	-7.23	<0.001
Subjective frequency	-0.03	0.007	170.17	-4.48	< 0.001
Orthographic length	0.009	0.004	165.93	2.03	0.04

Table 7: Results of model 3.

3 Discussion and conclusions

In this study we applied quantitative methods in the selection of experimental stimuli used in the two lexical decision tasks. In both tasks, results from the three models showed effects of subjective frequency and corpus frequency but not of category in written word recognition. For what concerns the plural dominance issue, this result was in line with previous literature. For what concerns the mass-count issue, our results are unexpected instead. Remind that frequency of occurrence in mass and count contexts was used to avoid biases in categorization of stimuli. Nevertheless, we did not observe differences in RTs between the two so categorized groups of nouns. Thus, we suggest that there is no need to postulate the computation of a lexical feature related to countability or uncountability in nouns. We propose that the fact that a noun is considered "mass" is better described as an epiphenomenon of the distribution of noun with respect of syntactic contexts. However the possibility for a noun to occur in the different syntactic contexts does not predict lexical decision RTs: frequency, as measured in the corpus and by the rating study, is the predictor of the lexical access times with respect to words presented in isolation. In this sense, the mass-count issue is similar to the plural dominance phenomenon: even in that case, there is no need to assume the presence of a feature marking plurality, as the frequency of the inflected form is sufficient to account for the observed effects in lexical decision tasks.

The frequency of occurrence of nouns considered as a continuous variable is a better predictor of RTs than a distinction attributed to alleged lexical categories both in the case of phenomena

seemingly unrelated to core grammar rules, like the plural dominance, as well as in phenomena that have traditionally been described as grammar based, like the mass-count issue.

References

Acquaviva, P. (2013). *Il nome*. Roma: Carocci.

Baayen, H., Burani, C., & Schreuder, R. (1996). Effects of semantic markedness in the processing of regular nominal singulars and plurals in Italian. *Yearbook of morphology*, Springer Netherlands, 13-33.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, 37(1), 94-117.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412.

Baayen, R., Levelt, W., Schreuder, R., & Ernestus, M. (2007). Paradigmatic structure in speech production. *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, 43(1): 1-29. Chicago Linguistic Society.

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition* 29(4), 639-647.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation* 43 (3), 209-226.

Biedermann, B., Beyersmann, E., Mason, C., & Nickels, L. (2013). Does plural dominance play a role in spoken picture naming? A comparison of unimpaired and impaired speakers. *Journal of Neurolinguistics*, 26(6), 712-736.

Borer, H. (2005). *In name only*. Oxford: OUP.

Cheng, C.-Y. (1973). Response to Moravcsik. J. Hintikka, J.M.E. Moravcsik, & P. Suppes (eds.). *Approaches to Natural Language*. Dordrecht: Reidel, 286-288.

Chierchia, G. (2010). Mass nouns, vagueness and semantic variation. *Synthèses* 174, 99-149.

Ferrand, L., Bonin, P., Méot, A., Augustinova, M., New, B., Pallier, C., & Brysbaert, M. (2008). Age-of-acquisition and subjective frequency estimates for all generally known monosyllabic French words and their relation with other psycholinguistic variables. *Behavior Research Methods* 40 (4), 1049-1054.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time.*Journal of verbal learning and verbal behavior*, *12*(6), 627-635.

Gillon, B., Kehayia, E., & Taler, V. (1999). The mass/count distinction: Evidence from on-line psycholinguistic performance. *Brain and Language* 68, 205-211.

Jackendoff, R. (1991). Parts and boundaries. *Cognition* 41, 9-45.

Katz, G. & Zamparelli, R. (2012). Quantifying Count/Mass Elasticity. Choi, J. et al. (eds). *Proceedings of the 29th West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla Proceedings Project, 371-379.

Kulkarni, R., Rothstein, S., & Treves, A. (2013). A Statistical Investigation into the Cross-Linguistic Distribution of Mass and Count Nouns: Morphosyntactic and Semantic Perspectives. *Biolinguistics* 7, 132-168.

Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance* 39(3), 802.

Marcantonio, A. & Pretto, A. M. (2001). Il nome. L. Renzi, G. Salvi, & A. Cardinaletti (eds.). *Grande grammatica italiana di consultazione*. Bologna: Il Mulino, 329-346.

Mondini, S., Kehaya, E., Gillon, B., Arcara, G., & Jarema, G. (2009). Lexical access of mass and count nouns. How word recognition reaction times correlate with lexical and morphosyntactic processing. *The Mental Lexicon* 4, 354-379.

Pelletier, F. J. (2012a). Lexical Nouns are Neither Mass nor Count, but they are Both Mass and Count. D. Massam (ed.). *A Cross-Linguistic Exploration of the Count-Mass Distinction*. Oxford: OUP, 9-26.

Williams, R., & Morris, R. (2004). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology* 16(1/2), 312–339.

Semantic priming effects in Italian verbs recognition: the role of grammatical classes and semantic categories

Azzurra Mancuso
University of Salerno
Via Giovanni Paolo II, 132
Fisciano, SA, 84084, Italy
amancuso@unisa.it

Maria De Martino
University of Salerno
Via Giovanni Paolo II, 132
Fisciano, SA, 84084, Italy
mdemartino@unisa.it

Alessandro Laudanna
University of Salerno
Via Giovanni Paolo II, 132
Fisciano, SA, 84084, Italy
alaudanna@unisa.it

Abstract

English. The hypothesis that grammatical class information is represented in the mental lexicon and that it is activated during lexical access has generated a wide literature about the differences between nouns and verbs. However, the available evidences are discordant.
In this study we tried to disentangle grammatical class effects from semantic categories effects during visual word recognition by exploiting the semantic priming paradigm.
Semantically related prime-target pair were arranged. They could share (verb-verb) or not (noun-verb) grammatical class information. A third condition was included where noun primes and verb targets had both an action as a referent (*delitti/uccide*, crimes/he-she kills).
Only prime/target pairs sharing grammatical class information showed significant semantic priming effects.
Results are compatible with the hypothesis that grammatical class is an organizational criterion in the mental lexicon and it is activated during lexical access.

Italiano. *L'ipotesi secondo cui l'informazione di classe grammaticale sia rappresentata nel lessico mentale e attivata durante l'accesso lessicale ha dato origine a un'ampia letteratura sulle differenze tra nomi e verbi. La base empirica è, tuttavia, ancora incerta.*
In questo studio abbiamo usato il paradigma del priming semantico per distinguere il ruolo della classe grammaticale da quello della categoria semantica di appartenenza delle parole in un compito di riconoscimento visivo.
Sono state impiegate coppie prime-target semanticamente collegate che condividevano (verbo-verbo) o meno (nome-verbo) la classe grammaticale. In una terza condizione abbiamo usato prime-nome e target -verbo che avevano entrambi un referente appartenente alla categoria semantica delle azioni (delitti/uccide).
L'effetto di priming semantico è risultato significativo solo per le coppie prime-target che condividevano l'informazione di classe grammaticale.
I dati sono compatibili con l'idea che la classe grammaticale sia un criterio organizzativo nel lessico mentale e che sia attivata durante l'accesso lessicale.

1 Introduction

Psychological and neural evidence revealed that the distinction between parts of speech, mainly nouns and verbs, occurring in all languages (Sapir, 1921) affects speaker's performance: grammatical class is preserved in speech errors (Garrett, 1982) and nouns and verbs can be selectively disrupted in aphasic populations (Collina, Marangolo, and Tabossi, 2001; Miceli, Silveri, Villa, and Caramazza, 1984; Miceli, Silveri, Nocentini, and Caramazza, 1988).
An influential hypothesis states that the grammatical class is an organizing principle in the mental lexicon (Caramazza and Hillis, 1991) but the picture of empirical data is actually multifaceted and often inconsistent (for a review, see Vigliocco, Vinson, Druks, Barber, and Cappa, 2011).
For the sake of conciseness, here we focus on two examples of alternative interpretations about the role of grammatical class in lexical representation and processing of words.
A first position, mostly grounded on word production data (Pechmann, Garrett, and Zerbst, 2004; Pechmann and Zerbst, 2002; Vigliocco, Vinson, and Siri, 2005), is that grammatical class information, although lexically represented, is

only retrieved under specific circumstances, namely in sentence or phrasal contexts (Levelt, Roelofs, and Meyer, 1999; Garrett, 1982). However, grammatical class effects are significantly reported in word production even in tasks not requiring syntactic integration (Mahon, Costa, Peterson, Vargas, and Caramazza, 2007; De Simone and Collina, 2016). A stronger lexicalist view is held in the field of recognition and comprehension processes and conceives grammatical class as a feature of words that is automatically retrieved during lexical access. Comparisons between noun/verb homographs (*condannato* N (the convict) vs. *condannato* V (past participle, convicted), Postiglione and Laudanna, 2016) and homonymic nominal and verbal forms (*saliva* N, spittle vs. *saliva* V, he/she went up, Mancuso and Laudanna, 2013) revealed the possibility that separate, grammatical class-specific representations are present in the lexicon. Also in this case, different patterns of data have been described (Vigliocco, Vinson, Arciuli and Barber, 2008).

A possible reason for such a divergence relies on the fact that grammatical class effects are often not clearly distinguishable from the influence of confounding variables, i.e. the imageability of words, the number of inflectional alternatives for nominal and verbal stems, the argumental structure of nouns and verbs and so on. A challenging issue is that noun/verb distinction is not lexical in nature but relies on an object/action distinction (Vigliocco et al., 2005).

Here we aim at verifying whether lexical access to input orthographic representations of Italian verbs can be affected by the pre-activation of grammatical class information. The semantic priming paradigm was exploited and the expected facilitation effect on target verbs elicited by semantically related primes was compared across prime/target pairs sharing or not grammatical class information, i.e., noun/verb pairs vs. verb/verb pairs. In order to disentangle the possible confound between grammatical class (nouns vs. verbs) and semantic categories which nouns and verbs belong to (objects vs. actions), two different types of noun-verb pairs were used: object nouns denoting objects (*candela*, candle) vs. nouns denoting actions (*sberla*, slap).

We reasoned as follows: if grammatical class informs input orthographic representations of words, its pre-activation through primes should speed up targets recognition even in a lexical decision task where any process of syntactic integration is not involved. On the contrary, semantic priming effects are expected to equally affect prime/target pairs regardless their grammatical relation. No interaction between grammatical class and semantic relation is expected because the two variables are supposed to affect lexical selection with distinct modalities (Yudes, Domínguez, Cuetos, and de Vega, 2016).

2 Experiment

2.1 Method

Participants: Seventy-six undergraduate students (36 females) from University of Salerno voluntarily took part in the experiment. They were all native speakers of Italian, free of speech-language and hearing disorders and they all had normal or corrected-to-normal vision. Their age ranged from 18 to 31 years (AV: 22 years). They served for a session lasting about 30 minutes. Each pair of participants constituted one data point in the statistical analyses.

Materials: Sixty Italian unambiguous verbs were selected as targets and subdivided into 3 lists on the basis of the type of prime word adopted:

1. Object Noun/Verb Condition, ON/V: 20 targets were preceded by semantically related object nouns (*bottega*/*acquista*, atelier/he-she buys);
2. Action Noun/Verb Condition, AN/V: 20 targets were preceded by semantically related action nouns (*furto*/*ruba*, theft/he-she steals);
3. Verb/Verb Condition, V/V: 20 targets were preceded by semantically related verbs (*colpiva*/*spara*, he-she struck/he-she fires).

The semantic distance between prime and target was calculated on the basis of an off-line rating (on a 7-points Likert scale), previously submitted to 54 participants (who did not take part into the experiment) and balanced among conditions.

Each experimental list was matched with a control list:

1. Object Noun/Verb Control Condition, ON/V_C: 20 targets were preceded by unrelated object nouns (*polmone/acquista*, lung/he-she buys);
2. Action Noun/Verb Control Condition, AN/V_C: 20 targets were preceded by unrelated action nouns (*dormita/ruba*, the sleep/he-she steals);
3. Verb/Verb Control Condition, V/V_C: 20 targets were preceded by unrelated verbs (*variava/spara*, he modified /he-she fires).

Targets of the three lists were matched for the following variables:

- cumulative written frequency of the verb paradigm;
- written form frequency;
- length calculated in number of letters;
- lexical decision latencies and percentage errors[1].

The mean values for the controlled parameters of targets are shown in Table 1.

	Paradigm Frequency	Form Frequency	Length	LD latencies	LD %errors
ON/V	236	27	6.8	551ms	4%
AN/V	248	23	6.4	545ms	6%
V/V	238	28	6.5	553ms	5%

Table 1. Summary of targets characteristics

Primes of the both experimental and control lists were matched for:

- written form frequency;
- length calculated in number of letters.

The mean values for the controlled parameters of primes are shown in Table 2. Values for frequency were taken from the CoLFIS database (Bertinetto, Burani, Laudanna, Marconi, Ratti, Rolando & Thornton, 2005).

	Form frequency (related prime)	Length (related prime)	Form frequency (unrelated prime)	Length (unrelated prime)	Prime/target Semantic distance
ON/V	20	6.1	17	6.2	5.5
AN/V	16	7	17	7	5.3
V/V	16	7.4	11	7.7	5.2

Table 2. Summary of primes characteristics

Procedure: The participants were tested individually; an experimental session consisted of two parts: a practice and an experimental phase.
A semantic priming lexical decision task was used as experimental paradigm. Participants were asked to be as fast and accurate as possible. They had to press on two buttons: the button corresponding to their dominant hand for the decision 'word', the other for the decision 'non-word'.
Stimuli appeared in lower case letters (12-point size) in the center of the computer screen. Each experimental trial was composed by: fixation point (200 ms), blank (300 ms), prime (200 ms), blank (50 ms), target (1 sec). If the participant did not respond within 1000 ms, the feedback *"Fuori tempo"* (out of time) was given and the trial was recorded as an error. Following the participant's response (or non-response), the next trial was presented after a delay of 1 sec.
Reaction times (ms) and accuracy constituted the dependent variables.

*Equipment***:** Response box, connected to a PC running the E-Prime software 2.0 (Psychology Software Tools, Inc., Pittsburgh, PA).

2.2 Results

A repeated measures ANOVA was performed on the averaged correct response latencies and on errors with the Condition (two levels (i.e., semantically related vs. unrelated) and the Experimental List (3 levels, ON/V, AN/V and V/V) as variables. Separate analyses were carried out for participants and items, yielding F1 and F2 statistics, respectively.

Data from two items (*bombarda* (he/she bombs) and *contagia* (he/she infects) were excluded from the analyses because they elicited a number of errors exceeding the sample's mean more than 2.5 standard deviations.

Data on reaction times (reported in Table 3) revealed significant main effects of Condition [$F1(1,75)=21.3$, $p<.01$; $F2(1, 55)=5,73$, $p<.05$] and Experimental List [$F1(2,150)=21.6$, $p<.01$; $F2(2, 55)=1.2$, $p<.1$].

No significant interaction between the two variables was observed.

Interestingly, planned comparisons revealed that the observed semantic priming effect is mainly elicited by prime/target pairs sharing grammatical class information: V/V Condition ($p<.05$). On the contrary, both conditions where primes do not share grammatical class information with the targets (i.e., AN/V and ON/V) exhibit a weak semantic priming not reaching statistical significance.

[1]Prior to the study, a simple visual lexical decision experiment was administered to 35 participants (who did not take part into the priming experiment), in order to verify whether targets of the three conditions were exactly balanced with each other.

Condition	ON/V	AN/V	V/V	Overall
Related	554 (-5)	545 (-12)	562 (-18)**	554 (-11)
Unrelated	559	557	580	565

Table 3. Correct lexical decision response latencies as a function of the Condition and Experimental List

On accuracy data (reported in Table 4) only a significant effect of the Experimental List was detected [$F1(2,250) = 10.53$, $p < .01$; $F2(2, 55) = 3.98$, $p < .02$].

Condition	ON/V	AN/V	V/V	Overall
Related	1.8%	1.5%	2.6%	2%
Unrelated	1.6%	1.6%	3.2%	2.1%

Table 4. Lexical decision percentage of errors as a function of Condition and Experimental List

3 Conclusion

Our purpose here was to clarify whether grammatical class works as an organizational criterion of word representations within the mental lexicon. In particular, we aimed at demonstrating that words from different grammatical classes tend to be processed differently by speakers not only because of their differences in terms of semantic categories they belong to (actions vs. objects) or of semantic features (imageability) but also because their lexical representations specify their role as different parts of speech.

From an empirical point of view, our purpose was to verify:

- whether grammatical class information is automatically activated when orthographic representations of Italian verbs are accessed;
- whether grammatical class effects can be detected in tasks that do not explicitly require syntactic integration processes, that is during the processing of isolated words;
- whether grammatical class effects are an epiphenomenon of the semantic categories to which nouns and verbs belong to or if they are truly grammatical in nature.

We addressed the issue by exploiting the semantic priming effect, a robust and well-known effect in word recognition consisting in the advantage in lexical decision tasks exhibited by target words when preceded by semantically related primes and compared to an unrelated baseline.

Our experimental design was suitable to investigate the problem for two main reasons:

- it allows to pre-activate a definite linguistic feature, i.e. grammatical class (nouns vs. verbs) information and/or a semantic category (actions vs. objects), and to observe whether such a property can affect word processing;
- it rules out the intervention of any confound due to syntactic integration process because it focuses on lexical access to single word representations.

With that aim, we manipulated the congruency of grammatical class in different kinds of prime/target pairs. The rationale of the experiment was the following: if grammatical class informs lexical representation of words, its preactivation through the prime should modulate the expected effects of semantic priming.

In order to specifically disentangle the role of grammatical class from the influence of semantic category of referents of nouns and verbs, we observed the effect in different conditions: grammatically congruent prime/target pairs, prime/target pairs from incongruent grammatical classes but both belonging to the semantic category of actions, and prime/target pairs from incongruent grammatical classes and different semantic categories (objects for nouns and actions for verbs).

Our results showed that semantic priming is effective only for prime/target pairs sharing grammatical class information; much weaker effects were detected for noun/verb pairs, regardless of the semantic category of the referents.

This pattern of data seems to indicate that grammatical class informs lexical representations in the orthographic input lexicon since its preactivation through the prime modulates the expected facilitation induced by semantically related primes. In other words, grammatical class is likely to be automatically activated during lexical access to written representation of Italian verbs and, in addition, it is effective during processing of verbal forms presented outside a sentence context. This effect seems to have a truly grammatical basis as it is not elicited by grammatically incongruent prime/target pairs. Moreover, in our experiment the congruency of grammatical class between prime and target does not interact with the semantic similarity between prime and target:

this suggests that the two sources of information affect the word recognition process with distinct modalities.

This pattern of data, although preliminary, adds new challenging details to the debate about lexical representation of grammatical class information and provides evidence in favor of the lexicalist models that conceive grammatical class as an intrinsic property of the lexical representation consulted during lexical access which is necessarily and automatically accessed at least during written word recognition processes.

Reference

Bertinetto, P. M., Burani, C., Laudanna, A., Marconi, L., Ratti, D., Rolando, C., and Thornton, A. M. 2005. *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)*. http://linguistica.sns.it/CoLFIS/Home.htm

Caramazza, A. and Hillis, A. E. 1991. Lexical organization of nouns and verbs in the brain. *Nature*, 349(6312):788-790.

Collina, S., Marangolo, P., and Tabossi, P. 2001. The role of argument structure in the production of nouns and verbs. *Neuropsychologia*, 39(11):1125-1137.

De Simone, F. and Collina, S. 2016. The Picture–Word Interference Paradigm: Grammatical Class Effects in Lexical Production. *Journal of Psycholingustic Research*, 45(5):1003-19.

Garrett, M. F. 1982. Production of speech: Observations from normal and pathological use. In A. W. Ellis (Ed.), *Normality and pathology in cognitive functions*. Academic Press, London.

Levelt, W. J., Roelofs, A., and Meyer, A. S. 1999. A theory of lexical access in speech production. *Behavioral and brain sciences*, 22(01):1-38.

Mahon, B. Z., Costa, A., Peterson, R., Vargas, K. A., and Caramazza, A. 2007. Lexical selection is not by competition: a reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(3):503-535.

Mancuso, A. and Laudanna, A. 2013. Revisiting the ambiguity effect in word recognition: part of speech and meaning dominance effects. *Proceedings of AMLaP, Architectures and Mechanisms for Language Processing*, Marseille, 2-4 September 2013.

Miceli, G., Silveri, M. C., Nocentini, U., and Caramazza, A. 1988. Patterns of dissociation in comprehension and production of nouns and verbs. *Aphasiology*, 2(3-4):351-358.

Miceli, G., Silveri, M. C., Villa, G., and Caramazza, A. 1984. On the basis for the agrammatic's difficulty in producing main verbs. *Cortex*, 20(2):207-220.

Pechmann, T. and Zerbst, D. 2002. The activation of word class information during speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(1):233-243.

Pechmann, T., Garrett, M., and Zerbst, D. 2004. The time course of recovery for grammatical category information during lexical processing for syntactic construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(3):723-728.

Postiglione, F. and Laudanna, A. 2016. Competition in lexical processing of Italian noun/verb homographs. *Journal of Cognitive Psychology*, 28:1-16.

Sapir, E. 1921. *Language. An Introduction to the Study of Speech*. Harcout, Brace, New York.

Yudes, C., Domínguez, A., Cuetos, F., and de Vega, M. 2016. The time-course of processing of grammatical class and semantic attributes of words: Dissociation by means of ERP. *Psicológica*, 37(2):105-126.

Vigliocco, G., Vinson, D. P., and Siri, S. 2005. Semantic similarity and grammatical class in naming actions. *Cognition*, 94(3):91-100.

Vigliocco, G., Vinson, D. P., Arciuli, J., and Barber, H. 2008. The role of grammatical class on word recognition. *Brain and language*, 105(3):175-184.

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., and Cappa, S. F. 2011. Nouns and verbs in the brain: a review of behavioral, electrophysiological, neuropsychological and imaging studies. *Neuroscience & Biobehavioral Reviews*, 35(3):407-426.

Sentiment Analysis: applicazione in un dominio psico-forense

Francesca Guglielmi[1] and **Pierpaolo Basile**[2] and **Antonietta Curci**[1] and **Giovanni Semeraro**[2]
[1]Dipartimento di Scienze della Formazione, Psicologia, Comunicazione, Università di Bari Aldo Moro
[2]Dipartimento di Informatica, Università di Bari Aldo Moro
[1]{francesca.guglielmi,antonietta.curci}@uniba.it
[2]{pierpaolo.basile,giovanni.semeraro}@uniba.it

Abstract

English. This study aims to apply the sentiment analysis to a psycho-forensic context through the analysis of narrative transcriptions related to the crimes committed by violent detainees. We investigate the presence of psychopathy through the Psychopathic Personality Inventory-Revised (PPI-R). Psychopathy is a personality disorder that is characterized by emotional detachment, the lack of empathy and it is often found in the prison population for the brutality of committed crimes. Our study explores possible associations between psychopathy and emotional content present in the narratives. Results show a neutral polarity for both psychopathic and not psychopaths offenders, however it is possible to identify significant emotional characteristics that diversify the narrative transcriptions.

Italiano. *Il presente studio ha l'obiettivo di applicare la sentiment analysis ad un contesto psico-forense attraverso l'analisi di resoconti narrativi relativi ai crimini commessi da detenuti, autori di reati violenti. Attraverso lo strumento Psycopathic Personality Inventory-Revised (PPI-R), è stata indagata la presenza di psicopatia, disturbo di personalità che si caratterizza per il distacco emotivo, l'assenza di empatia e lo stile di vita antisociale riscontrabile nella popolazione carceraria per l'efferatezza dei crimini commessi, al fine di esplorare eventuali associazioni tra il disturbo e i contenuti emotivi presenti nelle narrazioni. Dai risultati emerge un tono affettivo neutro sia per gli offender psicopatici sia per gli offender non psicopatici, tuttavia è possibile individuare peculiarità emotive significative che diversificano i resoconti narrativi.*

1 Introduzione

La psicopatia è un disturbo di personalità caratterizzato da deficit emotivi, stile di vita antisociale, mancanza di senso di colpa ed empatia. Negli psicopatici la coloritura affettiva che generalmente serve a differenziare gli eventi emotivi da quelli neutri non esiste nella stessa misura degli altri individui (Cleckley, 1941).

Gli psicopatici hanno un'attenuata risposta emotiva agli eventi di cui fanno esperienza e nei confronti delle persone con cui interagiscono. I tratti caratterizzanti la psicopatia, si riscontrano in alcuni offender detenuti (Serin, 1996), proprio per le peculiarità relative alla violenza degli atti esperiti da queste persone (Yang et al., 2010). Inoltre, la callosità emotiva, tratto caratterizzante il disturbo, pare possa essere associata ad una percezione minimizzata dell'esperienza emotiva (positiva o negativa) dell'evento vissuto, relativa sia al momento in cui si è verificato sia al richiamo dell'evento stesso (Osumi and Ohira, 2010).

L'incapacità nell'elaborare e nel fare esperienza del contenuto affettivo degli stimoli e degli eventi può avere delle conseguenze sulle capacità di memoria e quindi sulla possibilità di elaborare e di recuperare l'informazione emotiva relativa all'evento esperito. Evidenze empiriche (Burrow et al., 2014) sostengono che le persone psicopatiche soffrono di un deficit emotivo che può avere un impatto sul loro modo di percepire il ricordo degli eventi. Alcuni clinici e ricercatori sostengono

che gli psicopatici generalmente non sono in grado di apprendere il significato emotivo degli eventi; pertanto si ipotizza che i racconti degli offender psicopatici siano scarni di dettagli emotivi rispetto alle narrazioni degli offender non psicopatici (Glass and Newman, 2006; Wilson et al., 2008; Christianson et al., 1996)

La ricerca in questi ambiti necessita di essere implementata, infatti negli ultimi anni la ricerca di base in psicologia cognitiva si è ampliata mediante l'adattamento delle procedure, delle variabili e il coinvolgimento di altri domini scientifici, tra cui la psicolinguistica. Tali adattamenti consentono di allargare le conoscenze sulle caratteristiche di personalità e le loro influenze sui comportamenti in contesti diversificati, come quello penitenziario, per rendere fruibile la messa in atto di programmi di intervento mirati e adatti a supportare persone con tali disturbi. Il presente studio è il risultato di queste integrazioni di domini concettuali e di ricerca, nella fattispecie si applica la sentiment analysis, una metodologia per l'elaborazione automatica del testo al fine di estrarre informazioni soggettive in un contesto psico-forense. A 48 offender detenuti, autori di reati violenti, è stato chiesto, attraverso un compito di free-recall, di narrare il ricordo autobiografico relativo all'evento criminoso commesso per il quale sono stati condannati, tale ricordo è stato trascritto dagli sperimentatori simultaneamente alla narrazione. Inoltre, a ciascun partecipante è stato somministrato il Psychopatic Personality Inventory-Revised (PPI-R) (Lilienfeld and Widows, 2005) per analizzare la presenza di psicopatia. Ciascun resoconto narrativo è stato successivamente analizzato con la metodologia della sentiment analysis attraverso un algoritmo descritto in dettaglio nel paragrafo 2. I risultati di questa analisi sono riportati nel paragrafo 3.

2 Metodologia

La metodologia da noi proposta si compone di cinque distinte fasi:

1. reperimento dei resoconti narrativi. Ad un campione composto da 48 uomini ($M_{età} = 38,69$; $Ds = 8,5$) autori di reati violenti, detenuti presso l'Istituto penitenziario di Turi (BA) e la Casa circondariale di Trani (BA) è stato chiesto di narrare, attraverso un compito di free-recall, il ricordo autobiografico relativo al crimine per il quale sono stati condannati. Le narrazioni sono state trascritte dagli sperimentatori in simultanea;
2. rilevazione della presenza/assenza di psicopatia misurata attraverso il PPI-R (Lilienfeld and Widows, 2005) che valuta i tratti di personalità del soggetto attraverso l'indagine di quei comportamenti e peculiarità cognitive, emotive e percettive che costituiscono i tratti tipici della psicopatia. Il test è composto da 154 item, da cui si ricavano 8 scale di contenuto, aggregate in 3 fattori sovraordinati: *Dominanza priva di paura*, *Impulsività auto centrata* e *Freddezza emotiva*. Ogni item prevede 4 alternative di risposta (vero, abbastanza vero, abbastanza falso, falso). Le somministrazioni si sono tenute previa autorizzazione del P.R.A.P. della Regione Puglia e previa autorizzazione individuale all'uso dei dati personali nei limiti stabiliti dalla legge. Le somministrazioni si sono svolte in una cella protetta, alla presenza di due somministratori e una guardia penitenziaria;
3. elaborazione dei resoconti narrativi, come descritto nel paragrafo 2.1, per l'estrazione automatica di attributi emotivi e di sentimento;
4. analisi dei risultati per l'individuazione di una correlazione tra i punteggi del PPI-R e gli attributi estratti dall'analisi automatica del testo;
5. analisi degli attributi estratti dall'elaborazione del testo con tecniche di apprendimento automatico.

La metodologia proposta ha l'obiettivo di verificare se l'analisi automatica del testo può rilevare alcune caratteristiche emozionali o di sentimento associabili alla psicopatia nei detenuti che hanno commesso crimini particolarmente violenti.

2.1 Elaborazione del testo

Ogni trascrizione è stata processata da una pipeline di strumenti per l'analisi automatica del linguaggio per la lingua italiana. In particolare, ogni trascrizione è sottoposta alle seguenti operazioni:

1. **Tokenizzazione e PoS-tagging:** il testo è suddiviso in token (parole) e ad ogni token è associato un tag che identifica la parte del discorso (sostantivo, aggettivo, verbo, avverbio, preposizione, articolo, ...). Queste operazioni sono realizzate attraverso il tool

open-source Apache OpenNLP[1];

2. **Lemmatizzazione:** sfruttando le informazioni sul PoS-tag, riconduce ogni parola al suo lemma. Questa operazione è necessaria per poter accedere a dizionari elettronici, ad esempio quelli utilizzati nelle operazioni successive per l'analisi del sentiment. L'operazione è stata realizzata attraverso l'utilizzo della risorsa linguistica Morph-it![2] (Zanchetta and Baroni, 2005);
3. **Analisi del sentiment:** l'analisi del sentiment prevede l'associazione di uno score di polarità ad ogni parola presente nella trascrizione, ed il calcolo di uno score di sentiment sull'intera trascrizione. Lo score di sentiment sull'intera trascrizione è calcolato seguendo due differenti approcci: 1) come media degli score di sentiment associati ad ogni parola; 2) come differenza tra gli score massimi di polarità positiva e negativa. Questo processo è realizzato utilizzando la risorsa linguistica Sentix (Basile and Nissim, 2013). Inoltre, la polarità dei termini è invertita se compaiono in un contesto di negazione. Il contesto di negazione è individuato dalla parola "non" e si estende fino all'individuazione di uno dei seguenti caratteri $\{.\ :;!?\}$;
4. **Analisi delle emozioni:** ad ogni parola sono associate delle etichette emotive utilizzando la risorsa linguistica WordNet-Affect (Strapparava and Valitutti, 2004).

L'analisi del sentiment e quella delle emozioni necessitano di maggiori dettagli implementativi. Poiché le risorse linguistiche utilizzate sono dei dizionari basati sul significato si rende necessaria una strategia per poter accedere a questi dizionari attraverso il lemma. Un lemma può avere più significati; nel caso di lemma polisemico si è scelto il significato più frequente. La frequenza dei significati per la lingua italiana è calcolata utilizzando la risorsa MultiSemCor (Bentivogli and Pianta, 2005). Lo stesso procedimento viene applicato a WordNet-Affect in presenza di lemmi polisemici.

2.2 Analisi con approcci di apprendimento automatico

Al fine di analizzare l'associazione tra gli attributi linguistici e la psicopatia, abbiamo formalizzato un problema di apprendimento di tipo binario in cui ogni racconto è classificato come *psicopatico* o *non psicopatico* in base ai risultati del PPI-R. Ogni racconto è descritto da una serie di attributi: polarità media, score massimo di polarità positiva, score massimo di polarità negativa, differenza tra gli score massimi di polarità e, per ogni classe emotiva prevista da WordNet-Affect, numero di parole che ricadono in tale classe.

Lo scopo di questa analisi non è quello di misurare la bontà predittiva degli attributi estratti dal testo, ma di capire il potere discriminante di ogni attributo. Per questo motivo si è scelto l'algoritmo C4.5 (Quinlan, 1993) basato sugli alberi di decisione. In particolare l'algoritmo genera un albero di decisione basandosi sul concetto di entropia dell'informazione.

Ad ogni nodo dell'albero, l'algoritmo C4.5 sceglie l'attributo che divide più efficacemente i dati in una o nell'altra classe. Il criterio di suddivisione è il guadagno d'informazione (differenza di entropia). L'attributo con il più alto guadagno d'informazione viene scelto per prendere la decisione. L'algoritmo C4.5 viene applicato in maniera ricorsiva ai nuovi sottoinsiemi di dati che si vengano a creare ad ogni passo.

L'idea è che l'albero di decisione possa darci un'indicazione di quali siano gli attributi più discriminanti per l'individuazione dei racconti appartenenti a soggetti affetti da psicopatia.

3 Analisi sperimentale

Il punteggio totale di psicopatia, che ha consentito la discriminazione dei due gruppi: offender psicopatici (n. 20) e offender non psicopatici (n. 28), è stato calcolato attraverso moduli di autoscoring sommando, per ciascun soggetto, i punteggi ottenuti nei tre fattori (*Dominanza priva di paura*, *Impulsività auto centrata* e *Freddezza emotiva*). Tale punteggio grezzo è stato poi convertito in punti t, quindi standardizzato. Attraverso il t-test, test statistico parametrico che permette di verificare se il valore medio di una distribuzione si discosta significativamente da un certo valore di riferimento ad un livello di significatività inferiore a 0,05, emerge che sia i punteggi totali di Psicopatia ($t(46) = -9,86$), sia dei tre fattori: *Dominanza priva di paura* ($t(46) = -8,24$); *Impulsitvità autocentrata* ($t(46) = -9,86$) e *Freddezza emotiva* ($t(46) - 8,24$) hanno delle differenze significative con valori di p $<$ 0,05. Questo risultato la-

[1] https://opennlp.apache.org/

[2] http://sslmitdev-online.sslmit.unibo.it/linguistics/morph-it.php

Tabella 1: Correlazione di Pearson tra gli indicatori della Sentiment Analysis e i fattori del PPI-R nel gruppo di offender psicopatici (n = 20). (*) $p < 0.05$

Sentiment Analysis	**Psicopatia: PPI-R**			
Polarità	Impulsività autocentrata	Dominanza priva di paura	Freddezza emotiva	Psicopatia totale
M_{pol}	-0,260	0,103	0,097	-0,095
Max_{pos}	-0,006	0,507*	0,335	0,305
Max_{neg}	-0,116	0,104	-0,121	-0,054
$Max_{pos} - Max_{neg}$	0,069	0,291	0.311	0,249

scia ragionevolmente supporre che i due gruppi si differenziano per i tratti caratterizzanti la psicopatia relativi allo stile di vita antisociale, la callosita emotiva e le caratteristiche legate alla mancanza di paura, gestione dello stress e influenza sociale.

Successivamente, i resoconti narrativi sono stati analizzati mediante tecniche di elaborazione del linguaggio naturale al fine di estrarre attributi emotivi e di polarità del sentimento (*positiva*, *negativa* e *neutra*). Gli attributi così estratti sono stati utilizzati per impostare il problema di apprendimento automatico descritto nel paragrafo 2.2. Per realizzare ciò è stata utilizzata l'implementazione J48 dell'algoritmo C4.5 presente nel tool open-source Weka[3] (Hall et al., 2009).

3.1 Risultati

Per analizzare l'associazione tra psicopatia e punteggi ottenuti negli score della Sentiment Analysis abbiamo utilizzato analisi correlazionali attraverso le quali è emersa una correlazione significativa tra Max_{pos} e il fattore *Dominanza priva di paura* del PPI (Tabella 1) nel gruppo composto da offender psicopatici. Questo risultato sembra sostenere l'ipotesi che la freddezza emotiva dello psicopatico predisponga ad una ricerca attiva di tutte quelle situazioni e sensazioni forti ed eccitanti che possano in qualche modo indurre uno stato di eccitazione in loro (Hervé et al., 2007), come il commettere i crimini violenti di cui si sono resi protagonisti e la loro la tendenza a correre rischi (Ellis, 1987). Al contrario nel gruppo composto da offender non psicopatici non è emersa alcuna correlazione significativa.

Per quanto riguarda l'output fornitoci dall'albero di decisione, si evince che la differenza di polarità è l'attributo più discriminante, in grado di riconoscere da solo quasi il 50% degli individui affetti da psicopatia. Ancora più interessante è il valore per il quale avviene la suddivisione del campione. In particolare, se la differenza di polarità è maggiore di -0,125, l'individuo viene classificato come affetto da psicopatia. Questo risultato è molto rilevante in quanto la differenza di polarità nel campione assume valori nell'intervallo [-0,625; 0,25]. Questo significa che le narrazioni neutre, che hanno una differenza vicina allo zero, risultano scritte da soggetti affetti da psicopatia, in linea con l'ipotesi che gli psicopatici abbiano un ricordo povero di dettagli emotivi che lascia ragionevolmente presupporre un'attenuata risposta emozionale agli eventi.

È anche importante sottolineare come la polarità sia sbilanciata verso la negatività. Ciò suggerisce che, in assenza di psicopatia, il tipo di ricordo associato all'evento criminoso risulta essere negativo. Continuando ad analizzare l'albero di decisione si scopre che i nodi successivi si basano sugli attributi di polarità massima positiva e negativa, e solo uno dei nodi fa riferimento all'etichetta "piacere" di WordNetAffect.

4 Conclusioni

In questo lavoro abbiamo cercato di capire se i risultati prodotti da un algoritmo di sentiment analysis possano in qualche modo essere associati alla psicopatia. In particolare, abbiamo analizzato la presenza di psicopatia, in un campione di offender detenuti poichè i tratti caratterizzanti di tale disturbo, distacco emotivo, mancanza di senso di colpa, stile di vita antisociale, è possibile riscontrarli negli offender detenuti per le peculiarità relative alla violenza degli atti esperiti e allo stile di vita condotto.

I risultati ottenuti dalla sentiment analysis sono stati confrontati con quelli ottenuti attraverso il test PPI-R che valuta i tratti di personalità del sog-

[3] `http://www.cs.waikato.ac.nz/ml/weka/`

getto attraverso l'indagine di quei comportamenti e peculiarità cognitive, emotive e percettive che costituiscono i tratti tipici della psicopatia.

I risultati ottenuti attraverso l'utilizzo di un approccio basato sull'induzione di alberi di decisione dimostrano una dipendenza tra la presenza di psicopatia e la differenza tra la massima polarità positiva e quella negativa individuate nella narrazione del detenuto. Inoltre, è stata riscontrata una correlazione tra lo score massimo di polarità e il fattore *Dominanza priva di paura* attraverso l'utilizzo di test statistici.

I risultati ottenuti sono incoraggianti e dimostrano l'efficacia di ampliare la ricerca di base in psicologia cognitiva con strumenti innovativi come quello della sentiment analysis e in genere dell'analisi automatica del testo. Lo studio qui riportato, seppur rilevante, è comunque da considerarsi preliminare per due ragioni: la ridotta dimensione del campione e l'utilizzo di tecniche abbastanza semplici per l'analisi del sentiment.

Acknowledgement

Questo lavoro è parzialmente supportato dal progetto "Multilingual Entity Liking" finanziato dalla Regione Puglia con il programma FutureInResearch.

References

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta.

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. *Natural Language Engineering*, 11(03):247–261.

Ashley N Burrow, Nichole Currence, Diana Lemus, Amber E DeBono, Matthew T Crawford, and W Richard Walker. 2014. Psychopaths view autobiographical memories as less memorable, important, and emotional than normal individuals. *International Journal of Humanities and Social Science*, 4:1–9.

Sven-Åke Christianson, Adelle E Forth, Robert D Hare, Catherine Strachan, Lars Lidberg, and Lars-Håkan Thorell. 1996. Remembering details of emotional events: A comparison between psychopathic and nonpsychopathic offenders. *Personality and Individual Differences*, 20(4):437–443.

H Cleckley. 1941. The mask of sanity. st. louis. *MO: Mosby*.

Lee Ellis. 1987. Relationships of criminality and psychopathy with eight other apparent behavioral manifestations of sub-optimal arousal. *Personality and Individual Differences*, 8(6):905–925.

Samantha J Glass and Joseph P Newman. 2006. Recognition of facial affect in psychopathic offenders. *Journal of abnormal psychology*, 115(4):815.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Hugues Hervé, Barry S Cooper, and JOHN C Yuille. 2007. Memory formation in offenders: Perspectives from a biopsychosocial model of eyewitness memory. *Offenders' memories of violent crimes*, pages 38–74.

Scott O Lilienfeld and Michelle R Widows. 2005. *PPI-R: Psychopathic personality inventory revised: Professional Manual*. Psychological Assessment Resources, Incorporated.

Takahiro Osumi and Hideki Ohira. 2010. The positive side of psychopathy: Emotional detachment in psychopathy and rational decision-making in the ultimatum game. *Personality and individual differences*, 49(5):451–456.

J Ross Quinlan. 1993. *C4. 5: programs for machine learning*. Morgan Kaufmann Publishers.

Ralph C Serin. 1996. Violent recidivism in criminal psychopaths. *Law and Human Behavior*, 20(2):207.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet Affect: an Affective Extension of WordNet. In *LREC*, volume 4, pages 1083–1086.

Kevin Wilson, Sabrina Demetrioff, and Stephen Porter. 2008. A pawn by any other name? social information processing as a function of psychopathic traits. *Journal of research in personality*, 42(6):1651–1656.

Yaling Yang, Adrian Raine, Patrick Colletti, Arthur W Toga, and Katherine L Narr. 2010. Morphological alterations in the prefrontal cortex and the amygdala in unsuccessful psychopaths. *Journal of abnormal psychology*, 119(3):546.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! A free corpus-based morphological resource for the Italian language. *Corpus Linguistics 2005*, 1(1).

Stepping out of the Chinese Room: Word meaning with and without consciousness

Roberto Bottini
Center for Mind/Brain Sciences,
University of Trento, Italy
bottini.r@gmail.com

Daniel Casasanto
University of Chicago, IL
USA
casasanto@uchicago.edu

Andrea Nadalini
International School for Advanced Studies (SISSA)
Trieste, Italy
anadalini@sissa.it

Davide Crepaldi
Milan Center for Neuroscience
Milan, Italy
davide.crepaldi@sissa.it

Abstract

English. What is the role of consciousness in language processing? Unconscious priming experiments show that words can prime other words with related meanings (cat – dog), and these priming effects are assumed to reflect the activation of conceptual knowledge in semantic memory. Alternatively, however, unconscious priming effects could reflect predictive relationships between the words' forms, since words that are semantically related are also statistically related in language use. Therefore, unconscious "semantic" priming effects could be due to relationships between words' forms mimicking conceptual relationships, as in Searle's Chinese Room thought experiment. To distinguish word-form-based and semantics-based accounts of priming we conducted an experiment in which temporal words (e.g., earlier, later) were preceded by spatial words that were processed either consciously or unconsciously. Time is typically conceptualized as a spatial continuum extending along either the sagittal (front-back) or the lateral (left-right) axis, but only the sagittal space-time mapping is encoded in language (e.g. the future is *ahead*, not *to the right*). Results showed that temporal words were primed both by sagittal words (back, front) and lateral words (left, right) when primes were perceived consciously, as predicted by both wordform-based and semantics-based accounts. Yet, only sagittal words produced an unconscious priming effect, as predicted by the word-form-based account. Unconscious word processing appears to be limited to relationships between words' forms, and consciousness may be needed to activate words' meanings.

Italiano. *Qual è il ruolo della coscienza nell'elaborazione semantica delle parole? Esperimenti di masked priming semantico mostrano che la vista di una parola può facilitare il riconoscimento di un'altra parola dal contenuto semantico simile (gatto – cane). Questo effetto di priming è solitamente interpretato come evidenza che la parola inconscia è processata a livello semantico. Tuttavia, tale effetto può essere spiegato anche sulla base di relazione tra forme lessicali (senza attivazione di informazione nella memoria semantica). Infatti, parole che sono semanticamente legate sono anche legate statisticamente nel linguaggio. Il priming semantico inconscio potrebbe semplicemente emulare relazioni concettuali, come nel famoso esperimento mentale della stanza cinese di Searle. Per distinguere il priming lessicale dal priming semantico abbiamo condotto un esperimento in cui parole temporali (ieri, domani) erano precedute da parole spaziali mostrate sia a livello subliminale che supraliminale. Il tempo è tipicamente concettualizzato attraverso mappe spaziali che si estendono lungo l'asse sagittale (il passato è dietro, e il futuro davanti) e lungo l'asse laterale (il passato è a sinistra, futuro a destra). Solo la mappatura sagittale è però codificata nel linguaggio (il futuro è davanti, non a destra). I risultati mostrano come sia le parole sagittali (dietro, davanti) che quelle laterali (sinistra, destra) facilitano l'elaborazione di parole temporali (ad esempio prima e dopo), quando percepite consciamente. Al contrario, quando i prime sono elaborati al di fuori della coscienza, l'effetto sull'asse laterale viene meno. Il processo inconscio delle parole sembra dunque essere limitato a relazioni tra forme lessicali; la coscienza*

potrebbe essere necessaria per attivarne il significato.

1 Introduction

What role does consciousness play in word meaning's construction? As previous literature has pointed out, lexical items seem to be processed up to the semantic level even when processed out of awareness (Quinn & Kinoshita, 2008; Ansorge, Kiefer, Khalid, Grassl, & König, 2010). Most evidence for this claim comes from masked priming: When two words are sequentially presented, the recognition of the latter is made easier if the two are semantically related (cat-dog), even when the visibility of the former (the prime) is prevented by displaying it very briefly, embedded between visual masks (Forster, 2006; Dehaene et al., 1998). For instance, participants are likely to classify more quickly the word *dog* as referring to a living entity when it is preceded by the semantically related word *cat*, rather than by a semantically unrelated word like *apple*. As similar effects are attested when the prime word is clearly visible, it has been suggested that lexical items can be processed up to the semantic level irrespective of their visibility. We will refer to this perspective as the *semantic-based account* of masked priming, as it assumes that words are processed beyond their surface structure and activate conceptual knowledge about their referents. Such knowledge is thought to be stored within the semantic memory, where concepts and concepts' features are represented in an interconnected network (Tulving, 1972; Masson, 1995). In such a view, *cat* would prime *dog* as both words refer to mammals that have four legs, have a tail, can be pet, and so on; and therefore they are more closely related to each other than to *apple*.

However, there is an aspect of semantic similarity that has been largely overlooked in the priming-related literature, that words with similar meaning tend to have a similar contextual distribution. As corpus-based studies have pointed out, words referring to entities with similar perceptual and conceptual attributes tend be used together (e.g., *dog* and *cat* are more likely to co-occur in the same sentence than *dog* and *apple*), and to be used in similar contexts (e.g., both *cat* and *dog* tend to appear when speaking about pets, whether or not they co-occur within a given utterance; Louwerse, 2011; Landauer & Dumais, 1997). Based on this fact, unconscious priming may be alternatively explained through predictive relationships between words' forms established in language use. According to the *word-form-based account*, *cat* would prime *dog simply* because the two words share a similar contextual distribution. No conceptual representation is involved, as the locus of the unconscious semantic priming would be the lexical system, not semantic memory (Collins & Loftus, 1975).

This latter interpretation of unconscious word processing somehow resembles the Chinese Room thought experiment developed by the philosopher John Searle, where an English-speaking man is closed in a room receiving message written in Chinese characters (Searle, 1980). Due to a set of norms that determine the relationships between those characters (if you see X followed by Y, than reply Z), he is able to provide answers that would look perfectly appropriate to a native speaker. From the outside, it would appear that the man has a good understanding of the language, while instead he is acting on the basis of associations between word forms. Indeed, if he received a message saying that the room is about to explode, he would reply appropriately; but would not leave the room.

For the *semantic-based* and *wordform-based* accounts of masked priming to be distinguished, it is necessary to find concepts that are related in the semantic system, but not in the lexical system. This is the case for the metaphorical relationship linking time to space. The two domains are strictly intertwined in the human mind, in such a way that space is often used to think about time (Lakoff & Johnson, 1980). Time conceptualization involves both the sagittal and the lateral axis (Casasanto & Bottini, 2014; Bonato, Zorzi & Umiltà, 2012). For example, participants are faster in responding to past-related words by providing a leftward response, and to future-related words by providing a rightward response, relative to the opposite pattern. The same holds for the sagittal arrangement, with backward response associated with past-related words and forward response associated with future-related words. Moreover, neurological evidence shows that patients with hemispatial neglect have also impairments in temporal judgments (i.e. if they neglect the left side on space, they also show worst memory for past-related events; Saj, Fuhrman, Vuilleumier, & Boroditsky, 2013). Finally, people have been found to use hand gestures along both the lateral and the sagittal line when speaking about time (Casasanto & Jasmin,

2012). Critically, while the sagittal mapping is linguistically encoded in sentences such as "a bright future *in front of* you", the lateral one is not. The existence of these two mental timelines, and the fact that only one of them is linguistically expressed, offers the ideal test-bed for contrasting the *semantic-based* and *wordform-based* accounts of unconscious word processing. The latter predicts that space–time priming would only emerge along the linguistically encoded sagittal axis when primes are kept outside of awareness; while the former would predict priming to emerge along both axes, both supra- and subliminally.

We tested these predictions in a priming study with spatial words related to the lateral (*left-right*) and the sagittal (*ahead-behind*) axis as primes, and temporal words referring to either the past (*yesterday*) or the future (*tomorrow*) as target stimuli. In the first experiment, primes were clearly visible In the second experiment, prime visibility was prevented by means of a masking procedure.

2 Experiments

2.1 Experiment 1a - Visible primes

Participants: 60 volunteers were recruited for the experiment; all subjects were right-handed, and they all stated being native Italian speakers, with normal or corrected-to-normal vision and no history of neurological disorders. Each subject gave written informed consent for participation.

Stimuli, apparatus and procedure: Primes were 2 spatial words related to the lateral axis ("sinistra", *left*, and "destra", *right*) and 2 spatial words related to the sagittal axis ("davanti", *front*, and "dietro", *back*). Target stimuli were 8 temporal words. Four of them refer to the past ("prima", *earlier*, "ieri", *yesterday*, "passato", *past*, "scorso", *previous*), and four refers to the future ("dopo", *later*, "domani", *tomorrow*, "futuro", *future*, "successivo", *next*).

Each trial consisted of a fixation point (+) displayed for 750 ms. Then a blank screen was shown for 200ms, followed by the prime and by another blank screen, both lasting 50 ms. Finally, the target word was presented for 1500 ms, or until a response was provided.

Participants engaged in a GO/NO_GO task: They had to press a key if the target word referred to the past and do nothing if the target word referred to the future, or vice versa, according to the block instructions.

Results and discussion: analyses were conducted only on "GO" trials. Inaccurate trials (less than 1%) were excluded. In order to reduce the effect of outliers, those individual datapoints standing at more than 2 standard deviations from each participant's mean (~5% of the correct trials) were also removed from the analyses. A 2–by–2 ANOVA on the log-transformed RTs revealed a significant main effect of Congruity, $F(1, 59)= 11.47$, $p= 0.001$, indicating that participants were faster in congruent trials (535 ms) compared to incongruent ones (540 ms). We found no effect of Axis, $F(1, 59)= 0.41$, $p> .250$, and no Axis by Congruity interaction, $F(1, 59)= 0.06$, $p>0.250$. Pairwise comparisons showed that the priming effect was *significant* both in the sagittal (4 ms; $F(1, 59)= 5.79$, $p= 0.02$) and the lateral axis (6 ms; $F(1, 59)= 6.76$, $p= 0.01$).

Thus, significant congruity effects were produced both by sagittal and lateral spatial prime words, consistent with previous studies that provide evidence for sagittal and lateral mental timelines.

2.2 Experiment 1b - Subliminal primes

Participants: 60 volunteers from the same population as in Experiment 1a were recruited into the experiment. None of them took part in the previous experiment.

Stimuli, apparatus and procedure were the same as in Experiment 1a with one exception, i.e the blank screens that were displayed before and after the prime word were replaced with two visual masks in order to make the prime invisible (subjects were not informed of the presence of the primes).

Prime visibility task: after the end of the last part of the experiment, all subjects were informed about the presence of the prime word between the masks. Then, they performed a prime visibility test that included 10 practice and 128 experimental trials. The stimuli to be detected were the same spatial words we used in the previous experiment in half of the trials, and a string of identical lowercase letters (<aaaaaaaa>) in the other half.

Results and discussion: only the "GO" trials, in which participants provided a response, were analyzed. Inaccurate trials (less than 1%) were excluded. In order to reduce the effect of extremely long and short RTs, those individual datapoints standing at more than 2 standard deviations from each participant's mean (~4% of the correct trials) were also removed from the analyses.

A 2–by–2 ANOVA on the log-transformed RTs revealed a significant main effect of Congruity, $F(1, 58)= 27.63$, $p< .001$, and an Axis by Congruity interaction, $F(1, 58)= 14.986$, $p< 0.001$, which we followed up through pairwise comparisons showing that priming was significant for the sagittal axis (9 ms; $F(1, 58)= 40.21$, $p< 0.001$), but not for the lateral axis (2 ms; $F(1, 58)= 1.52$, $p= 0.22$).

No participant reported having noticed the prime. Overall, the average d-prime value was 0.33 (SD= 0.37). Although significantly different from zero, $t(58)= 7.03$, $p< 0.001$, this value is widely taken to indicate that the prime was effectively masked from perceivers' awareness (Kouider & Dupoux, 2005).

Experiment 1b clearly suggests that spatio–temporal masked priming is limited to the sagittal axis, with no apparent effect on the lateral axis. Thus, the pattern of results provides evidence in favor of the *wordform-based* account of unconscious word processing.

3 Conclusion

In this study we looked at the nature of word processing with and without awareness. Using the relationships between space and time, we were able to disentangle the *wordform-based* from the *semantic-based* account of masked priming. When words were clearly visible, we found priming effect on both axes, which reflects the sagittal, linguistically encoded, timeline, as well as the lateral mapping, which relies only on conceptual knowledge. Conversely, subliminal priming was obtained only with the sagittal words, matching the predictions of the *wordform-based* account. Therefore, our data suggests that when people read words unconsciously, activation spreads only between predictively related wordforms.

Unconscious priming between semantically related words may mimic semantic priming, much in the same way as the man inside Searle's Chinese Room mimics knowledge of Chinese, on the basis of "meaningless" wordform-wordform relationships. Unconscious word processing appears to be limited to relationships between words' forms, and consciousness may be needed to activate words' meanings.

Reference

Ansorge, U., Kiefer, M., Khalid, S., Grassl, S., & König, P. (2010). Testing the theory of embodied cognition with subliminal words. *Cognition*, 116(3),303–320.

Bonato, M., Zorzi, M., & Umiltà, C. (2012). When time is space: evidence for a mental time line. *Neuroscience and Biobehavioral Reviews*, *36*(10), 2257–73.

Casasanto, D., & Bottini, R. (2014). Mirror Reading Can Reverse the Flow of Time. *Journal of Experimental Psychology. General*, *143*(2).

Casasanto, D., & Jasmin, K. (2012). The Hands of Time: Temporal gestures in English speakers. *Cognitive Linguistics*, *23*(4), 643 – 674.

Collins, A. M., & Loftus, E. F. (1975). A Spreading-Activation Theory of Semantic Processing, *Psych Review*, 82(6), 407–428.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., Van de Mortele, P-F, & Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, *395*(6702), 597–600.

Fodor, J. (1983). *The modularity of mind*. Cambridge. MIT Press.

Forster, K. I. (2006). Early activation of category information in visual word recognition More on the turple effect, *The Mental Lexicon*, 1, 35–58.

Kouider, S., & Dupoux, E. (2005). Subliminal speech priming. *Psychological Science*, 16(8), 617–25.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: University of Chicago Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato ' s problem : The Latent Semantic Analysis Theory of Acquisition , Induction , and Representation of Knowledge. *Psychological Review*, 104(2), 211–240.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2), 273–302.

Masson, M. E. J. (1995). A Distributed Memory Model for Semantic Priming. Journal of Experimental Psychology.

Quinn, W. M., & Kinoshita, S. (2008). Congruence effect in semantic categorization with masked primes with narrow and broad categories. *Journal of Memory and Language*, 58(2), 286–306.

Saj, A., Fuhrman, O., Vuilleumier, P., & Boroditsky, L. (2013). Patients With Left Spatial Neglect Also Neglect the "Left Side" of Time. *Psychological Science*, (November).

Searle, J. R. (1980). *Minds , brains , and programs*, *BBS*, 3, 1–19.

Tulving, E. (1972). *Episodic and semantic memory*. In Organization of memory (Vol. 381). London: Academic.

Context-aware Spoken Language Understanding for Human Robot Interaction

Andrea Vanzo(†), **Danilo Croce**(‡), **Roberto Basili**(‡) **and Daniele Nardi**(†)
(†)Sapienza University of Rome
Department of Computer, Control and Management Engineering "Antonio Ruberti"
Via Ariosto 25, 00185 Roma, Italy
(‡)University of Roma, Tor Vergata
Department of Enterprise Engineering, Via del Politecnico 1, 00133 Roma, Italy
{vanzo,nardi}@dis.uniroma1.it, {croce,basili}@info.uniroma2.it

Abstract

English. Robots operate in specific environments and the correct interpretation of linguistic interactions depends on physical, cognitive and language-dependent aspects triggered by the environment. In this work, we present LU4R - *adaptive spoken Language Understanding 4 Robots*, a Spoken Language Understanding chain for the semantic interpretation of robotic commands, that is sensitive to the operational environment. The system has been designed according to a Client/Server architecture in order to be easily integrated with the vast plethora of robotic platforms.

Italiano. *L'interpretazione di comandi espressi nei confronti di piattaforme robotiche è un processo strettamente legato al contesto operativo in cui avviene l'interazione. In questo lavoro, presentiamo LU4R - adaptive spoken Language Understanding 4 Robots, un sistema per l'elaborazione automatica di comandi vocali, dipendente dall'ambiente in cui il comando viene espresso. Il sistema proposto, implementato come una cascata di passi di elaborazione semantica, è stato progettato seguendo un'architettura Client/Server, per ridurre i requisiti di integrazione con le piattaforme robotiche esistenti.*

1 Introduction

End-to-end communication in natural language between humans and robots is challenging for the different cognitive abilities involved during the interaction. As an example, for a robot to react to a command like "*take the book on the table*", a number of implicit assumptions should be met. First, at least two entities, a `book` and a `table`, must exist in the environment and the speaker must be aware of such entities. Accordingly, the robot must have access to an inner representation of the objects, e.g., an explicit map of the environment. Second, mappings from lexical references to real world entities must be developed or made available. In this respect, the *Grounding* process (Harnad, 1990) links symbols (e.g., words) to the corresponding perceptual information. Hence, robot interactions need to be *grounded*, as meaning depends on the state of the physical world and the interpretation crucially interacts with perception, as pointed out by psycho-linguistic theories (Tanenhaus et al., 1995). To this end, the integration of perceptual information derived from the robot's sensors with an ontologically motivated description of the world has been adopted as an augmented representation of the environment, in the so-called *semantic maps* (Nüchter and Hertzberg, 2008). In this maps, the existence of real world objects can be associated to *lexical* information, in the form of entity names given by a knowledge engineer or spoken by a user for a pointed object, as in Human-Augmented Mapping (Diosi et al., 2005). While SLU for Interactive Robotics have been mostly carried out over the only evidences specific to the linguistic level (see, for example, (Chen and Mooney, 2011; Matuszek et al., 2012)), we argue that such process should be context-aware, in the sense that both the user and the robot access and make references to a shared environment. For example, in the above command, "*taking*" is the intended action whenever a book is actually on the table, so that "*the book on the table*" refers to a single argument. On the contrary, the command may refer to a "*bringing*" action, when no book is on the table and *the book* and *on the table* correspond to different semantic roles.

In this paper, we present LU4R an adaptive spoken language understanding chain for the auto-

matic interpretation of robotic spoken commands that is coherent with the above assumptions. The resulting chain is based on the approach proposed in (Bastianelli et al., 2016) that allows to produce interpretations that are consistent with (*i*) the world (with all the entities composing it), (*ii*) the Robotic Platform (with all its inner representations and capabilities), and (*iii*) the linguistic information derived from the user's utterance. LU4R is fully implemented in Java and is released according to a Client/Server architecture, in order to decouple the chain from the specific robotic platform that will use it. It receives as input one or more transcriptions of a spoken command and produces one or more linguistic predicates reflecting the actions intended by the user. Predicates, as well as their arguments, are consistent with a linguistically-motivated representation and coherent with the environment perceived by the robot.

The rest of the paper is structured as follows. Section 2 provides an architectural description of the entire system, as well as an overall introduction about its integration with a generic robot. In Section 3 we demonstrate the applicability of the chain in the interpretation of commands in English and Italian.

2 The overall architecture

The architecture of the proposed system is decoupled into two main macro-components, as shown in Figure 1: the *Robotic Platform* and *LU4R*.

The Client-Server communication schema between the Robotic Platform (the Client) and LU4R (the Server) allows to maintain the former independent from the latter. It is obvious that the interpretation process must be achieved even when no information about the domain/environment is available, i.e., a scenario involving a *blind* but speaking robot. This is the case when the command "*take the book on the table*" is paired with any additional information and the ambiguity with respect to the evoked predicate, i.e., *Taking* vs. *Bringing*, cannot be resolved. At the same time, the platform allows to specialize the semantic interpretation process to individual situations when contextual information is available. In this case, whenever the sentence "*take the book on the table*" is provided along with information about the presence and position of a `book` on a `table`, the above disambiguation can be solved.

In the following, each macro-component of the architecture in Figure 1 is discussed and analyzed.

2.1 The Robotic Platform

The overall architecture contemplates a generic Robotic Platform, whose task, domain and physical setting are not necessarily specified. In order to make LU4R independent from the above specific aspects, we will assume that the platform requires at least the following modules: (*i*) an Automatic Speech Recognition (ASR) system; (*ii*) a SLU Orchestrator; (*iii*) a Grounding and Command Execution Engine; (*iv*) a Physical Robot. Additionally, the optional component *Support Knowledge Base* is expected to provide the contextual information discussed above. While the discussion about the Physical Robot is out of the scope of this work, all the other components are hereafter shortly summarized.

ASR system. An ASR engine allows to transcribe a spoken utterance into one or more possible transcriptions. In the actual release, the ASR is here performed through an *ad-hoc* Android application that can be deployed on both Android smartphones and tablets. It relies on the official *Google ASR API*[1] that offers valuable performances for an off-the-shelf solution. The acoustic model is here based on deep learning techniques, i.e. Recurrent Neural Networks (Hinton et al., 2012). The Google ASR is publicly available and is rather robust toward some of the complexities of spoken language, such as disfluencies and repetitions. This allowed us to focus on other challenges of spoken language, such as linguistic variations (e.g. synonymy, phonetically similar words, interrogative vs. imperative sentences, ...). Advantages of our lexicalized grounding approach (Bastianelli et al., 2015) is the robustness against variability and sense ambiguity.

SLU Orchestrator. The SLU Orchestrator implements a TCP Server for the Android App, here coded as a ROS node (Quigley et al., 2009) waiting for Client requests. Once a new request arrives (a list of transcriptions for a given spoken sentence), this module is in charge of extracting the perceived entities from a structured representation of the environment (here, a sub-component of the Support Knowledge Base) and sending the list of hypothesized transcriptions to LU4R along with the list of the perceived entities. The communication protocol requires the serialization of such

[1] `https://cloud.google.com/speech/`

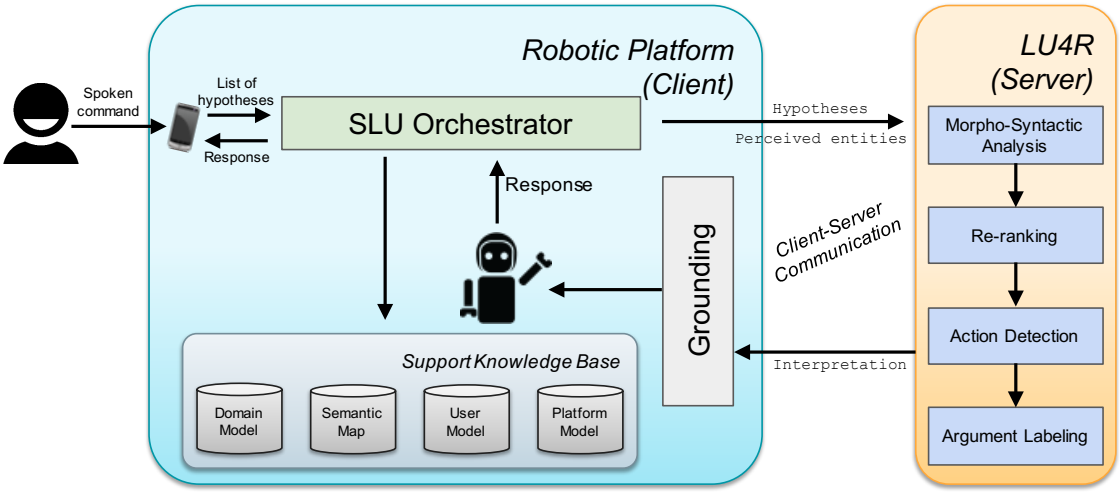


Figure 1: Overall architecture of the system

information in two different JSON objects. Even though this module is actually a TCP Server for the Android App, it represents also the Client interface toward LU4R.

Grounding and Command Execution. Even though the grounding process is placed at the end of the loop, it is discussed here as it represents part of the Robotic Platform. In fact, this process has been completely decoupled from the SLU, as it may involve perception capabilities and information unavailable to LU4R or, in general, out of the linguistic dimension. Nevertheless, this situation can be partially compensated by defining mechanisms to exchange some of the grounding information with the linguistic reasoning component. However, grounding is always carried out on board of the robot, as it represents the most general situation. The grounding carried out by the robot is triggered by a logical form expressing one or more actions through (linguistic) predicates. The output of the SLU process embodies the produced logical form: this latter exposes the recognized actions that are then linked to specific robotic operations (primitive actions or plans). Correspondingly, the predicate arguments (e.g., objects and location involved in the targeted action) are detected and linked to the objects/entities of the current environment. A fully grounded command is obtained through the complete instantiation of the robot action (or plan) and its final execution.

2.2 The LU4R system

The language understanding process produces an interpretation of the user's utterance in terms of linguistic predicates as defined in Frame Semantics (Fillmore, 1985). Specifically, we consider the formalization adopted in the FrameNet (Baker et al., 1998) database. According to such theory, actions expressed in user utterances can be modeled as *semantic frames*. These are micro-theories about real world situations, e.g., the action of *taking*. Each frame specifies also the set of participating entities, called *frame elements*, e.g., the THEME representing the object that is taken during the *Taking* action. For example, for the sentence "*take the book on the table*", a potential corresponding parsed version is:

$$[\textbf{\textit{take}}]_{Taking}\ [\textit{the book on the table}]_{\text{THEME}} \quad (1)$$

In a robotic perspective, semantic frames provide a cognitively sound bridge between the actions expressed in the language and the implementation of such actions in the robot world.

As shown in Figure 1, LU4R is composed of four main steps.

Morpho-syntactic analysis is performed over each available utterance transcription, by applying Part-of-Speech tagging and syntactic parsing, providing morphological and syntactic information, essential for further processing.

Whenever more than one hypothesized transcription is available, a **Re-ranking** module can be activated to evaluate a new sorting of the hypotheses, in order to get the best one out of the original ranking. It allows to reuse existent ASR solutions, making the final rank sensitive to the specific robotic domain.

The selected transcription is the input of the **Action Detection** (AD) component. Here all the

frames (i.e. intended actions) evoked in a sentence are detected, according to their triggering lexical units. For instance, given the above example, the AD would produce the following interpretation: $[\textbf{\textit{take}}]_{\textit{Taking}}$ *the book on the table*.

The final step is the **Argument Labeling** (AL). Here a set of frame elements is retrieved for each frame, detected during the AD step. Such process is, in turn, realized in two sub-steps. First, the *Argument Identification* (AI) aims at finding the spans of all the possible frame elements. Then, the *Argument Classification* (AC) assigns the suitable frame element label to each span identified during the AI, producing the final tagging shown in (1).

An off-the-shelf tool is used for the morpho-syntactic analysis, namely the Stanford CoreNLP suite (Manning et al., 2014). Re-ranking is performed using a learn-to-rank approach, where a Support Vector Machine exploiting a combination of linguistic kernels is applied, as discussed in (Basili et al., 2013). The AD, AI and AC steps are modeled as a sequential labeling task, as in (Bastianelli et al., 2016). The Markovian formulation of a structured SVM proposed in (Altun et al., 2003) is applied to implement the sequential labeler, known as *SVM*^{hmm} . In general, this learning algorithm combines a local discriminative model, which estimates the individual observation probabilities of a sequence, with a global generative approach to retrieve the most likely sequence, i.e. tags that better explain the whole sequence. In other words, given an input sequence $\mathbf{x} = (x_1 \dots x_l) \in \mathcal{X}$ of feature vectors $x_1 \dots x_l$, *SVM*^{hmm} learns a model isomorphic to a k-order Hidden Markov Model, to associate $\mathbf{x}$ with a set of labels $\mathbf{y} = (y_1 \dots y_l) \in \mathcal{Y}$.

A sentence s is here intended as a sequence of words w_i, each modeled through a feature vector x_i and associated to a dedicated label y_i, specifically designed for each interpretation process. During training, the SVM algorithm is devoted to associating words to step-specific labels: linear kernel functions are applied to different types of features, ranging from linguistic to perception-based features, and linear combinations of kernels are used to integrate independent properties. At classification time, given a sentence $s = (w_1 \dots w_{|s|})$, the *SVM*^{hmm} efficiently predicts the tag sequence $\mathbf{y} = (y_1 \dots y_{|s|})$ using a Viterbi-like decoding algorithm.

Both the re-ranking and the language understanding phases can work in two different settings.

In the so-called *basic* scenario, only linguistic information is used during the interpretation task. Perceptual information from the environment is thus neglected and evidences from the user's utterances or linguistic resources are considered.

Conversely, when perceptual information is made available to the chain, a context-aware interpretation is triggered. Such a perceptual knowledge is mainly exploited through a *linguistic grounding* mechanism (Bastianelli et al., 2015). This lexically-driven grounding is estimated through distances between filler (i.e. argument heads) and entity names. Such a semantic distance integrates metrics over word vectors descriptions and phonetic similarity. Word semantic vectors are here acquired through corpus analysis, as in Distributional Lexical Semantic paradigms (Turney and Pantel, 2010). They allow to map referential elements, such as lexical fillers, e.g. *desk*, to entities, e.g. a `table`, by thus modeling synonymy or co-hyponymy. Conversely, phonetic similarities ar smoothing factors against possible ASR transcription errors, e.g. *pitcher* and *picture*. Once links between fillers and entities have been activated, the sequential labeler is made sensitive to additional features, that inject perceptual information both in the learning and the tagging process, e.g. the presence/absence of referred objects in the environment. As a side effect, the above mechanism provides the robot with the set of linguistically-motivated groundings, that can be potentially used for any further grounding process.

Overall, the service provided by LU4R is performed as a black-box component, so that the complexity of each inner sub-task is hidden to the user. The service is realized through a server accepting connections on a predefined port. LU4R is entirely coded in Java and released as a single Jar file[2], along with the required folders containing linguistic models, configurations files and other resources. Hence, it can be run through command line, so that it is easier to integrate it within any architecture.

The LU4R system takes three input parameters: *type* of the understanding process, *output format* and *listening port*. The first parameter defines the type of the interpretation process to be initialized: the `basic` value activates the setting where only linguistic information is adopted, while `simple`

[2] `http://sag.art.uniroma2.it/sluchain.html`

refers to the interpretation where perceptual information is considered. The second parameter specifies the desired output format, e.g., *eXtended Dependency Graph* (Basili and Zanzotto, 2002) or the *Abstract Meaning Representation*, proposed in (Banarescu et al., 2013).

3 Evaluating LU4R

In order to provide evidences about the effectiveness of the proposed solution, we report here a preliminary evaluation of the interpretation process w.r.t. robotic commands in two languages, i.e. English and Italian.

	AD	**AI**	**AC**
English	95.91%	94.29%	95.31%
Italian	82.29%	79.46%	84.49%

Table 1: Experimental evaluation of the semantic interpretation process, in terms of F1

Table 1 shows results obtained over the Human Robot Interaction Corpus (HuRIC) (Bastianelli et al., 2014), a dataset of semantically annotated commands typical of Service Robotics. HuRIC contains 527 sentences in English. Moreover, results w.r.t. to a subset of 188 commands from HuRIC translated in Italian are reported. The results, expressed in terms of F1 measure, focus on the semantic interpretation process, in particular Action Detection (AD), Argument Identification (AI) and Argument Classification (AC) steps. In fact, F1 scores measures the quality of a specific module. While in the AD step the F1 refers to the ability to extract the correct frame(s) evoked by a sentence, in the AI step it evaluates to the correctness of the predicted argument spans. Finally, in the AC step the F1 measures the accuracy of the classification of individual arguments.

We tested each sub-module in isolation, feeding each step with gold information provided by the previous step in the chain. Moreover, the evaluation has been carried out considering the correct transcriptions, i.e., not contemplating the error introduced by the Automatic Speech Recognition system. The results over the Italian dataset refer to the `basic` setting of LU4R, i.e. leveraging just linguistic information. Conversely, the experiments over the English dataset have been carried out with the LU4R setting exploiting also perceptual knowledge. Results against the commands in English are encouraging for the application of LU4R in realistic applications, with a F1 higher than 94% in the recognition of semantic predicates used to express intended actions as well as the involved entities. We speculate that the gap w.r.t. results against the Italian dataset is mainly due to the lack of perceptual knowledge as well as the reduced size of the dataset. A more detailed description of the above evaluation over the English and Italian dataset is available in (Bastianelli et al., 2016) and (Vanzo et al., 2016), respectively.

4 Conclusions

In this paper, we presented LU4R, an adaptive SLU processing chain focused on the interpretation of commands in the Mobile Service Robotics domain. The proposed solution relies on Frame Semantics and Distributional Lexical Semantics to support example-driven machine learning algorithms that map individual sentence transcriptions to meaningful commands. Statistical learning, i.e. SVM^{hmm} , is applied by transforming the interpretation process into a cascade of sentence annotation tasks. The use of Frame semantics enables the reuse of large repositories of examples (i.e. (Baker et al., 1998)), supporting the recognition of up to more than 1,000 different semantic frames, currently defined in the FrameNet database. The robustness of the sentence interpretation, as measured in this paper, is rather good, where language variability is tackled by relying on distributional lexical models that generalize semantics for large vocabularies, well beyond the training set dictionaries. Moreover, even though LU4R is completely decoupled from the Robotic Platform, the final interpretation is made dependent on the robot's environment, by designing perceptual knowledge through feature modeling. Perceptual knowledge is derived from the robot's semantic map and translated into feature values. The corresponding space allow to synthesize information about existence and position of named entities, useful to disambiguate predicates and role assignment.

The results gathered during repeated empirical investigation campaigns confirm the effectiveness of the proposed tool and, in particular, the benefits provided by the injection of perceptual Knowledge into the understanding (i.e. labeling) process. LU4R is thus an effective example of grounded language learning framework, whereas

predicates and semantic roles are acquired through an integration between linguistic (e.g. lexical and grammatical) properties and perceptual information (e.g. distances between entities in a map): this makes LU4R an interesting topic for future research, such as the extension of command interpretation process to complex interaction patterns, such as in interactive question answering or dialogue.

References

Yasemin Altun, I. Tsochantaridis, and T. Hofmann. 2003. Hidden Markov support vector machines. In *Proc. of ICML*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of ACL and COLING*, pages 86–90.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August. Association for Computational Linguistics.

Roberto Basili and Fabio Massimo Zanzotto. 2002. Parsing engineering and empirical robustness. *Nat. Lang. Eng.*, 8(3):97–120, June.

Roberto Basili, Emanuele Bastianelli, Giuseppe Castellucci, Daniele Nardi, and Vittorio Perera. 2013. Kernel-based discriminative re-ranking for spoken command understanding in hri. In *AI*IA*, volume 8249, pages 169–180. Springer.

Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, Roberto Basili, and Daniele Nardi. 2014. Huric: a human robot interaction corpus. In *Proceedings of LREC 2014*, Reykjavik, Iceland, may.

Emanuele Bastianelli, Danilo Croce, Roberto Basili, and Daniele Nardi. 2015. Using semantic models for robust natural language human robot interaction. In *AI* IA 2015, Advances in Artificial Intelligence*, pages 343–356. Springer International Publishing.

Emanuele Bastianelli, Danilo Croce, Andrea Vanzo, Roberto Basili, and Daniele Nardi. 2016. A discriminative approach to grounded spoken language understanding in interactive robotics. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York*.

David L. Chen and Raymond J. Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the 25th AAAI Conference on AI*, pages 859–865.

Albert Diosi, Geoffrey R. Taylor, and Lindsay Kleeman. 2005. Interactive SLAM using laser and advanced sonar. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation, ICRA 2005, April 18-22, 2005, Barcelona, Spain*, pages 1103–1108.

Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254.

S. Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Geoffrey Hinton, Li Deng, Dong Yu, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath George Dahl, and Brian Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, November.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Cynthia Matuszek, Evan Herbst, Luke S. Zettlemoyer, and Dieter Fox. 2012. Learning to parse natural language commands to a robot control system. In Jaydev P. Desai, Gregory Dudek, Oussama Khatib, and Vijay Kumar, editors, *ISER*, volume 88 of *Springer Tracts in Advanced Robotics*, pages 403–415. Springer.

Andreas Nüchter and Joachim Hertzberg. 2008. Towards semantic maps for mobile robots. *Robot. Auton. Syst.*, 56(11):915–926.

Morgan Quigley, Ken Conley, Brian P. Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y. Ng. 2009. Ros: an open-source robot operating system. In *ICRA Workshop on Open Source Software*.

M. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, and J. Sedivy. 1995. Integration of visual and linguistic information during spoken language comprehension. *Science*, 268:1632–1634.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, January.

Andrea Vanzo, Danilo Croce, Giuseppe Castellucci, Roberto Basili, and Daniele Nardi. 2016. Spoken language understanding for service robotics in italian. In *15th International Conference of the Italian Association for Artificial Intelligence*, page to appear.

Argument Mining on Italian News Blogs

Pierpaolo Basile
University of Bari
pierpaolo.basile@uniba.it

Valerio Basile
Université Côte d'Azur, Inria,
CNRS, I3S, France
valerio.basile@inria.fr

Elena Cabrio, Serena Villata
Université Côte d'Azur, CNRS,
Inria, I3S, France
firstname.lastname@unice.fr

Abstract

English. The goal of *argument mining* is to extract structured information, namely the *arguments* and their *relations*, from unstructured text. In this paper, we propose an approach to argument relation prediction based on supervised learning of linguistic and semantic features of the text. We test our method on the CorEA corpus of user comments to online newspaper articles, evaluating our system's performances in assigning the correct relation, i.e., *support* or *attack*, to pairs of arguments. We obtain results consistently better than a sentiment analysis-based baseline (over two out three correctly classified pairs), and we observe that sentiment and lexical semantics are the most informative features with respect to the relation prediction task.

Italiano. *L'estrazione automatica di argomenti ha come scopo recuperare informazione strutturata, in particolare gli argomenti e le loro relazioni, a partire da testo semplice. In questo contributo proponiamo un metodo di predizione delle relazioni tra argomenti basato sull'apprendimento supervisionato di feature linguistiche e semantiche del testo. Il metodo è testato sul corpus di commenti di news CorEA, ed è valutata la capacità del sistema di classificare le relazioni di supporto ed attacco tra coppie di argomenti. I risultati ottenuti sono superiori ad una baseline basata sulla sola analisi del sentimento (oltre due coppie di argomenti su tre è classificata correttamente) ed osserviamo che il sentimento e la semantica lessicale sono gli indicatori più informativi per la predizione delle relazioni tra argomenti.*

1 Introduction

The argument mining (Peldszus and Stede, 2013; Lippi and Torroni, 2016) research area has recently become very relevant in computational linguistics. Its main goal is the automated extraction of natural language arguments and their relations from generic textual corpora, with the final goal of providing machine-readable structured data for computational models of argument and reasoning engines. Two main stages have to be considered in the typical argument mining pipeline, from the unstructured natural language documents towards structured (possibly machine-readable) data: *(i)* argument extraction, i.e., to detect arguments within the input natural language texts, and *(ii)* relation extraction, i.e., to predict what are the relations holding between the arguments identified in the first stage. The relation prediction task is extremely complex, as it involves high-level knowledge representation and reasoning issues. The relations between the arguments may be of heterogeneous nature, like attack, support or entailment (Cabrio and Villata, 2013).

The increasing amount of data available on the Web from heterogeneous sources, e.g., social network posts, forums, news blogs, and the specific form of language adopted there challenge argument mining methods, with the aim to support users to understand and interact with such a huge amount of information.

In this paper, we address this issue by presenting an argument relation prediction approach for Italian. We test the method on the CorEA corpus (Celli et al., 2014) of user comments to the news articles of an Italian newspaper, annotated with *agreement* (i.e., support) and *disagreement* (i.e., attack) relations. We extract argument-level features from the CorEA comment (i.e., argument)

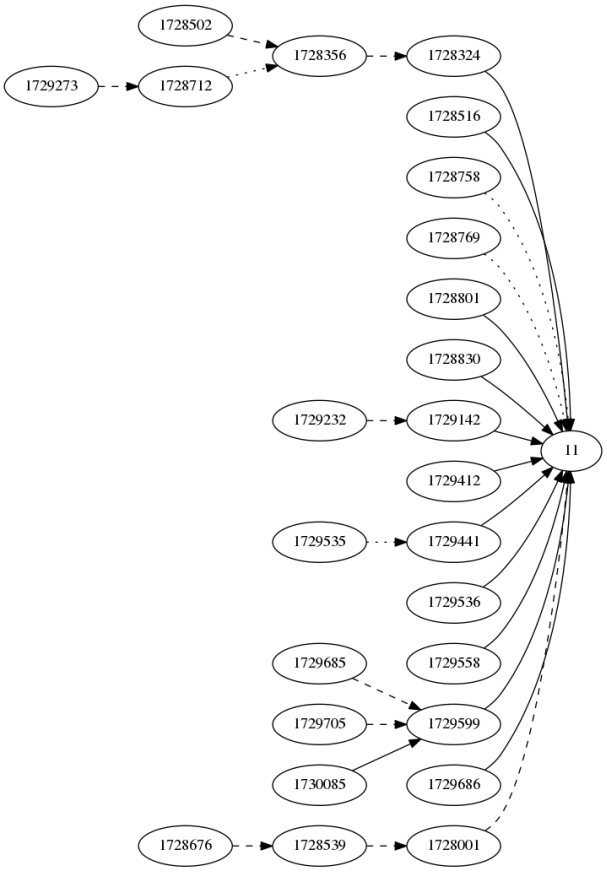


Figure 1: Example of debate structure.

pairs, and we train our system to predict the support and attack relations.

2 Mining Arguments

A debate, whether it happens online or in person, can be modeled as a set of *arguments* proposed by the *participants*. Arguments can be independent, for instance expressing the participant's stance on a particular topic, but often they are replies to previous arguments put forward in the debate. This results in a network structure of the debate, that is, a (possibly disconnected) directed graph where nodes are arguments, and the two kinds of edges are the support and attack relations between them. In Figure 1, each node represents an argument with a numeric identifier, filled and dashed edges represent respectively support and attack relations, and dotted edges are neutral relations. The hub-like node labeled 11 is a news article, thus attracting many first-level comments.

The goal of our work is to be able to predict the relations between the arguments in a given debate, thus reconstructing the relation graph. We therefore cast the problem as a classification task: given two arguments from a debate, we aim to predict whether one argument attacks the other, supports it, or there is no relation between the two arguments. The construction of the graph structure is then straightforward, resulting from the combination of all the argument pairs we considered.

2.1 Features

We extract argument-level features from the CorEA comment pairs, that we group into the following categories:

Lexical We take into account several lexical features: tokens, bi-grams, and the first bi-gram and tri-gram of each argument.

Syntactic We exploit the output of a dependency parser. We consider two kinds of dependency features: the former is the original output, the latter generalizes a word to its POS tag. For instance, "amod(denaro, pubblico)" is generalized as the "amod(NN, pubblico)" and "amod(denaro, ADJ)". We adopt the Malt parser (Nivre, 2003) trained on the Universal Dependency Treebank[1].

Message info We extract the argument size, the number of uppercase words, the number of negations[2], the number of sequences of two or more punctuation characters, the number of citations. A citation is a quoted sequence of words in the second argument that occurs in the first argument.

Message overlap Cosine similarity between two arguments is computed exploiting TF/IDF.

Word-embedding We build word-embeddings relying on the Paisà corpus through the word2vec (Mikolov et al., 2013) tool. We use a vector dimension equal to 50, and we consider only words that occur at least 20 times. For each argument, we use the vector components as features directly.

Sentiment We extract the sentiment from the arguments with two separate tools. Alchemy API[3], the sentiment analysis feature of IBM's Semantic Text Analysis API, returns a polarity label (positive, negative or neutral) and a

[1] http://universaldependencies.org/it/overview/introduction.html

[2] The occurrences of the word "non"

[3] http://www.alchemyapi.com/

polarity score between -1 (totally negative) and 1 (totally positive). The UNIBA system (Basile and Novielli, 2014), one of the most successful participants in the Sentipolc task at Evalita 2014 (Basile et al., 2014), returns a subjectivity label (subjective or objective) and a polarity label (positive, negative, neutral or mixed).

Topic model We train a domain-independent topic model for Italian and compute, for each argument, its representing vector in the topic space. The 300-dimensional topic model is created with Gensim[4] using the ItWaC corpus (Baroni et al., 2009). We use the vector components as features directly, i.e., each comment has 300 topic-based features.

3 Evaluation

The goal of the evaluation is twofold: *i)* to compute the performance of several machine learning methods and compare them with respect to some baselines, and *ii)* to investigate the importance of each group of features through an ablation test.

3.1 Data

We test our approach on the CorEA corpus (Celli et al., 2014), a collection of text from Italian news blogs. It contains 27 news articles, about 1,660 unique authors and more than 2,900 comments. The corpus is annotated with emotions and, most interestingly for our work, the comments are annotated pair-wise with agreement information (Celli et al., 2016). We extracted such comment pairs for a total of 1,275 pairs: 682 disagreement, 106 neutral, 180 agreement (307 pairs are not classified, examples in Figure 2).

The CorEA dataset provides several information about each message. Beside the features described in Section 2.1, we also extract the following dataset-dependent features: the set of manually annotated topics, the news category of the article, the count of replies to the message, the count of message likes, the participant's activity score, the participant's interests, the participant's page views, the participant's total comments, the participant's total shares, the participant's likes received, and the overall emotion declared by the participant after reading the articles.

[4] https://radimrehurek.com/gensim/

3.2 System setup

We exploit two kinds of learning algorithms: 1) different configurations of SVM based on linear kernel (SVM_{lin}), degree-2 polynomial kernel (SVM_{poly}), and RBF kernel (SVM_{rbf}); 2) Random Forest (RF).

The baseline method always predicts the most frequent class, in this case "attack". Moreover, we test the two simple sentiment analysis systems already described in 2.1, $SA_{alchemy}$ and SA_{uniba}. In particular, these systems exploit the result of the sentiment analysis in terms of polarity (positive, negative, or neutral) for predicting the relation between two arguments: if two arguments have the neutral polarity, they are tagged as neutral, while they are tagged as "support" in case they have the same polarity, otherwise the "attack" class is predicted. The system is implemented in JAVA relying on the Weka tool (Hall et al., 2009). All the experiments are performed by adopting the 10-folds cross-validation. For all the learning methods, we adopt the default Weka parameters since the goal of our work is not to optimize the classification performance but to provide a features study.

3.3 Results

Table 1 reports on the best results obtained by each method. Regarding RF the best result is obtained using 10 trees, while for SVM we optimize only the C parameter using default values for the other ones. The best C value for SVM_{lin} is 1, 2 in all the other settings.

Each one of the supervised systems performs better than the baseline. The good performance of the linear kernel classifier is likely to be ascribed to the high number of features. The performance of Random Forest is also quite good, considering that only ten trees are employed.

Table 1: Results

System	P	R	F
baseline	0.4964	0.7045	0.5824
$SA_{alchemy}$	0.3553	0.3616	0.3584
SA_{uniba}	0.2942	0.3286	0.3105
SVM_{lin}	0.6789	0.7169	**0.6719**
RF	0.6607	0.7180	0.6491
SVM_{poly}	0.6609	0.7097	0.6486
SVM_{rbf}	0.6414	0.7076	0.6120

As can be seen from the results of ablation tests (see Table 2), the features that contribute the most

Relation	Example
Attack	"in certi paesi 100 sterline a settimana permettono di vivere come un pascià" "si ma in certi altri no..;-) la cifra mi sembra davvero esigua.."
Support	"Caro Renzi , hai visto com'è semplice restituire i soldi? Basta una firmetta... perchè non lo fai anche tu invece di promettere e promettere e promettere?" "Bisogna prendere atto che il movimento 5 stelle sta davvero restituendo i soldi agli Italiani. Questo è un fatto, tutto il resto sono chiacchere."
Neutral	"E le riforme?" "le riforme cominciano dl'atteggiamento dei parlamentari. con il cambiamento del mind-set . il punto di partenza."

Figure 2: Examples of relations between pairs of comments in CorEA.

to the argument classification task are the semantic features (i.e., embeddings) and the sentiment features. This confirms our hypothesis that sentiment is a key information for argument mining, and more specifically for the relation prediction task. The results also confirm that lexical and semantic features are useful for the task, as expected. Table 2 reports also the number of features (Feat.Size) and the F1 (F1-f) achieved by exploiting the respective feature in isolation. It is important to note that, despite the bad performance obtained by both embedding and sentiment features, their contribution in the overall performance is relevant.

Table 2: Ablation test

Features	F1	$\Delta\%$	Feat.Size	F1-f
all	0.6719	-	220,499	-
-lexical	0.6624	-1.42	140,443	0.66
-syntactic	0.6702	-0.26	80,909	0.65
-info	0.6691	-0.42	220,490	0.58
-CorEA	0.6674	-0.68	220,218	0.64
-embedding	0.6525	-2.89	220,399	0.59
-overlap	0.6724	0.07	220,498	0.58
-sentiment	0.6622	-1.45	220,491	0.58
-topic	0.6673	-0.69	220,045	0.59

4 Related Work

(Lippi and Torroni, 2016) and (Peldszus and Stede, 2013) provide an overview about the argument mining research area. In particular, some approaches have been recently proposed to address the same task addressed in this paper, i.e. predicting relations between arguments, even if ours is the first effort for the Italian language. (Aharoni et al., 2014) assume that evidence is always associated with a claim, enabling the use of information about the claim to predict the evidence. The support relations are thus obtained by definition when predicting the evidence. (Mochales and Moens, 2011) have addressed the problem by parsing with a manually-built context-free grammar to predict relations between argument components. The grammar rules follow the typical rhetorical and structural patterns of sentences in juridical texts. This is a highly genre-specific approach, and its direct use in other genres would be unlikely to yield accurate results. (Stab and Gurevych, 2014) instead employ a binary SVM classifier to predict relations in a claim/premise model. (Biran and Rambow, 2011) apply the same method adopted for the detection of premises also for the prediction of relations between premises and claims. (Wang and Cardie, 2014) apply an isotonic Conditional Random Fields based sequential model to make predictions on sentence- or segment-level on discussions on Wikipedia Talk pages. Finally, (Cabrio and Villata, 2013) adopt Textual Entailment to infer whether a support or attack relation between two given arguments holds.

5 Conclusions

In this paper, we have presented a supervised approach for argument relation prediction for Italian, mainly relying on features including semantics and sentiment. We tested such approach on the CorEA corpus, extracted from user comments to online news. Our experimental results are good, and foster future research in the direction of including semantics as well as sentiment analysis in the argument mining pipeline. It will be also interesting, as future work, to refine the model in order to consider the full sequence of interactions between arguments.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland, June. Association for Computational Linguistics.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, Eros Zanchetta, Springer, and Science+business Media B. V. 2009. The wacky wide web: A collection of very large linguistically processed webcrawled corpora. language resources and evaluation.

Pierpaolo Basile and Nicole Novielli. 2014. Uniba at evalita 2014-sentipolc task: Predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features. *Proceedings of EVALITA*, pages 58–63.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, Pisa, Italy.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs by classifying text as argumentative. *Int. J. Semantic Computing*, 5(4):363–381.

Elena Cabrio and Serena Villata. 2013. A natural language bipolar argumentation approach to support users in online debate interactions†. *Argument & Computation*, 4(3):209–230.

Fabio Celli, Giuseppe Riccardi, and Arindam Ghosh. 2014. Corea: Italian news corpus with emotions and agreement. In *CLIC-it 2014*, pages 98–102.

Fabio Celli, Giuseppe Riccardi, and Firoj Alam. 2016. Multilevel annotation of agreement and disagreement in italian news blogs. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Trans. Internet Techn.*, 16(2):10.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop at ICLR, 2013*.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *IJCINI*, 7(1):1–31.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 46–56.

Lu Wang and Claire Cardie. 2014. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 97–106, Baltimore, Maryland, June. Association for Computational Linguistics.

Twitter Sentiment Polarity Classification using Barrier Features

Anita Alicante, Anna Corazza, Antonio Pironti
Department of Electrical Engineering and Information Technologies (DIETI)
Università di Napoli Federico II
via Claudio 21, 80125 Napoli, Italy
anita.alicante@unina.it, anna.corazza@unina.it,
antonio.pironti@gmail.com

Abstract

English. A crucial point for the applicability of sentiment analysis over Twitter is represented by the degree of manual intervention necessary to adapt the approach to the considered domain. In this work we propose a new sentiment polarity classifier exploiting *barrier features*, originally introduced for the classification of textual data. Empirical tests on SemEval2014 competition data sets show that such approach overcomes performance of baseline systems in nearly all cases.

Italiano. *Un punto cruciale per l'applicabilità della sentiment analysis su Twitter è rappresentato dal livello di intervento manuale necessario per adattare l'approccio al dominio considerato. In questo lavoro proponiamo un nuovo classificatore di sentiment polarity che sfrutta le* barrier features*, originariamente introdotte per la classificazione di relazioni estratte da testi. Test empirici sui data sets usati nella competizione SemEval2014 mostrano che l'approccio proposto supera le performance dei sistemi baseline nella maggioranza dei casi.*

1 Introduction

Sentiment analysis (SA) (Pang and Lee, 2008), or opinion mining, is mainly about finding out the feelings of people from data such as product reviews and news articles.

Most methods adopt a two-step strategy for SA (Pang and Lee, 2008): in the subjectivity classification step, the target is classified to be subjective or neutral (objective), while in the polarity classification step the subjective targets are further classified as positive or negative. Therefore, two classifiers are trained for the whole SA process: the subjectivity classifier and the polarity classifier. Polarity is an aspect of sentiment analysis which can be faced as a three-way classification problem, in that it aims to associate either a positive, negative or neutral polarity to each tweet.

Expressions in tweets are often ambiguous because they are very informal messages no longer than 140 characters, containing a lot of misspelled words, slang, modal particles and acronyms. The characteristics of the employed language are very different from more formal documents and we expect statistical methods trained on tweets to perform well thanks to an automatic adaptation to such specificities.

As evidenced by tasks included in competitions (Rosenthal et al., 2015) and (Nakov et al., 2016), twitter sentiment analysis is a relevant topic for scientific research. To the best of our knowledge (Ravi and Ravi, 2015; Kolchyna et al., 2015; Silva et al., 2016) present a comprehensive, State-of-the-Art (SoA) review on the research work done in various aspects of SA. Furthermore some approaches, as described in (Gonçalves et al., 2016), are based on the combination of several existing SoA "off-the-shelf" methods for sentence-level sentiment analysis[1].

(Saif et al., 2016) proposes an approach based on the notion that the sentiment of a term depends on its contextual semantics and some trigonometric properties on SentiCircles, that is a 2D geometric circle. These properties are applied to amend an initial sentiment score of terms, according to the context in which they are used. The sentiment identification at either entity or tweet-level is then performed by leveraging trigonometric identities

[1] A point of strength of this kind of systems is that combining several classification methods in an ensemble approach results to be very strong with respect to the input vocabulary size and to the amount of available training.

on SentiCircles.

The approach we are proposing has been experimentally assessed by comparing its performance with two baseline systems. In addition to that, the capability of adaptation of the approach to slightly different domains has been tested by comparing on a web-blog data set the performances of two systems in which the *Barrier Feature*(BF) dictionary has been respectively built on a collection of tweets and Wikipedia webpages. Eventually, the contribution of BF has been evaluated.

2 Proposed approach

Some automatic machine learning approaches recently applied to Twitter sentiment polarity classification try new ways to run the analysis, such as performing sentiment label propagation on Twitter follower graphs and employing social relations for user-level sentiment analysis (Speriosu et al., 2011). Others, not differently from the one we are proposing here, investigate new sets of features to train the model for sentiment identification, such as microblogging features including hashtags, emoticons etc. (Barbosa and Feng, 2010; Kouloumpis et al., 2011). Indeed, we are proposing to add *Barrier Features* (BFs) (Alicante and Corazza, 2011) to unigrams, bigrams and input parse tree and to provide them as input to a Support Vector Machine (SVM) classifier.

Introduced in the context of another application of text mining, namely relation classification, BFs are inspired by (Karlsson et al., 1995) for Part-of-Speech (PoS) tagging, but they have been completely redesigned as features rather than rules. BFs have also been exploited in (Alicante et al., 2016) for Italian Language in a unsupervised entity and relation extraction system, proving the language portability of these features. BFs describe a linguistic binding between the entities involved in each relation.

BFs require PoS tagging of the considered texts, which can be automatically performed with very high accuracy (Giménez and Màrquez, 2004). In fact, they consist of sets of PoS tags occurring between a predefined PoS pair, namely (endpoint, trigger). Similarly to unigrams and bigrams of words, these features are Boolean: for each tweet, their value is *true* if the feature occurs in the tweet, *false* otherwise.

Given a set of (endpoint, trigger) pairs *P* and a sentence (or tweet, in our case) *s*, the BFs extraction algorithm loops over the PoS tags in *s* and, for each trigger tag *t*, it looks backward in the sentence finding the closest occurrence of a PoS tag *e* such that $(e, t) \in P$. If such endpoint is found, then the algorithm extracts the barrier feature *(e, t,* $PT_{e,t}$*)*, where $PT_{e,t}$ is the set of PoS tags occurring between *e* and *t*. Otherwise it extracts as many barrier features as the number of the elements in *P* having *t* as trigger tag and, for each of them, the related tag set is the set of POS tags of all the words in the sentence preceding the trigger.

While in the preceding work (Alicante and Corazza, 2011) (endpoint, trigger) pairs were predefined, in this work we apply an innovative approach: we choose such pairs in a completely automatic and unsupervised way, starting from an unannotated data set, not necessarily in the same domain as the final task. In fact, BFs are unlexicalized as they only depend on PoS tags: for any text collection, we can perform this analysis basing on a different one which has to be similar in the kind of language but not necessarily in the domain. For instance, we expect the pairs which are more effective for the language adopted in tweets to be generally different from the ones adopted for standard texts.

In choosing the (endpoint, trigger) pairs, our purpose is two-fold: we aim to obtain a high variability of the identified sets of tags while only considering statistically significant patterns, that is, patterns having a rather large number of occurrences. In addition to this, we do not want to penalize longer patterns, although they usually correspond to larger and then more infrequent sets.

Table 1: Endpoints and triggers of the BFs employed for the tweet and text messages task.

Endpoint	**Trigger**
DT	JJR or NNPS
NNP	NNP or VBZ
IN	NNS
NN	NN or VBG or VBN
RB	RBR
PRP	VBD or VBP
TO	VB

For each possible trigger, we therefore choose the endpoint ep which maximizes the *expected information per tag* of the set corresponding to the

Table 2: BFs built considering the (endpoint, trigger) pairs listed in Table 1 and the following text: *Now/RB I/PRP can/MD see/VB why/WRB Dave/NNP Winer/NNP screams/NNS about/IN lack/NN of/IN Twitter/NNP API/NNP ,/, its/PRP limitations/NNS and/CC access/NN throttles/NNS !/.*

Barrier Feature	**Combined Text**
(TO, VB, {MD, PRP, RB})	Now I can see
(NNP, NNP, {MD, PRP, RB, VB, WRB})	Now I can see why Dave
(NNP, NNP, {})	Dave Winer
(IN, NNS, {MD, NNP, PRP, RB, VB, WRB})	Now I can see why Dave Winner screams
(NN, NN, {IN, MD, NNP, NNS, PRP, RB, VB, WRB})	Now I can see why Dave Winner screams about lack
(NNP,NNP, {IN, NN, NNS})	Winer screams about lack of Twitter
(NNP, NNP, {})	Twitter API
(IN, NNS {,, NNP, PRP})	of Twitter API, its limitations
(NN, NN, {,, CC, IN, NNP, NNS, PRP})	lack of Twitter API, its limitations and access
(IN,NNS, {,, CC, NN, NNP, NNS, PRP})	of Twitter API, its limitations and access throttles

BF, that is:

$$\mathrm{sc}(\mathrm{ep}) = -\sum_{\mathrm{BF}} \Pr(\mathrm{BF}) \frac{1}{\mathrm{len}(\mathrm{BF})} \log \Pr(\mathrm{BF}) \quad (1)$$

where $\Pr(\mathrm{BF})$ has been estimated by the corresponding frequency. In order to cut off insignificant cases, a threshold has been put on the minimum number of occurrences of the considered BF candidates. The normalization on the set size len(BF) has been introduced to avoid penalizing larger sets.

Table 1 reports the pairs resulting from this new approach and adopted for the experiments described in Section 3, Table 2 shows an example of BF extraction based on those pairs.

While in the system presented in (Alicante and Corazza, 2011) BFs were collected by only using the training set, in this work we consider an additional feature reduction step: we only take into account the BFs contained in a BFs dictionary, which is built by only considering the BFs whose number of occurrences within an unannotated data set is greater or equal than a threshold value. The data set employed in the BFs dictionary construction step is not necessarily constrained to the training set [2].

Being unlexicalized, BFs lead us to improve the portability of our approach not only towards new languages but also towards new kinds of applications. In particular, this and the dictionary construction steps are decisive both for the automation of the process and for its performance.

[2]Specifically, we used the same data set employed for the identification of the PoS (endpoint, trigger) pairs. This aspect is not as trivial as it might seem: such strategy allows the application of the approach also to tasks where the size of the annotated training set is limited.

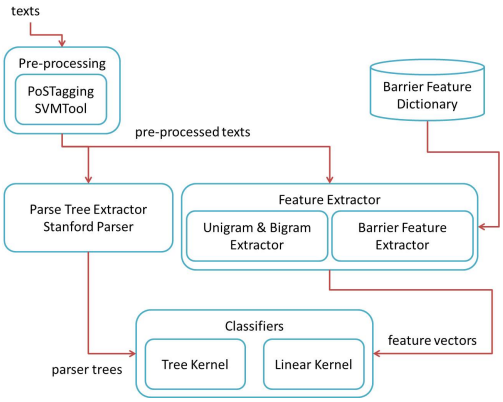


Figure 1: Architecture of the system for Twitter sentiment polarity classification.

3 Experimental Assessment

In order to evaluate our system performance, we implemented a solution for the *Message Polarity Classification* subtask of *SemEval-2014 Task 9 (Sentiment Analysis in Twitter)*[3] (Rosenthal et al., 2014). For each input tweet, our classification system decides whether it expresses a positive, negative, or neutral sentiment. According to the competition rules, the only training data we used are the ones that have been provided by the task organisers. We used a training set of about 8,000 tweets, a subset of the training and the development data released by the organisers[4].

After training the classifier on this training set, the performance of the obtained system have been evaluated against the test datasets provided for

[3]The SemEval-2014 Task 9 competition website: `http://alt.qcri.org/semeval2014/task9/`

[4]The only way to collect the data is by using the downloader script available to the participants to the competition and some of the tweets were no longer available on Twitter at the time we ran the script.

the competition: Twitter2013 (T2013), tweets provided for testing the task in 2013; Twitter2014 (T2014), a new test set delivered in 2014; Twitter2014Sarcasm(T2014Sa), a dataset of sarcastic tweets; LiveJournal2014 (LJ2014), a set of sentences extracted from the *LiveJournal* blog; SMS13, text messages provided for testing the same task in 2013. The statistics for each test datasets are shown in Table 3.

Table 3: Dataset Statistics of SemEval2014-taskB, *Message Polarity Classification*

Data Set	Positive	Negative	Neutral	Total
LiveJournal2014	427	304	411	1142
SMS2013	492	394	1207	2093
Twitter2013	1281	542	1426	3249
Twitter2014	633	125	453	1211
Twitter2014Sarcasm	33	40	13	86

Table 4: Experimental results to compare the performance of our systems with the two baseline systems. Bold cases correspond to the best performance, while symbol ‡ indicates statistical significance of the comparison with the confidence interval.

Data Set	LJ2014	SMS2013	T2013	T2014	T2014Sa
BLS1	**74.84**	70.28	70.75	69.85	58.16
BLS2	69.44	57.36	72.12	70.96	56.50
BFS	68.91	**72.01**‡	**72.88**‡	**72.10**‡	**58.79**‡

Table 5: Experimental results to evaluate the BF contribution. Bold cases correspond to the best performance. Symbol † indicates the improvement of BFS is statistical significant, verified with approximate randomisation.

Data set	System	P	R	F1
LiveJournal	WOBFS	68.94	62.34	65.32
2014	BFS	**74.69**†	**63.97**†	**68.91**†
SMS	WOBFS	64.25	**72.92**	67.14
2013	BFS	**74.80**†	69.43	**72.01**†
Twitter	WOBFS	74.30	67.32	70.62
2013	BFS	**77.96**†	**68.45**†	**72.88**†
Twitter	WOBFS	76.09	65.81	70.57
2014	BFS	**78.19**†	**66.98**†	**72.10**†
Twitter2014	WOBFS	61.12	52.92	55.28
Sarcasm	BFS	**64.75**†	**54.59**†	**58.79**†

The Barrier Features System (BFS) implements the approach we are proposing and follows the schema depicted in Figure 1. Input is tagged by using SVMtool[5] (Giménez and Màrquez, 2004) an SVM-based tagger able to achieve a very competitive accuracy on English texts. Although accuracy is likely to be lower on tweets, classification performance does not appear to be affected; this is probably due to the robustness of the statistical learner against such kind of errors. In order to reduce syntactical irregularities, we remove hashtags from tweets before providing them to the PoS-tagger component.

In the BFS system, we use the STS data set[6] to build both the (endpoint, trigger) PoS pairs and the BFs dictionary. For the BFs dictionary construction step we considered a threshold value of 10, chosen by 5-fold cross validation on the SemEval2014 training set. This resulted in $44,536$ different BFs. In conclusion, once BFs are extracted from the SemEval2014 datasets, a vector of binary features which encodes all the related unigrams, bigrams and BFs is associated to every tweet.

We use the Stanford Parser[7] (Klein and Manning, 2003a; Klein and Manning, 2003b) to extract the parse trees for each of the sentences contained in the datasets. Since a tweet can be composed by several sentences, we use Tsurgeon[8] (Levy and Andrew, 2006) to build a single parse tree for each dataset's item (tweet, text messages, etc.).

The classification module based on Support Vector Machines has been implemented using the SVMLight-TK[9] (Moschitti, 2006) package. This module takes as input both the feature vectors and the parse trees. Moreover, by applying SVMs with a combination of two different kernel functions, we can handle at the same time both structured and non-structured information. Indeed, as in (Alicante et al., 2014; Alicante and Corazza, 2011), we applied tree kernels to the parse trees and a linear kernel to the vector of binary fea-

[5]The software can be freely downloaded from `http://www.lsi.upc.edu/~nlp/SVMTool/`

[6]The Stanford Twitter Sentiment (STS) data set can be freely downloaded from `http://help.sentiment140.com/for-students/`

[7]The parser can be freely downloaded from `http://nlp.stanford.edu/software/lex-parser.shtml`

[8]Tsurgeon can be freely downloaded from `http://nlp.stanford.edu/software/tregex.shtml`

[9]The software package can be freely downloaded from `http://disi.unitn.it/moschitti/Tree-Kernel.htm`

tures described above. We build three binary classifiers, one for each sentiment/class (positive, negative, neutral). Moreover, for each classifier, the training phase has been performed by considering gold positive examples for the considered class, while negative examples are represented by all the other messages. In this way, the number of negative examples is much larger than the positive ones. SVMLight allows to balance the number of positive and negative examples by using a cost factor given by the rate between the number of negative and positive training examples. In order to assess our classification system performance, we consider two baseline systems (BLS), namely the two systems that won the SemEval2014 competition (Rosenthal et al., 2014). The former, BLS1 (Zhu et al., 2014), is based on an SVM classifier and a feature set composed by some lexical and syntactical features, while the latter, BLS2 (Miura et al., 2014), exploits a Logistic Regression trained with features based on lexical knowledge.

4 Results

Performance is assessed by adopting the same evaluation metrics as in the SemEval2014 competition (Rosenthal et al., 2014). As usual, they are based on F_1-*measure*, which is separately computed for each class (positive, negative and neutral). Table 4 compares the classification performance of our tweet system, namely BFS, and the baseline systems, namely BLS1 (Zhu et al., 2014) and BLS2 (Miura et al., 2014) by adopting the same evaluation protocol used in the SemEval2014 competition (Rosenthal et al., 2014). Our system performs significantly better on all data sets except LiveJournal2014. However, additional experiments, whose results are here omitted due to space constraints, showed that our approach performs better than BLS2 on this data set when the BF dictionary is built on Wikipedia.

We think that the explanation for this behaviour depends on the capability of the approach to adapt to the employed data set. In fact, our strategy is based on the use of unsupervised mining of text to maximize the adaptation to the specificity of the type of the language. This also explains why BFS performs worse than the others on LiveJournal2014: the syntactical structure of the structured sentences contained in a weblog is quite different from the tweets' one. It is worth highlighting that the difference in performance is not statistically significant though. The main innovation of our system is the introduction of BFs and the way in which it learns them from data. We assess the BFs contribution to the overall classification performance by comparing the performance between the Barrier Features System (BFS) and Barrier Features System (WOBFS) systems we described in Section 3 and report results in Table 5. Note that this table is more detailed than Table 4 because in this case we can run both systems and collect all the different parameters. Barrier features almost always improve performance both in terms of precision and recall, and thus also in terms of F_1. In a few cases, the introduction of BFs improves precision while decreasing recall: however, in all these cases F_1 improves in BFS with respect to WOBFS.

In conclusion, the introduction of BFs always comes with an improvement in terms of F_1 and such improvement is nearly always statistically significant. We can therefore conclude that BFs provide a crucial contribution to sentiment polarity classification.

5 Conclusions and future work

We explored the effectiveness of BFs for sentiment polarity classification in Twitter posts and we showed on SemEval2014 data sets that they can be very effective. In our approach, the need of a manual intervention is really minimum. Indeed, the BFs dictionary can be built from any collection of tweets, even one that do not belong to the same domain of the considered task. This is quite interesting because it suggests that BFs are able to capture hints about the polarity of the expressions in a domain independent way.

References

Anita Alicante and Anna Corazza. 2011. Barrier features for classification of semantic relations. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nicolas Nicolov, editors, *RANLP*, pages 509–514. RANLP 2011 Organising Committee.

Anita Alicante, Massimo Benerecetti, Anna Corazza, and Stefano Silvestri. 2014. A distributed information extraction system integrating ontological knowledge and probabilistic classifiers. In *Proceedings of the 9th International 3PGCIC-2014 Conference*, Guangzhou, CHINA. In Press.

Anita Alicante, Anna Corazza, Francesco Isgrò, and Stefano Silvestri. 2016. Unsupervised entity and

relation extraction from clinical records in Italian. *Computers in Biology and Medicine*.

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jesús Giménez and Lluís Màrquez. 2004. SVMTool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC)*, pages 43–46, Lisbon, Portugal.

Pollyanna Gonçalves, Daniel Hasan Dalip, Helen Costa, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. On the combination of off-the-shelf sentiment analysis methods. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 1158–1165. ACM.

Fred Karlsson, Atro Voutilainen, Juha Heikkila, and Arto Anttila, editors. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Mouton de Gruyter.

Dan Klein and Christopher D. Manning. 2003a. Accurate unlexicalized parsing. In *Proc. of ACL 03 of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430, Morristown, NJ, USA. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2003b. Fast exact inference with a factored model for natural language parsing. In *Proc of NIPS03: In Advances in Neural Information Processing Systems*, pages 3–10. MIT Press.

Olga Kolchyna, Thársis TP Souza, Philip Treleaven, and Tomaso Aste. 2015. Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955*.

E. Kouloumpis, T. Wilson, and J. Moore. 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG! In *Fifth International AAAI Conference on Weblogs and Social Media*.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *Proceedings of the fifth international conference on Language Resources and Evaluation*, pages 2231–2234. Citeseer.

Yasuhide Miura, Shigeyuki Sakaki, Keigo Hattori, and Tomoko Ohkuma. 2014. Teamx: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 628–632, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Alessandro Moschitti. 2006. Making tree kernels practical for natural language learning. In *EACL*.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), San Diego, US (forthcoming)*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. *Proceedings of SemEval-2015*.

Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani. 2016. Contextual semantics for sentiment analysis of twitter. *Information Processing & Management*, 52(1):5–19.

Nadia Felix F Da Silva, Luiz FS Coletta, and Eduardo R Hruschka. 2016. A survey and comparative study of tweet sentiment analysis via semi-supervised learning. *ACM Computing Surveys (CSUR)*, 49(1):15.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, EMNLP '11, pages 53–63, Stroudsburg, PA, USA. Association for Computational Linguistics.

Xiaodan Zhu, Svetlana Kiritchenko, and Saif Mohammad. 2014. Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 443–447, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

The DiDi Corpus of South Tyrolean CMC Data: A multilingual corpus of Facebook texts

Jennifer-Carmen Frey **Aivars Glaznieks** **Egon W. Stemle**
Institute for Specialised Communication and Multilingualism
EURAC Research
Bolzano/Bozen, Italy
{jennifer.frey,aivars.glaznieks,egon.stemle}@eurac.edu

Abstract

English. The DiDi corpus of South Tyrolean data of computer-mediated communication (CMC) is a multilingual sociolinguistic language corpus. It consists of around 600,000 tokens collected from 136 profiles of Facebook users residing in South Tyrol, Italy. In conformity with the multilingual situation of the territory, the main languages of the corpus are German and Italian (followed by English). The data has been manually anonymised and provides manually corrected part-of-speech tags for the Italian language texts and manually normalised data for German texts. Moreover, it is annotated with user-provided socio-demographic data (among others L1, gender, age, education, and internet communication habits) from a questionnaire, and linguistic annotations regarding CMC phenomena, languages and varieties. The anonymised corpus is freely available for research purposes.

Italiano. *DiDi è un corpus di comunicazione mediata dal computer (CMC), che raccoglie dati linguistici di area sudtirolese. Il corpus, multilingue e sociolinguistico, è composto da circa 600,000 occorrenze raccolte (previo consenso all'utilizzo dei dati) dai profili di 136 iscritti a Facebook e residenti in Alto Adige. Le principali lingue del corpus, tedesco e italiano (seguite dall'inglese), riflettono lo spazio plurilingue del territorio. I dati sono stati manualmente anonimizzati e i testi in lingua italiana sono corredati da etichette (manualmente corrette) per le parti del discorso. Inoltre, DiDi è annotato con dati sociodemografici forniti dall'utente (fra gli altri: L1, genere, età, istruzione e modalità di comunicazione via Internet) attraverso un questionario e contiene ulteriori annotazioni linguistiche relative a fenomeni legati alla CMC e agli usi di varietà linguistiche. Il corpus anonimizzato è liberamente disponibile a fini di ricerca.*

1 The DiDi Project

The autonomous Italian province of South Tyrol is characterized by a multilingual environment with three official languages (Italian, German, and Ladin), an institutional bi- or trilingualism (depending on the percentage of the Ladin population), and diverse individual language repertoires (Ciccolone, 2010).

In the regionally funded DiDi project,[1] the goal was to build a South Tyrolean CMC corpus to document the current language use of residents and to analyse it socio-linguistically with a focus on age. The project initially focused on the German-speaking language group. However, all information regarding the project, e.g. the invitation to participate, the privacy agreement, the project web site, and the questionnaire for socio-demographic data was published in German and Italian. Hence, we attracted speakers of both Italian and German. Accordingly, the collected data is multilingual, with major parts in German but with a substantial portion in Italian (100,000 of 600,000 tokens).

The collected multilingual CMC corpus combines Facebook status updates, comments, and private messages with socio-demographic data (e.g. language biography, internet usage habits, and general parameters like age, gender, level of education) of the writers. The data was enriched with linguistic annotations on thread, text and token level including language-specific part-

[1] For further information see www.eurac.edu/didi.

of-speech (PoS) and lemma information, normalisation, and language identification.

In this paper, we describe the corpus with respect to its multilingual characteristics and give special emphasis to the Italian part of the corpus to which we added manually corrected PoS annotations. Hence, it presents a continuation of Frey et al. (2015) which was restricted to German texts of the corpus, not taking into account the full variety of data collected for the total corpus.

2 Corpus Construction

For the purpose of the DiDi project, we collected language data from social networking sites (SNS) and combined it with socio-demographic data about the writers obtained from a questionnaire. We chose to collect data from Facebook as this SNS is well known in South Tyrol, hosts a wide variety of different communication settings, and is used over the whole territory by nearly all groups of the society.

Related research mainly draws on public data such as public Facebook groups, Twitter or chat data (e.g. Celli and Polonio (2013), Basile and Nissim (2013), Burghardt et al. (2016), Beißwenger (2013)), excluding the possibility to analyse discourse patterns of non-public everyday language use.

Collecting non-public and personal data for the DiDi corpus raised technical issues regarding Italian privacy regulations (which require user consent incl. privacy statement), the time-saving acquisition of authentic and complete language data, and the assignment of language data to questionnaire data. These issues have been solved by developing a Facebook application[2] that allowed for the gathering of all three sorts of data (user consent, language data, questionnaire data) at once. In addition, the application was easy to share via Facebook which helped to promote the project and to reach many potential participants. While data collection was solely managed by the Facebook application, we relied on Facebook's in-platform means (i.e. users' sharing and liking) to recruit participants. In order to reach older users (> 50 years) it was necessary to additionally resort to Facebook advertisment.[3]

With the consent of each participant, the data was downloaded via the Facebook Graph API[4] and from the used questionnaire service[5], and stored in a local MongoDB[6] data base. Both entities were linked via randomised unique identifiers. A python interface provided access points to retrieve user and text data from the data base in a linked and structured format, and also allowed to rebuild the conversational structure of threads by linking successive text objects together. This information can now be used to analyse turn-taking and language choices within threads.[7]

3 Corpus Annotations

This section describes the annotations added during the process of corpus construction.[8]

3.1 Socio-demographic Information about Participants

The corpus provides the following socio-demographic information about the participants obtained from the online questionnaire: gender, education, employment, internet communication habits, communication devices in use, internet experience, first language(s) (L1), and usage of a South Tyrolean German or Italian dialect and its particular origin.

3.2 Linguistic Annotation of Texts

The corpus was annotated on text and token level with a series of information.

- *Language identification:*
  The used languages of a text were identified in a semi-automatic approach: Firstly, using the language identification tool *langid.py* (Lui and Baldwin, 2012), and secondly, manually correcting short texts and texts with a low confidence score.

- *Tokenisation:*
  The corpus was tokenized with the Twitter tokenizer *ark-twokenize-py*[9] and subse-

[2] The source code is available at `https://bitbucket.org/commul/didi_app`.

[3] For details regarding the technical and strategical design of the data collection and methods of user recruitment see Frey et al. (2014).

[4] `https://developers.facebook.com/docs/graph-api`

[5] `http://www.objectplanet.com/opinio/`

[6] `https://www.mongodb.com/`

[7] The source code is available at `https://bitbucket.org/commul/didi_proxy`.

[8] See Frey et al. (2015) for detailed information on the anonymisation procedure and the normalisation and processing of German texts, including identification of languages and varieties.

[9] `https://github.com/myleott/ark-twokenize-py`

quently corrected manually for non-standard language tokenisation issues.

- *Part-of-speech tagging and lemmatization:* (Corrected) tokens were annotated with PoS tags and lemma information considering the predominant language of the text at hand. We tagged Italian texts with the Italian tag set of the Universal Dependencies project[10] using the *RDR PoS Tagger* (Nguyen et al., 2014). Subsequently, we manually corrected PoS annotations to handle bad tagging accuracy for social media texts. Additionally, we used the *TreeTagger* (Schmid, 1994; Schmid, 1995) to assign PoS tags for German, English, Spanish, French and Portuguese texts applying the standard tagsets for each language. No manual correction was performed for these languages.

- *Normalisation:* So far, we have manually normalised non-standard language to word-by-word standard transcriptions only for German texts.

- *Variety of German:* We classified German texts as dialect, non-dialect or unclassifiable texts applying a heuristic approach based on the normalisation.

- *Untranslatable dialect lexemes:* We have created a lexicon for untranslatable dialect words encountered during manual normalisation. The dialect lexicon was used to post-process out-of-vocabulary (OOV) tokens in the corpus.

- *Foreign language insertions:* The most common OOV tokens that we manually classified as foreign language vocabulary have been annotated with information about their language origin.

- *CMC phenomena:* Emoticons, emojis, @mentions, hashtags, hyperlinks, and iterations of graphemes and punctuation marks were annotated automatically using regular expressions.

- *Topic of the text:* In order to investigate context factors of language choice we annotated texts as either political or non-political according to a list of politicians, political parties and political terms.

[10] http://universaldependencies.org/it/pos/index.html

3.3 Conversation-related Annotations

We rebuilt conversation threads by linking successive texts and created thread objects containing ordered lists of texts that are accessible via the Python interface. Thread objects contain information about the used languages and the number of active interlocutors and recipients of a message as well as the time passed between two texts.

As described in Frey et al. (2015), no text content of non-participants of the DiDi project was stored, but general information about the publishing time and the language of the text was kept. If all interlocutors of a thread were participants of the project, the whole conversation is available.

3.4 User-related Annotations

In addition to socio-demographic data, we added information about the users' (multilingual) communicational behaviour, i.e. their primary language, used languages and the number of interlocutors.

4 Corpus Data

4.1 Corpus Size

The DiDi corpus comprises public and non-public language data of 136 South Tyrolean Facebook users. The users could choose to provide either their Facebook wall communication (status updates and comments), their chat (i.e. private messages) communication or both. In the end, 50 people provided access to both types of data. 80 users only provided access to their Facebook wall and 6 users gave us only their chat communication. In total, the corpus consists of around 600 thousand tokens that are distributed over the text categories status updates (172 ,66 tokens), comments (94,512 tokens) and chat messages (328,796 tokens).

4.2 Multilingualism in the Corpus

The corpus is highly multilingual. Although the initial intention of the project was to document the use of German in South Tyrol, German language content comprises only 58% of the corpus. 13% are written in Italian and 4% in English (the remainder of the messages was either classified as unidentifiable language, non-language or other language). The distribution of the languages is

based on the language backgrounds of the participants and is comparable to the multilingual community of South Tyrol. The following tables show the distribution of profiles, texts and tokens (table 1) and text type (table 2) by L1.

User L1	Profiles	Texts	Tokens
IT	9	4,260	80,368
DE	108	29,883	421,262
other	3	407	8,643
IT + DE	11	4,165	75,359
DE + other	5	1,110	10,642
Total	136	39,825	596,274

Table 1: Distribution of profiles, texts and tokens by L1.

User L1	SU	CO	PM
IT	1,682	1,063	1,515
DE	7,286	4,890	17,707
other	172	45	190
IT+DE	1,962	343	2,791
DE+other	1,031	166	13
Total	11,102	6,507	22,216

Table 2: Distribution of texts by text type (SU = status updated, CO = comments, PM = private messages) by L1.

While very few users wrote only in their first language, most users used at least two (88%), very often even three (73%) or more (51%) languages. Table 3 shows the number and proportion of German, Italian and English texts written as first or second/foreign language.

Text written	as L1	as L2
IT	4,761 (57%)	3,566 (42%)
DE	23,191 (99%)	170 (1%)
EN	166 (4%)	3,625 (96%)
All languages	28,120 (78%)	7,842 (22%)

Table 3: Distribution of text language by L1 or L2 use.

In terms of multilingual language use in the DiDi corpus, we observe a slight difference between Italian and German-speaking users. L1 Italian speakers stick more to their L1 compared to the German-speaking participants, who are characterized by a higher usage of L2 Italian. The comparison of L1 and L2 usage in status updates, comments and private messages (c.f. Table 4) shows that the respective L1 is preferred in all messages types. We find the highest percentage of second or foreign language use in status updates, whereas in comments and private messages around 75% of the texts are written in L1.

Text written	as L1	as L2
Status updates	6,774 (61%)	3,032 (27%)
Comments	5,089 (78%)	924 (14%)
Messages	16,257 (73%)	3,886 (17%)
Total	28,120 (71%)	7,842 (20%)

Table 4: Distribution of L1 and L2 use by text types.

Finally, we observed 4,295 code-switching instances on conversation level and at least 1,653 texts that contain multiple languages[11]. The average number of code-switching instances per user is 10%, meaning that every tenth text does not continue the language of the previous text in the thread (the maximum was around every second text, i.e. 42%). The average proportion of text with multiple languages per user is 4% (max. 25%).

5 Issues in Corpus Creation

In addition to general issues of working with social media texts (e.g. text processing on noisy, short texts as described for example in (Baldwin et al., 2013; Eisenstein, 2013)) , the high diversity in used languages and varieties in our corpus led to various restraints in corpus creation and processing as cross-lingual annotation and information extraction are still crucial problems in natural language processing. We tried to address the demands of a multilingual corpus by providing language specific PoS tagging and by applying language independent annotations. We are aware of the fact that this is by no means sufficient to deal with linguistic research questions that exceed language boundaries. Moreover, manual correction tasks occupied a significant part of the work on the corpus as automatic annotation (e.g. for language identification) does not yet provide the accuracy expected for linguistic studies (Carter et al., 2013; Lui and Baldwin, 2014).

[11]Texts were annotated as mixed-language texts during the correction of the language identification, therefore this annotation has not been done for the whole corpus. A further word-level identification of languages could detect even more mixed-language content(Nguyen and Dogruoz, 2013)

6 Conclusion and Future Work

In this paper we presented a freely available language corpus of Facebook user profiles from South Tyrol, Italy. The multilingual corpus is anonymised and annotated with sociodemographic data of users, language specific (and for Italian manually corrected) PoS tags, lemmas and linguistic annotations mainly related to used languages, varieties and multilingual phenomena. The corpus is accessible for querying via ANNIS[12] or can be obtained as processable data for research purposes on `http://www.eurac.edu/didi`.

Acknowledgements

The project was financed by the Provincia autonoma di Bolzano – Alto Adige, Ripartizione Diritto allo studio, università e ricerca scientifica, Legge provinciale 13 dicembre 2006, n. 14 "Ricerca e innovazione".

References

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107.

Michael Beißwenger. 2013. Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik*, 41(1):161–164.

Manuel Burghardt, Daniel Granvogl, and Christian Wolff. 2016. Creating a Lexicon of Bavarian Dialect by Means of Facebook Language Data and Crowdsourcing. In *Proceedings of LREC 2016*, pages 2029–2033.

Simon Carter, Wouter Weerkamp, and Manos Tsagkias. 2013. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1):195–215.

Fabio Celli and Luca Polonio. 2013. Relationships between personality and interactions in facebook. *Social Networking: Recent Trends, Emerging Issues and Future Outlook*, pages 41–54.

[12] `http://annis-tools.org/`

Simone Ciccolone. 2010. *Lo standard tedesco in Alto Adige*. Il segno e le lettere. LED Edizioni Universitarie, Milan.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of NAACL-HLT*, pages 359–369.

Jennifer-Carmen Frey, Egon W. Stemle, and Aivars Glaznieks. 2014. Collecting language data of non-public social media profiles. In Gertrud Faaß and Josef Ruppenhofer, editors, *Workshop Proceedings of the 12th Edition of the KONVENS Conference*, pages 11–15, Hildesheim, Germany, October. Universitatsverlag Hildesheim, Germany.

Jennifer-Carmen Frey, Egon W. Stemle, and Aivars Glaznieks. 2015. The DiDi Corpus of South Tyrolean CMC Data. In *Workshop Proceedings of the 2nd Workshop on NLP4CMC at GSCL2015*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2014. Accurate language identification of twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)@ EACL*, pages 17–25, Gothenburg. Association for Computational Linguistics.

Dong-Phuong Nguyen and A Seza Dogruoz. 2013. Word level language identification in online multilingual communication. Association for Computational Linguistics.

Dat Quoc Nguyen, Dang Duc Pham Dai Quoc Nguyen, and Son Bao Pham. 2014. RDRPOSTagger: A ripple down rules-based part-of-speech tagger. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 17–20.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Sequenze N+pN (nome comune + nome proprio): descrizione linguistica da un corpus dell'italiano.

Felicia Logozzo
Università di Roma 'Tor Vergata'
via Columbia, 1 – 00133 Roma
logozzo@lettere.uniroma2.it

Abstract

English. This paper describes the most important N+pN (noun + proper noun) structures in Italian from the corpus of La Repubblica 2002-2005

Italiano. *Il contributo descrive le strutture N+pN più significative estratte dal corpus de La Repubblica 2002-2005.*

1 Introduzione

I dati oggetto dell'analisi sono stati ricavati dal corpus de *La Repubblica* 2002-2005 (circa 20 milioni di token), mediante T2K[1] (*Text to Knowledge*), sistema automatico per l'estrazione e l'organizzazione della conoscenza da testi.
Al software è stato richiesto di selezionare le sequenze N+N (nome comune + nome comune) e N+pN (nome comune + nome proprio) con frequenza pari o superiore a 18 occorrenze.
Dei 557 risultati estratti (24.013 occorrenze complessive) ne sono stai considerati validi 437 (78%[2] circa); i restanti 120 sono prevalentemente sequenze N+A (nome + aggettivo) o A+N tra cui composti nominali con aggettivo[3], sia con testa a sinistra (*maglia rosa, giudice istruttore, gruppo dirigente, piano regolatore* ecc.), sia con testa a destra (*sostituto procuratore* ecc.); tra questi ultimi, anche *anni Sessanta, anni Novanta, anni Trenta, anni Venti, anni Quaranta* - classificati dal software come sequenze con nomi propri, per l'uso convenzionale della maiuscola.
Il gruppo delle 437 sequenze potenzialmente utili ai fini dell'analisi N+N/N+pN si riduce ulteriormente, se si eliminano le 14 sequenze non complete (*calo rispetto, tv via, pistola calibro, via G, via de'*), le 2 locuzioni avverbiali (*passo passo, fin fine*) e i 41 prestiti stranieri ed espressioni latine (*talk show, pole position, week end, par condicio, alter ego, call center, best seller, tour operator* ecc.).
Escludendo infine un appellativo - *signor Presidente* - e la parte di polirematica - *parola fine*-, l'insieme delle 380 sequenze rimanenti è costituito da 147 composti N+N e da 231 sequenze N+pN, oggetto, queste ultime, di analisi specifica nel prossimo paragrafo.
Quanto al primo gruppo, sono rappresentati nel corpus composti N+N[4] di tutte le categorie: coordinati endocentrici (*decreto legge*) ed esocentrici (*centro sinistra*), subordinati (*ufficio stampa*). Tra questi, ben 53 sequenze su 147 sono rappresentate da espressioni di spazio o di tempo, formate con pochi e diffusissimi elementi base (*centro, metà, mattina, pomeriggio, sera, inizio, fine,* punti cardinali[5]).
Sebbene l'italiano non esprima tradizionalmente rapporti di subordinazione mediante giustapposizione, ma prediliga sintagmi preposizionali e

[1](Dell'Orletta et al., 2014).
[2]Sul corpus in questione, l'accuratezza del *Part of Speech Tagger* impiegato risulta essere complessivamente pari al 96,34% (cfr. Dell'Orletta, 2009).
[3]Per la classificazione dei composti, là dove non diversamente specificato, si fa riferimento a Bisetto (2004: 33-50).
[4]Per i composti N+N dell'italiano, cfr. i lavori di Jan Radimský, in particolare Radimský (2013), sul corpus de *La Repubblica,* e Radimský (2015).
[5]Sono esempi prototipici di basi di composti grounding: 'Le teste dei composti grounding in italiano esprimono concetti intrinsecamente relazionali, che richiedono una specificazione ulteriore per convogliare un contenuto efficacemente informativo; esempi tipici di teste grounding sono i portatori/contenitori di informazione, organizzazioni, luoghi, aggregatori, puntatori nel tempo e nello spazio, proprietà misurabili' (Baroni, Guevara e Pirrelli, 2009)

aggettivi relazionali, la composizione N+N è, com'è noto, molto più diffusa di quanto appaia dalle grammatiche descrittive, che non prevedono, ad esempio, la categoria del nome attributivo, impiegata nella descrizione della lingua inglese, là dove convivono aggettivi denominali e nomi giustapposti corradicali con medesima funzione (*wool swater / woolen sweater*) [6].

Di contro, in alcuni dizionari[7], per superare le difficoltà descrittive delle strutture N+N, basi nominali di composti, prevalentemente ma non esclusivamente coordinati, vengono registrati come aggettivi invariabili (*parola **chiave**, punti **chiave**, cifra **record**, tempo **record**, motore **diesel**, rabbino **capo**, ruolo **chiave**, città **campione**, uomo **chiave**, campo **base**, livelli **record***).

2 Denominazione e composizione N+pN

Il rapporto prevalente per le sequenze N+pN è senza dubbio quello identificativo/appositivo (129 su 231 sequenze, 5451 occorrenze su 9430). La lingua italiana prevede infatti che ad un nome proprio possa/debba essere affiancata la propria categoria di appartenenza mediante apposizione.

N+pN

I casi prototipici di identificazione-apposizione sono rappresentati da sequenze titolo+pN (*presidente Bush, presidente Ciampi, ministro Tremonti, sindaco Veltroni, ministro Castelli* ecc.) o nomi societari (*società Autostrade*) in cui

pN è un N

come confermano usi quali: *'Autostrade è una società privatizzata negli anni '90 che opera in un regime di concessione statale'* (il giornale.it).

A differenza degli esempi di cui sopra, in molti casi l'apposizione è indispensabile e fa parte integrante dell'espressione 'nome proprio': è il caso, ad esempio, degli odonimi[8] che si compongono obbligatoriamente con struttura N+pN (*via Cristoforo Colombo ≠ Cristoforo Colombo*). Se col tempo una specifica strada riesce ad acquisire una sufficiente notorietà, potrà essere identificata anche soltanto col nome proprio preceduto dall'articolo, segno – quest'ultimo - della presenza silenziosa dell'omesso nome comune (*la Colombo*).

Gli odonimi infatti possono essere inclusi nella più generale categoria per cui vale l'espressione:

N intitolato a pN

premio Nobel, via Nazionale, palazzo Chigi, viale Mazzini, piazza Vittorio, piazza Navona, via Roma, corso Vittorio, teatro Politeama, palazzo Koch.

In sequenze di questo tipo, i nomi propri sono, per così dire, nomi di riuso; la relazione N+pN continua ad essere una relazione identificativo-appositiva, ma inizia ad acquisire aspetti più complessi che la proiettano verso la subordinazione.

Le vicende specifiche di ognuna di queste sequenze fa sì che sia sentita accettabile o meno l'equivalenza *pN è un N*. Tutti d'accordo, per esempio, che *il Nobel* sia un premio e che *il Politeama* sia un teatro; più difficile che *via Nazionale* o *via Roma* viaggino senza nome comune:

via Nazionale è una via / Piazza Venezia è una piazza / palazzo Chigi è un palazzo

?/ la Nazionale è una via / *(la) Venezia è una piazza / *il Chigi è un palazzo*

In questi casi, la motivazione che lega N a pN, sfugge ad un'analisi esclusivamente linguistica. Se si considera ad esempio *palazzo Chigi*, in rapporto ad altre sequenze dalla medesima struttura, come *palazzo Benetton e palazzo Fuksas*, emerge che solo conoscenze extralinguistiche permettono di sapere che i Chigi comprarono e abitarono l'omonimo palazzo nel 600; *palazzo Benetton* è stato progettato da Fuksas su commissione appunto di Benetton, e poi rivenduto a H&M; *palazzo Fuksas* è stato progettato anch'esso da Fuksas, a Pescara, e ospita la sede di una importante multinazionale.

Si tratta sostanzialmente di strutture analizzabili solo a livello pragmatico[9]: col passare del tempo, la relazione di motivazione che lega N+pN (il fatto di aver comprato, commissionato o progettato il palazzo) svanisce nella coscienza dei parlanti e l'intera sequenza N+pN diventa denominazione. Il processo è pressoché concluso nel caso di *palazzo Chigi*, sentita quasi solo come denominazione; appena iniziato per *palazzo Fuksas*.

A dispetto di quanto dichiarato dalle grammatiche, è cristallizzata nella lingua dell'uso la giustapposizione N+pN per l'espressione di relazio-

[6] Un'alternanza parzialmente ma non completamente sovrapponibile – non si tratta infatti di corradicali in senso stretto - si verifica in italiano in presenza di Scritture Brevi come in *tv/televisivo - programma tv / programma televisivo* (già oggetto di analisi in Logozzo, in stampa).

[7] Sabatini-Coletti, per esempio.

[8] Per denominazione e composizione con toponimi e marchionimi cfr. Logozzo (2015).

[9] Cfr. Bartning (2001: 150 ss).

ni morfosintattiche diverse da quella identificativa-appositiva.
Il corpus ci restituisce in tal senso 102 sequenze che coprono il 44% del totale delle sequenze N+pN.
Rendere conto accuratamente delle strutture soggiacenti a ciascuna sequenza è questione complessa; per questo, le sequenze estratte dal corpus verranno classificate mediante criteri volutamente inclusivi, volti ad individuare analogie più che a differenziare, al solo fine dell'analisi del ruolo dei nomi propri.
Il primo e più numeroso gruppo identificato (46 sequenze per 1475 occorrenze) è quello delle sequenze che rimandano in qualche modo ad un rapporto di appartenenza in senso lato, morale, sociale o figurato[10]:

N appartiene a pN

commissario Ue, soldati Usa, coppa Italia, amministrazione Usa, ispettori Onu, segretario Ds, vertice Rai, truppe Usa, segretario ds, dati Istat, paesi Ue, canone Rai, economia Usa, coppa Uefa, comando Usa, presidente Ue, reti Mediaset, leader Cgil, deputato Ds, ambasciata Usa, capogruppo Ds, militari Usa, commissione Ue, esercito Usa, leader Ds, senatore Ds, presidenza Ue, ex dc, reti Rai, forze Usa, stabilimento Fiat, governo Usa, norme Ue, giornalista Rai, agenti Cia, presidente Rai, vertice Ue, operai Fiat, consiglio Rai, consigliere Rai, base Usa, vertici Fiat, segretario Cgil, classifica Fifa, presidente ds, segretario Udc.

Andando ad osservare il rapporto semantico e sintattico che intercorre tra N e pN, appare evidente che si tratta di relazioni di subordinazione che potrebbero essere espresse in forma analitica mediante l'impiego di parole grammaticali – la preposizione 'di' primariamente – o di perifrasi verbali, riconducibili ai rapporti cosiddetti genitivali.
La mancanza della preposizione 'di' a collegare N e pN, fa sì che le strutture formalmente appaiano come composti. Nello specifico, si tratterà di composti determinativi che 'indicano una sottoclasse degli oggetti individuati dalla testa e nei quali il nome di destra si pone come una specificazione restrittiva del nome che lo precede'.[11]
Sulla base dei dati emersi dal corpus, le strutture che consentono questo tipo di costruzioni sembrano essere piuttosto uniformemente costituite.
La composizione N+pN per l'espressione di relazioni di appartenenza concrete o figurate, è dunque ammessa in italiano seppure con restrizioni riguardanti, e la forma, e la sostanza del determinante: quanto alla sostanza, entrano in composizione nomi propri di enti, associazioni, istituzioni o marchionimi; quanto alla forma, sono rappresentati determinanti di un'unica parola (*?/* deputato Unione di Centro*), meglio se in forma di sigla (Scritture Brevi[12]).
Alla base della grande produttività della composizione con Scritture Brevi vi è una motivazione semantica (il valore referenziale delle sigle coinvolte che, in quanto nomi propri dotati di riferimento, riducono le possibili ambiguità interpretative della relazione N+pN) e una motivazione formale: i composti sono strutturalmente formati da due elementi; la sigla permette di ridurre un nome proprio complesso, eventualmente costituito da un'espressione con coordinazioni e subordinazioni al suo interno (*Figc = Federazione Italiana Giuoco calcio*), ad elemento unico, evitando così lunghe successioni asindetiche di elementi (?/* *Presidente Federazione Italiana Giuoco Calcio*)[13].
Non sembrano ammesse giustapposizioni di nomi propri che abbiano generato nella lingua aggettivi relazionali denominali - si pensi agli etnici derivati da nomi di stato/nazione - tranne nel caso in cui essi siano siglati; si ricrea in tal modo l'alternanza già vista per *tv/televisivo* (cfr. nota 6)*: soldati USA / soldati statunitensi* ma non **soldati Stati Uniti.*
La composizione N+pN del tipo appena analizzato può essere considerata produttiva perché si presta alla composizione occasionale, pratica originariamente non diffusa nella lingua italiana[14], a differenza dell'inglese[15] che conia costantemente strutture pN+N.
E' importante, a questo proposito, distinguere l'imposizione di un nome proprio o di un'etichetta individuante (*Coppa Italia, via Marchesi*) dalla composizione occasionale (*presidente USA*). Se infatti nel primo caso le regole della

[10]'What is normally called possession is the linguistic expression of the relation between two entities, a Possessor and a Possessum, such that one, the Possessor, is seen as being in some way related to the other, the Possessum, as having it near or controlling it. The kind of relation between the two can be of various sorts' (Herslund and Baron, 2001: 2).

[11](Bisetto, 2004: 40).
[12]Cfr. Chiusaroli (2012).
[13] Cfr. Logozzo (in stampa).
[14]Dardano (1978: 143) parla di 'creazioni effimere'.
[15]Cfr. Bauer (1983: 202 ss).

grammatica si possono prescindere senza troppi imbarazzi e l'approccio di analisi può essere solo descrittivo, a posteriori, nel caso della composizione occasionale è possibile azzardare ipotesi predittive, assumendo che certe strutture saranno più probabilmente scelte, rispetto ad altre, sulla base dell'uso.

Merita un'attenzione particolare l'impiego degli antroponimi giustapposti in sequenze N+pN, per i quali il corpus ci restituisce i seguenti esempi:

governo Berlusconi, amministrazione Bush, legge Gasparri, legge Cirami, riforma Moratti, governo Sharon, lodo Schifani, legge Biagi, legge Bossi-Fini, governo Prodi, governo Blair, lodo Maccanico, governo Zapatero, riforma Dini, riforma Biagi, riforma Tremonti, governo Cuffaro, legge Boato

Sulla base dei dati estratti, si evince che non sono ammessi antroponimi giustapposti ad esprimere vere e proprie relazione di possesso[16] (*auto Antonio, *giacca Marcello).

Quando essi compaiono in sequenze N+pN, esprimono relazioni che oscillano dall'identificativo-appositivo al determinativo-specificativo e richiedono di spostare l'analisi dalla semantica degli elementi coinvolti, alla pragmatica del discorso.

E' infatti invalso l'uso di attribuire una denominazione a leggi e riforme, impiegando antroponimi come nomi propri di riuso. La riforma della scuola del 2003 è dunque 'intitolata' al Ministro (Moratti) che si è fatto promotore della stessa, così come si danno, per esempio, nomi ai governi (*Berlusconi 1, Berlusconi 4, Prodi 2*) in espressioni in cui la relazione N+pN può essere al contempo di subordinazione (*governo guidato da, nominato da, riforma ideata da, legge proposta da*) o di denominazione (*Il governo Berlusconi 3 è durato dal 23 aprile 2005 al 17 maggio 2006* – Wikipedia.it).

Come già detto a proposito degli odonimi di tipo *N intitolato a pN*, gli antroponimi non entrano bene in predicazione, nemmeno in presenza di un articolo che sottintenda l'omissione di un nome comune, per le evidenti possibili ambiguità: *la Bossi-Fini* è infatti una legge, ma *la Moratti* non è una riforma. Un elemento determinativo ulteriore di pN, che escluda il riferimento diretto ad un essere umano, consente il superamento del problema, come, ad esempio, un altro antroponimo in *Bossi-Fini*, un cardinale in *Berlusconi 3*.

Al limite tra relazione identificativo-appositiva e specificativo-determinativa anche quella tra N e pN in *gruppo Fiat, gruppo Parmalat, gruppo Telecom, gruppo Mediaset*, dal momento che, per esempio, Mediaset è una holding proprietaria di un gruppo di aziende, ma è anche l'etichetta dell'intero gruppo che prende il nome della capofila.

Denominate sulla base delle tematiche di cui si occupano, come testimonia la maiuscola dei determinanti che le pone nel gruppo N+pN e non in quello N+N, le *commissioni Giustizia, Lavoro, Bilancio e Affari (Costituzionali)* in cui però si mantiene un rapporto di subordinazione tra N e pN: **(La) Giustizia è una commissione / Quella della giustizia è una commissione.*

Altre sequenze del corpus si possono raccogliere infine in una categoria aspecifica:

si è verificato un N e riguarda pN

processo Sme, caso Parmalat, crisi Fiat, caso Sofri, caso Fiat, caso Cirio, crac Parmalat, caso Moro, crac Cirio, caso Fazio, caso Catania, processo Imi-Sir, caso Sme, affare Sme

Rientra in questo gruppo il tipo *caso*+pN, ispirato forse al modello francese dell'*affaire Dreyfus*[17], tipo estremamente produttivo in cui è possibile inserire anche il sottotipo *crisi Fiat, crac Parmalat*. Sebbene rappresentato nel corpus da un'unica occorrenza (*omicidio Biagi*) è bene citare in questo gruppo anche i genitivi oggettivi con antroponimi, che sembrano entrare in composizione con specifiche teste (es: *omicidio, rapimento*) che permettono uno scioglimento del tipo: *si è verificato un omicidio e ha riguardato Biagi*, probabilmente nel momento in cui un certo episodio di cronaca è talmente di dominio pubblico da diventare un 'caso' quasi identificabile con il solo pN.

3 Conclusioni

La presenza dei nomi propri nelle sequenze N+pN fa sì che si abbia a che fare, nella maggior parte dei casi, con relazioni di denominazione con motivazione più o meno trasparente, analizzabili, ora ricorrendo alla semantica degli elementi coinvolti, ora, più frequentemente, ricorrendo alla conoscenza del mondo extralinguistico. Completamente estranee alla denominazione, sono invece le sequenze del tipo *commissario Ue, soldati USA* che rappresentano per l'italiano esempi di composizione N+pN, alternativa all'impiego di aggettivi denominali o sintagmi preposizionali.

[16]Cfr. Heine (2006).

[17]Per l'influsso di modelli alloglotti sulle costruzioni N+pN cfr. Klajn (1972: 180 ss).

Bibliografia

Baroni M., Guevara E. e Pirrelli V. 2009. *Sulla tipologia dei composti N+N in italiano:principi categoriali ed evidenza distribuzionale a confronto,* in Ferrari G., Mosca M. e Benatti R. (a cura di) *Linguistica e modelli tecnologici di ricerca*. Bulzoni, Roma, 73-95.

Bartning I. 2001. *Towards a typology of French NP de NP structures or how much possession is there in complex noun phrases with* de *in French?*, in Baron I., Herslund M. and Sørensen F. (a cura di) *Dimensions of possessions*. Benjamins, Amsterdam – Philadephia, 147-167.

Bauer L. 1983. *English word formation.* Cambridge University Press, Cambridge.

Bisetto A. 2004. *Composizione con elementi italiani*, in Grossmann M. e Reiner F. (a cura di) *La formazione delle parole in italiano*. Max Niemeyer, Tübingen, 33-50.

Chiusaroli F. 2012. *Scritture brevi oggi: tra convenzione e sistema*, in Chiusaroli F. e Zanzotto F.M. (a cura di) *Scritture brevi di oggi*, Quaderni di Linguistica Zero, Napoli, 4-44.

Dardano M. 1978. *La formazione delle parole nell'italiano di oggi*. Bulzoni, Roma.

Dell'Orletta F., Venturi G., Cimino A. and Montemagni S. 2014. *T2K²: a System for Automatically Extracting and Organizing Knowledge from Texts* in *Proceedings of 9th Edition of International Conference on Language Resources and Evaluation* (LREC 2014), 26-31 May, Reykjavik, Iceland, 2062-2070 http://www.lrec-conf.org/proceedings/lrec2014/pdf/590_Paper.pdf

Dell'Orletta F. 2009. *Ensemble system for Part-of-Speech tagging*, in *Evaluation of NLP and Speech Tools for Italian*. Reggio Emilia. http://www.evalita.it/sites/evalita.fbk.eu/files/proceedings2009/PoSTagging/POS_ILC.pdf

Heine B. 2006. *Possession*. Cambridge University Press, Cambridge 2006.

Herslund M. and Baron I. 2001. *Introduction: Dimensions of possession*, in Baron I. Herslund M., Sørensen F. (a cura di), *Dimensions of possessions*. Benjamins, Amsterdam – Philadephia, 1-25.

Klajn I. 1972. *Influssi inglesi nella lingua italiana*, Olschki, Firenze.

Logozzo F. 2015. *Sull'apparente omissione della preposizione 'di' davanti a nomi propri, con particolare riguardo a toponimi e marchionimi*. in *Rivista italiana di onomastica* 21 (1): 93-116.

Logozzo F. (in stampa). *Scritture brevi tra sintagmi nominali, preposizionali e aggettivali.* Atti del convegno 'Scritture brevi e varietà diatecnica', Procida 26-28 giugno 2014.

Radimský J. 2013. *Tight N-N compounds in the Italian la Repubblica corpus*, in Baptista, J., Monteleone, M. (a cura di) *Actes du 32ème Colloque international sur le lexique et la grammaire (10-14 septembre 2013, Faro, Portugal)*, 41-47.

Radimský J. 2015. *Les composes de coordination en italien: esquisse d'une typologie*, in *Studia Romanica Posnaniensia* 42 (1): 97-111.

(Better than) State-of-the-Art PoS-tagging for Italian Texts

Fabio Tamburini
FICLIT - University of Bologna, Italy
fabio.tamburini@unibo.it

Abstract

English. This paper presents some experiments for the construction of an high-performance PoS-tagger for Italian using deep neural networks techniques (DNN) integrated with an Italian powerful morphological analyser. The results obtained by the proposed system on standard datasets taken from the EVALITA campaigns show large accuracy improvements when compared with previous systems from the literature.

Italiano. *Questo contributo presenta alcuni esperimenti per la costruzione di un PoS-tagger ad alte prestazioni per l'italiano utilizzando reti neurali 'deep' integrate con un potente analizzatore morfologico. I risultati ottenuti sui dataset delle campagne EVALITA da parte del sistema proposto mostrano incrementi di accuratezza piuttosto rilevanti in confronto ai precedenti sistemi in letteratura.*

1 Introduction

In recent years there were a large number of works trying to push the accuracy of the PoS-tagging task forward using new techniques, mainly from the deep learning domain (Collobert et al., 2011; Søgaard, 2011; dos Santos and Zadrozny, 2014; Huang et al., 2015; Wang et al., 2015; Chiu and Nichols, 2016).

All these studies are mainly devoted to show how to find the best combination of new neural network structures and character/word embeddings for reaching the highest classification performances, and typically present solutions that do not make any use of specific language resources (e.g. morphological analysers, gazetteers, guessing procedures for unknown words, etc.). This is, in general, a very desirable feature because it allows for the production of tools not tied to any specific language, but in various evaluation campaigns, at least for highly-inflected languages as Italian, the results showed quite clearly that this task would benefit from the use of specific and rich language resources (Tamburini, 2007; Attardi and Simi, 2009).

In this study, still work-in-progress, we set-up a PoS-tagger for Italian able to gather the highest classification performances by using any available language resource and the most up-to-date DNN. We used AnIta (Tamburini and Melandri, 2012), one of the most powerful morphological analysers for Italian, based on a wide lexicon (about 110.000 lemmas), for providing the PoS-tagger with a large set of useful information.

2 Input features

The set of input features for each token is basically formed by two different components: the word embedding and some morphological information.

2.1 Word Embeddings

All the embeddings used in our experiments were extracted from the CORIS corpus (Rossini Favretti et al., 2002), a 130Mw synchronic reference corpus for Italian, by using the tool `word2vec`[1] (Mikolov et al., 2013). We added two special tokens to mark the sentence beginning '<s>' and ending '</s>'.

2.2 Morphological features

One of the most useful kind of information that increases the performances of PoS-taggers concerns the list of all possible tags for a single word-form. Having a restricted list of possibility enable the tagger to reduce the search space and force it to take reasonable decisions. The results obtained

[1] https://code.google.com/archive/p/word2vec/

in past PoS-taggers evaluations on Italian agree in suggesting that powerful morphological analysers based on large lexica are invaluable resources to increase tagger accuracy. For these reasons, we extended the word embeddings computed in a completely unsupervised way by concatenating to them a vector containing the possible PoS-tags provided by the AnIta analyser. This tool is also able to identify, through the use of simple regular expressions, numbers, dates, URLs, emails, etc., and assign them the proper tag(s).

2.3 Unknown words handling and Sentence padding

The source of most tagging errors is certainly the presence of the so called 'unknown words', word-forms for which the tagger did not receive any information during the training phase. A morphological analyser based on a large lexicon could certainly alleviate this problem providing information also for word-forms not belonging to the training set, but there are large classes of tokens that cannot be successfully handled by the analyser, for example proper names, foreign words, etc.

In a previous work (Tamburini, 2007b) we showed that using such a powerful morphological analyser, the word-forms not covered by it in real texts belongs at 95% to the class of proper names, adjectives and common nouns and a simple heuristic correctly assigns most of the cases. In this way AnIta always provides one or more PoS-tag hypothesis for each word-form that can be transformed into a binary vector with 1s in correspondence of possible PoS-tags and 0s otherwise, but if the word-form did not have a computed embedding, the first part of the input features would not be defined. For solving such problem, instead of using the common solution of assigning a random vector to all unknown words, we averaged all the embeddings of the other word presenting exactly the same combination of possible PoS-tags.

It is also a common practice to pad sentences, at the beginning and at the end, using random vectors, but we, instead, used the real embeddings computed for the special tokens '<s>' and '</s>', added for this purpose, with the respective tag 'BoS' and 'EoS'. Due to the internal structuring of the used tensor manipulating application (see later), we were forced to add also an out-of-sentence vector to pad sentences to their maximal length, and the correspondent tag OoS.

2.4 Data structuring

We experimented two different ways of structuring the input features for processing:

- `Win`: this mode of organising input data is based on a sliding window that starts from the beginning of each sentence and concatenates word feature vectors into one single vector. Padding is inserted at sentence borders.
- `Seq`: each sentence is managed as one single sequence padded at the borders.

Each network experimented in this study uses one of these two data structuring type.

3 (Deep) Learning Blocks

All the experiments presented in this paper has been performed using Keras[2] a "a minimalist, highly modular neural networks library, written in Python and capable of running on top of either TensorFlow or Theano", two widely used tensor manipulation libraries. Keras provides some basic neural network blocks as well as different learning procedures for the desired network configuration and simple tools for writing new blocks. In our experiments we used some of them, namely multilayer-perceptrons (MLP) and Long Short-Term Memory (LSTM), and we wrote a new block to handle Conditional Random Fields (CRF).

MLP are simple feedforward neural networks with one or more fully-connected hidden layers. We obtained maximum performances using only one hidden layer.

LSTM networks (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) are a kind of recurrent neural network which received a lot of attention in recent years due to their ability of produce good classification results for sequence problems. Their property of preventing the vanishing (and exploding) gradient problem that affects standard recurrent neural networks made them the default choice for solving sequence classification problems inside the DNN framework. Usually this kind of units are arranged to form a bidirectional chain (BiLSTM) for gathering information both from the past and from the future of the input data sequence, a very desirable issue for such kind of classification problems. In all our experiments using BiLSTM we obtained maximum performances by stacking two layers of them, with

[2] https://github.com/fchollet/keras/tree/master/keras

a dropout layer after each of them (Srivastava et al., 2014), and a final dense softmax layer, or a time-distributed-dense softmax layer, feeded by the BiLSTM output.

Linear CRFs are the simpler Probabilistic Graphical Model (PGM) and it has been successfully used in NLP for sequence classification problems (Lafferty et al., 2001). We did some experiments stacking them after the softmax layer.

Figure 1 shows the most complex DNN structure used in out experiments.

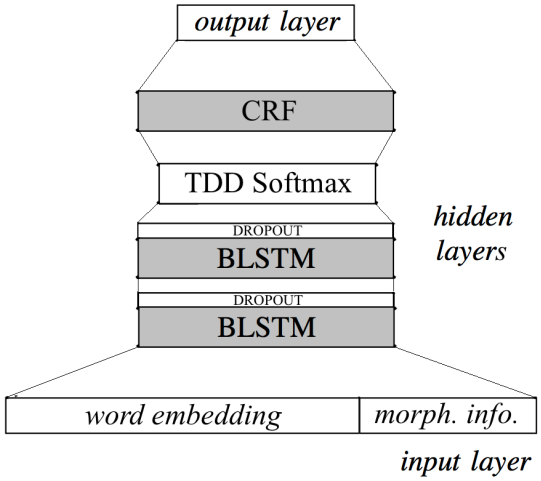


Figure 1: The most complex DNN used in our experiments.

4 Experiments

All the experiments presented in this paper to test the effectiveness of the proposed system refer to two evaluation campaigns organised inside the EVALITA[3] framework. In particular, in 2007 and 2009 were organised specific task to test Italian PoS-taggers performances.

4.1 The EVALITA 2007 evaluation

Two separate data sets were provided: the Development Set (DS), composed of 133,756 tokens, was used for system development and for the training phase, while a Test Set (TS), composed of 17,313 tokens, was used as a reference for systems evaluation. Both contain various documents belonging mainly to journalistic and narrative genres, with small sections containing academic and legal/administrative prose. Each participant was allowed to use any available resource or could freely induce it from the training data.

[3]http://www.evalita.it/

The original PoS-tagging task involved two different tagsets, but our experiments used only the tags and the annotation named 'EAGLES-like'.

The evaluation metrics were based on a token-by-token comparison and only one tag was allowed for each token. The EVALITA metric considered in this study is the *Tagging Accuracy*, defined as the number of correct PoS-tag assignments divided by the total number of tokens in the TS. See (Tamburini, 2007) for further details.

4.2 The EVALITA 2009 evaluation

The DS consisted in 113895 word forms (already divided in a training set - 108,874 tokens - and a validation set - 5021 tokens). The TS consisted of 5066 word forms. The training set is formed by newspaper articles from 'La Repubblica', while the validation and test set contain documents extracted from the Italian Wikipedia. This test the degree of system adaptation to new domains.

The organisers evaluated the results using a coarse grained (37 tags) and a morphed (336 tags) tagsets inserted in a closed/open task framework, but in this study all the results refer to the open task (one can use external resources) on the coarse grained tagset. The evaluation metric is the same described before in section 4.1. See (Attardi and Simi, 2009) for further details.

4.3 Hyper-Parameters

Considering the large number of hyper-parameters involved in the whole procedure, we did not test all the possible combinations; we used, instead, the most common set-up of parameters gathered from the literature. Table 1 outlines the whole set-up for the unmodified hyper-parameters.

`word2vec` **Embed.**		**Feature extraction**	
Hyperpar.	**Value**	**Hyperpar.**	**Value**
type	SkipGr.	window	5
size	100	**Learning Params.**	
(1/2) win.	5	batch (win)	1/4*NU
neg. sampl.	25	batch (seq)	1
sample	1e-4	Opt. Alg.	`Adam`
iter	15	Loss Func.	Categ.CE

Table 1: Unmodified hyper-parameters and algorithms used in our experiments. NU means the number of hidden or LSTM units per layer (the same for all layers). For `Adam` refer to (Kingma and Ba, 2015).

4.4 The Early Stopping Drama

There are some interesting studies (Bengio, 2012; Prechelt, 2012) dealing with the problem of stopping the learning process at the right point; this issue is known as the 'early stopping' problem. Choosing the correct epoch to stop the learning process helps avoiding overfitting on the training set and usually produces systems exhibiting better generalisations. But, how to choose the correct epoch is not simple. The suggestion given in various studies on this topic is to consider a validation set and stop the learning process when the performances on this set do not increase anymore or even decrease, a clear hint of overfitting.

The usual way to set up an experiment following this suggestions involves splitting the gold standard into three different instance sets: the training set, for training, the validation set, to determine the stopping point, and the test set to evaluate the system. However, we are testing our systems on real evaluation data that has been already split by the organisers into development and test set. Thus, we can divide the development set into training/validation set for optimising the hyperparameters and define the stopping epoch, but, for the final evaluation, we would like to train the final system on the complete development set to adhere to the evaluation constraints and to benefit from using more training data.

Having two different training procedures for the optimisation and evaluation phases leads to a more complex procedure for determining the stopping epoch. Moreover, the typical accuracy profile for DNN systems is not smooth and oscillate heavily during training. To avoid any problem in determining the stopping point we smoothed all the profiles using a bezier spline. The procedure we adopted to determine the stopping epoch is (please look at Fig. 2): (1) find the first maximum in the validation smoothed profile - A; (2) find the corresponding value of accuracy on the smoothed training profile - B; (3) find the point in the smoothed development set profile having the same accuracy as in B - C; (4) select the epoch corresponding at point C as the stopping epoch - D.

4.5 Results

Table 2 outlines the systems' accuracies for different configurations for both datasets. We can observe that by using AnIta morphological information, as well as all the techniques described in section 2.3, improves the systems' results by more than 1%. Considering the data structuring described in section 2.4, the management of an entire sentence as a complete sequence allows recurrent configurations to work with larger contexts producing better results. Adding a CRF layer after the BiLSTM seems to slightly improve the performances, but not in a significant way.

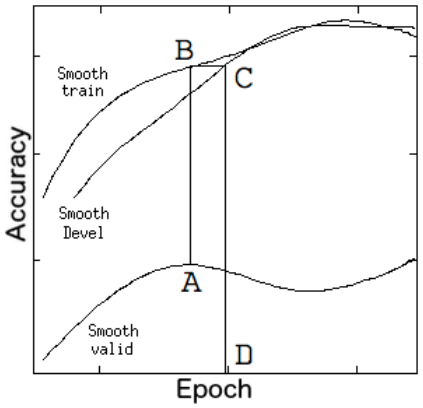


Figure 2: The early stopping procedure.

SYSTEM	TA		Notes
	E07	E09	
MLP-256	96.45	95.57	`Win`=5
MLP-256	97.75	96.84	M,`Win`=5
2-BiLSTM-256	98.12	97.30	M,`Win`=5
2-BiLSTM-256	98.14	97.45	M,`Seq`
2-BiLSTM-256-CRF	**98.18**	**97.48**	M,`Seq`

Table 2: Tagging accuracies (TA) for different configurations for both datasets. ('M' marks the use of AnIta morphological information).

In Table 3 we can see our best system performances, namely AnIta-BiLSTM-CRF, compared with the three best systems of the considered EVALITA campaigns. As you can see, in both cases the proposed system ranked first improving the scoring by large quantities.

5 Conclusions

The proposed system for PoS-tagging, integrating DNNs and a powerful morphological analyser, exhibited very good accuracy results when applied to standard Italian evaluation datasets from the EVALITA campaigns. The information from AnIta proved to be crucial to reach such accuracy values as well as stacked BiLSTM networks processing entire sentence sequences.

EVALITA 2007	
SYSTEM	**TA**
AnIta-BiLSTM-CRF	**98.18**
FBKirst_Zanoli	98.04
UniTn_Baroni	97.89
ILCcnrUniPi_Lenci	97.65
EVALITA 2009	
AnIta-BiLSTM-CRF	**97.48**
UniPi_SemaWiki 2	97.03
UniPi_SemaWiki 1	96.73
UniPi_SemaWiki 4	96.67

Table 3: Participants' results with respect to Tagging Accuracy (TA) at EVALITA 2007 and 2009.

We have to further test different DNN configurations and their integration with other kind of PGMs as well as make more experiments with different hyperparameters.

References

Giuseppe Attardi and Maria Simi. 2009. Overview of the EVALITA 2009 Part-of-Speech Tagging Task. In *Proc. of Workshop Evalita 2009*.

Yoshua Bengio. 2012. Practical Recommendations for Gradient-Based Training of Deep Architectures. In Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade: Second Edition*, pages 437–478. Springer Berlin Heidelberg, Berlin, Heidelberg.

Jason Chiu and Eric Nichols. 2016. Sequential Labeling with Bidirectional LSTM-CNNs. In *Proc. International Conf. of Japanese Association for NLP*, pages 937–940.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537.

Cicero dos Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *Proc. of the 31st International Conference on Machine Learning, JMLR*, volume 32. JMLR W&CP.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *ArXiv e-prints, 1508.01991*.

D.P. Kingma and J.L. Ba. 2015. Adam: a method for stochastic optimization. In *Proc. International Conference on Learning Representations - ICLR.*, pages 1–13.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proc. of Workshop at ICLR*.

Lutz Prechelt. 2012. Early Stopping — But When? In Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade: Second Edition*, pages 53–67. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rema Rossini Favretti, Fabio Tamburini, and Cristiana De Santis. 2002. CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In Andrew Wilson, Paul Rayson, and Tony McEnery, editors, *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, pages 27–38. Lincom-Europa, Munich.

Anders Søgaard. 2011. Semi-supervised condensed nearest neighbor for part-of-speech tagging. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 48–52, Portland, Oregon, USA.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Fabio Tamburini and Matias Melandri. 2012. AnIta: a powerful morphological analyser for Italian. In *Proc. 8th International Conference on Language Resources and Evaluation - LREC 2012*, pages 941–947, Istanbul.

Fabio Tamburini. 2007. EVALITA 2007: the Part-of-Speech Tagging Task. *Intelligenza Artificiale*, IV(2):4–7.

Fabio Tamburini. 2007b. CORISTagger: a high-performance PoS tagger for Italian. Intelligenza Artificiale. *Intelligenza Artificiale*, IV(2):14–15.

Peilu Wang, Yao Qian, Frank. K Soong, Lei He, and Hai Zhao. 2015. A Unified Tagging Solution: Bidirectional LSTM Recurrent Neural Network with Word Embedding. *ArXiv e-prints, 1511.00215*.

Tree Kernels-based Discriminative Reranker for Italian Constituency Parsers

Antonio Uva[†] and **Alessandro Moschitti**
[†]DISI, University of Trento, 38123 Povo (TN), Italy
Qatar Computing Research Institute, HBKU, 5825 Doha, Qatar
antonio.uva@unitn.it amoschitti@gmail.com

Abstract

English. This paper aims at filling the gap between the accuracy of Italian and English constituency parsing: firstly, we adapt the Bllip parser, i.e., the most accurate constituency parser for English, also known as Charniak parser, for Italian and trained it on the Turin University Treebank (TUT). Secondly, we design a parse reranker based on Support Vector Machines using tree kernels, where the latter can effectively generalize syntactic patterns, requiring little training data for training the model. We show that our approach outperforms the state of the art achieved by the Berkeley parser, improving it from 84.54 to 86.81 in labeled F1.

Italiano. *Questo paper mira a colmare il gap di accuratezza tra il constituency parsing dell'Italiano e quello Inglese: come primo miglioramento, abbiamo adattato il parser a costituenti per l'Inglese, Bllip, anche noto come Charniak parser, per l'Italiano e lo abbiamo addestrato sul Turin University Treebank. In seguito, abbiamo progettato un reranker basato sulle Macchine a Vettori di Supporto che usano kernel arborei, i quali possono efficacemente generalizzare pattern sintattici, richiedendo pochi dati di training per addestrare il modello. Il nostro approccio supera lo stato dell'arte ottenuto con il Berkeley parser, migliorando la labeled F1 da 84.54 a 86.81.*

1 Introduction

Constituency Syntactic parsing is one of the most important research lines in Computational Linguistics. Consequently, a large body of work has been also devoted to its design for Italian language (Bosco et al., 2007; Bosco et al., 2009; Bosco and Mazzei, 2011). However, the accuracy reported for the best parser is still far behind the state of the art of other languages, e.g., English.

One noticeable attempt to fill this technological gap was carried out in the EvalIta challenge, which proposed a parsing track on both dependency and constituency parsing for Italian. Among the several participant systems, the Berkeley parser (Petrov and Klein, 2007) gave the best result (Lavelli and Corazza, 2009; Lavelli, 2011).

At the beginning, the outcome for constituency parsing computed on TUT (Bosco et al., 2009) was much lower than the one obtained for English on the Penn Treebank (Marcus et al., 1993). In the last EvalIta edition, such gap diminished as the Italian parser labeled *F1* increased from 78.73% (EvalIta 2009) to 82.96% (EvalIta 2011). Some years later the parser F1 improved to 83.27% (Bosco et al., 2013). However, the performance of the best English parser (McClosky et al., 2006), i.e., 92.1%, is still far away. The main reason for such gap is the difference in the amount of training data available for Italian compared to English. In fact, while Penn Treebank contains $49,191$ sentences/trees, TUT only contains $3,542$ sentences/trees.

In presence of scarcity of training data, a general solution for increasing the accuracy of a machine learning-based system is the use of more general features. This way, the probability of matching training and testing instance representations is larger, allowing the learning process to find more accurate optima. In case of syntactic parsing, we need to generalize either lexical or syntactic features, or possibly both. However, modeling such generalization in state-of-the-art parser algorithms such as the Bllip[1] (Charniak, 2000; Charniak and Johnson, 2005) is rather challenging. In particular, the space of all possible syntactic patterns is very large and cannot be explicitly coded

[1] https://github.com/BLLIP/bllip-parser

in the model. An easier solution consists in using such features in a simpler model, which can be trained to improve the outcome of the main parser, e.g., selecting one of its best hypotheses. In particular, tree kernels (TKs) by Moschitti (2006) can be used for encoding an exponential number of syntactic patterns in parse rerankers.

In this work, we aim at filling the gap between English and Italian constituency parsing: firstly, we adapted Bllip parser, i.e., the most accurate constituency parser for English, also known as Charniak parser, for Italian and trained it on TUT. We designed various configuration files for defining specific labels for TUT by also defining their type, although we did not encode head-finding rules for Italian, needed to complete the parser adaptation.

Secondly, we apply rerankers based on Support Vector Machines (SVMs) using TKs to the *k*-best parses produced by Bllip, with the aim of selecting its best hypotheses. TKs allow us to represent data using the entire space of subtrees, which correspond to syntactic patterns of different level of generality. This representation enables the training of the reranker with little data. Finally, we tested our models on TUT, following the EvalIta setting and compare with other parsers. For example, we observed an improvement of about 2%, over the Berkeley parser, i.e., 86.81 vs. 84.54.

2 Bllip parser

The Bllip parser is a lexicalized probabilistic constituency parser. It can be considered a smoothed PCFG, whose non-terminals encode a wide variety of manually chosen conditioning information, such as heads, governors, etc. Such information is used to derive probability distributions, which, in turn, are utilized for computing the likelihood of constituency trees being generated. As described by McClosky et al. (2006), Bllip uses five distributions, i.e., the probabilities of the (i) constituent heads, (ii) constituent part-of-speeches (PoS), (iii) head-constituents, (iv) left-of-head and (v) right-of-head constituents. Each probability distribution is conditioned by five or more features and backed-off by the probability of lower-order models in case of rare feature configurations. The variety of information needed by Bllip to work properly makes its configuration much harder than for other parsers, e.g., the Berkeley' one. However, Bllip is faster to train than other off-the-shelf parsers.

2.1 Adapting Bllip to Italian Language

Bllip adaptation required to create various configuration files. For example, PoS and bracket labels observed in training and development sets must be defined in a file named *terms.txt*. As labels present in the TUT are different from those of the Penn Treebank[2], we added them in such file. Then, we specified the type of labels present in the data, i.e., constituent type, open-class PoS, punctuation, etc.

Finally, it should be noted that, since Bllip is lexicalized, head-finding rules for Italian should be specified in the file, *headInfo.txt*. For example, the rule, $ADJP \xrightarrow{r} JJ$, specifies that the head of an adjective phrase (ADJP) is the right-most adjective (JJ). Due to time restriction, we used the default Bllip rules and leave this task as our short-term future work.

3 Tree Kernel-based Reranker

We describe three types of TKs and the Preference Reranker approach using them.

3.1 Tree kernels

TKs can be used for representing arbitrary tree structures in kernel machines, e.g., SVMs. They are a viable alternative to explicit feature design as they implement the scalar products between feature vectors as a similarity between two trees. Such scalar product is computed using efficient algorithms and it is basically equal to the number of the common subparts of the two trees.

Syntactic Tree Kernels (STK) count the number of common tree fragments, where the latter (i) contain more than two nodes and (ii) each node is connected to either all or none of its children. We also used a variant, called STK_b, which adds the number of common leaves of the comparing trees in the final subpart count.

Partial Tree Kernels (PTK) counts a larger class of tree fragments, i.e., any subset of nodes, where the latter are connected in the original trees: clearly, PTK is a generalization of STK.

3.2 Preference Reranker

Preference reranking is cast as a binary classification problem, where each instance is a pair $\langle h_i, h_j \rangle$ of tree hypotheses and the classifier decides if h_i is better than h_j. The positive training examples are the pairs, $\langle h_1, h_i \rangle$, where h_1 has the highest *F1* with respect to the gold standard among the candidate hypotheses. The negative examples are

[2]For example, the PoS-tag NN in Penn Treebank corresponds to tag NOU∼CS in TUT

Models	sentences $\leq$ 40 words				All sentences			
	LR	LP	LF	EMR	LR	LP	LF	EMR
Berkeley (Bosco et al., 2013)	83.45	84.48	83.96	24.91	82.78	83.76	83.27	23.67
Berkeley (our model)	85.31	85.76	85.53	27.76	84.35	84.72	84.54	26.33
Bllip base model	85.90	86.67	86.28	29.54	85.26	85.97	85.61	28.00
STK	86.16	87.02	86.59	30.96	85.73	86.38	86.05	29.33
STK_b	86.36	87.21	86.78	**31.67**	85.89	86.53	86.21	**30.00**
PTK	**86.82**	**87.95**	**87.38**	30.96	**86.33**	**87.29**	**86.81**	29.67

Table 1: Comparative results on the test set. **LR/LP/LF** = labeled recall/precision/$F1$. **EMR** = percentage of sentences where recall and precision are 100%. STK- and STK_b-based rerankers use 20-best hypotheses, while PTK-based reranker use 30-best hypotheses.

obtained inverting the hypotheses in the pairs, i.e., $\langle h_i, h_1 \rangle$. If the hypotheses have the same score, the pair is not included in the training set. At classification time all pairs $\langle h_i, h_j \rangle$ generated from the *k*-best hypotheses are classified. A positive classification is a vote for h_i, whereas a negative classification is a vote for h_j. The hypothesis associated with the highest number of votes (or highest sum of classifier scores) is selected as the best parse.

4 Experiments

In these experiments, we first report on the performance of Bllip for Italian and compare it with the Berkeley parser. Then, we show that our parse reranker can be very effective, even in case of use of small training data.

4.1 Experimental Setup

Parsing data. The data for training and testing the constituency parsers come from the TUT project [3], developed at the University of Turin. There have been several releases of the dataset: we used the latest version from EvalIta 2011. The training set is composed of $3,542$ sentences, while the test set contains 300 sentences. The set of PoS-tags includes 97 tags: 68 encoding morphological features (out of which 19 basic tags) for pre-terminal symbols (e.g., ADJ, ADVB, NOUN, etc.) and 29 non-terminal symbols for phrase constituents (e.g., ADJP, ADVP, NP, etc.).

Reranking Data. To generate the data for training the reranker, we apply 10-fold cross validation to the official TUT training set: we train the based parser on 9 folds and applied it to the remaining fold to generate the n-best trees for each of its sentences. Then, we merge all the 10-labeled folds to produce the training set of the reranker. This way, we avoid the bias a parser would have if applied to the data used for training it. For generating the test data of the reranker, we simply apply the base parser (trained on all TUT training data) to the TUT test set and generate n-hypotheses for each sentence.

SVM Reranker. We train the reranker using SVM-light-TK, which takes both feature vectors and trees as input to learn a classification model. The features used for reranking constituency trees are: (i) the probability and the (inverse) rank of the hypotheses provided by Bllip and (ii) the entire syntactic trees used with two types of kernels, STK and PTK, described in Sec. 3.

Measures. For evaluating the parsers, we used the EVALB scoring program, which reports the Labeled Precision (LP), Labeled Recall (LR), Labeled *F1* (LF) and Exact Match Rate (EMR). According to the official EvalIta procedure for evaluating the participant system output, we did not score the TOP label, ignore all functional labels attached to non-terminals and include punctuation in the scoring procedure.

4.2 Bllip base parser results

We divided the training set in train and validation sets, where the latter is composed of the last 50 sentences of each of the six sections of the former for a total of 300 sentences. We train the models on the training set and tune parameters on the validation set. Then, we applied the learned model to the 300 sentences of the test set. Table 1 shows the results obtained by the Bllip base parser on the TUT test set. Our parser obtained an LF of 86.28% for sentences with less than 40 words and a score of 85.61% for all sentences.

4.3 Comparison with the Berkeley parser

Table 1 also reports the results of the Berkeley parser obtained by Bosco et al. (2013). For comparison purposes, we trained our own version of the Berkeley parser. In particular, we trained the parser for 5 split-merge cycles on the whole training set. We selected such number of cycles applying 10-fold cross validation on the training set. Similarly to Bosco et al. (2013), we specialized

[3] http://www.di.unito.it/~tutreeb/

Models	10-best				20-best				30-best			
	Tree		Tree + feat.		Tree		Tree + feat.		Tree		Tree + feat.	
	len≤40	All	≤40	All	len≤40	All	len≤40	All	len≤40	All	len≤40	All
Bllip base model	**86.28**	**85.61**	86.28	85.61	**86.28**	85.61	86.28	85.61	86.28	85.61	86.28	85.61
STK	84.95	84.49	86.31	85.69	84.70	84.16	86.59	86.05	84.99	84.45	86.55	86.00
STK_b	85.05	84.52	86.31	85.69	84.92	84.35	86.78	86.21	84.92	84.38	86.62	86.06
PTK	86.02	85.46	**87.34**	**86.65**	85.89	**86.41**	**87.37**	**86.79**	**86.42**	**85.92**	**87.38**	**86.81**

Table 2: Reranker performance: the first row reports the number n of the best hypotheses used during training. The second row shows the used group of features: Tree or Tree + feat, while the third row illustrates the parse results (LF) for two sentence groups: sentences with ≤ 40 words and all sentences.

punctuation symbols to more specific tags. However, we used full PoS-tags, as they gave the best results in cross-fold validation. Indeed, the table shows that our own version of the Berkeley parser outperforms the version the one of Bosco et al. (2013) by 1.27 absolute percent points (84.54 vs. 83.27). The table also reports the results of the Bllip parser, which outperforms the best result obtained by the Berkeley parser by 1.07% in LF, i.e., 85.61 vs. 84.54.

4.4 Reranking using different TKs

Table 2 reports the LF obtained by different reranking models, varying: (i) the type of TKs, (ii) the group of features (i.e., either trees or trees + feat.) and (iii) the number, n, of parse trees used to generate the reranker training data. More in particular, we experimented with three values for n, i.e., 10-, 20- and 30-best parse trees. As it can be seen from the table, PTK constantly outperforms STK and STK_b for any number of parse hypotheses. This indicates that the subtree features generated by PTK, which include nodes with any subset of the children in the original tree, are useful for improving the parser accuracy.

Very interestingly, the performance of all models when trained on 30-best trees give either worse results (e.g., STK_b and STK) or very little improvement (e.g., PTK) than training on 20-best parse trees. This may suggest that adding too many negative examples, largely populating the lower part of the n-best list may be detrimental.

The bottom part of Table 1 shows standard parser evaluation metrics for different reranking models using different kernel types: only the kernel models with the highest LF from Table 2 are reported. The method shows an 1.2% absolute improvement in LF (from 85.61% to 86.81%) on all the sentences over the base-parser model (i.e., the baseline) when using the most powerful kernel, PTK, and 30-best hypotheses. STK and STK_b show a lower improvement over the baseline of 0.44% and 0.6%, respectively. One interesting fact is the following: while PTK gives better results in terms of LF, STK and STK_b perform better in terms of EMR, i.e., the percentage of sentence parse completely matching gold trees. This is rather intuitive as the name suggests, Partial Tree Kernel generates partial subtrees, i.e., partial production rules as patterns. On one hand, this can improve the ability of matching syntactic patterns, thus capturing rules partially expressed by more than one support vector. On the other hand, the precision in capturing complete patterns, i.e., regarding a complete tree is intuitively decreased.

5 Related Work and Conclusions

This work was inspired by Collins and Duffy (2002) and Collins and Koo (2005), who explored discriminative approaches for ranking problems. Their studies were limited to WSJ, though, and did not explore the use of max-margin classifiers, i.e., SVMs. The first experiments with SVMs and TKs were conducted by Shen and Joshi (2003), who proposed a new SVM-based voting algorithm making use of preference reranking.

In this paper, we adapted the Charniak parser for Italian gaining an improvement of 1.07% over the Berkeley model (indicated by EvalIta as the state of the art for Italian). Then, our TK-based reranker further improved it up to 2 absolute percent points. It should also be noted that our best reranking result is 3.54 absolute points better than the best outcome reported in (Bosco et al., 2013), i.e., 83.27.

In the future, we would like to integrate (i) the features developed in the reranking software available by Johnson and Ural (2010) in our model for further improving it, (ii) generalizing lexical features (e.g., embeddings, brown clusters) and including similarity measures in PTK, i.e., SPTK (Croce et al., 2011).

Acknowledgments

A special thank is due to Alberto Lavelli and Alessandro Mazzei for enabling us to carry out an exact comparison with their parser.

References

[Bosco and Mazzei2011] Cristina Bosco and Alessandro Mazzei. 2011. The evalita 2011 parsing task: the constituency track. *Working Notes of EVALITA*.

[Bosco et al.2007] Cristina Bosco, Alessandro Mazzei, and Vincenzo Lombardo. 2007. Evalita parsing task: an analysis of the first parsing system contest for italian. *Intelligenza artificiale*, 12:30–33.

[Bosco et al.2009] Cristina Bosco, Alessandro Mazzei, and Vincenzo Lombardo. 2009. Evalita09 parsing task: constituency parsers and the penn format for italian. *Proceedings of EVALITA*, 9:1794–1801.

[Bosco et al.2013] Cristina Bosco, Alessandro Mazzei, and Alberto Lavelli. 2013. Looking back to the evalita constituency parsing task: 2007-2011. In *Evaluation of Natural Language and Speech Tools for Italian*, pages 46–57. Springer.

[Charniak and Johnson2005] Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 173–180. Association for Computational Linguistics.

[Charniak2000] Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 132–139. Association for Computational Linguistics.

[Collins and Duffy2002] Michael Collins and Nigel Duffy. 2002. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 263–270. Association for Computational Linguistics.

[Collins and Koo2005] Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1):25–70.

[Croce et al.2011] Danilo Croce, Alessandro Moschitti, and Roberto Basili. 2011. Structured lexical similarity via convolution kernels on dependency trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1034–1046. Association for Computational Linguistics.

[Johnson and Ural2010] Mark Johnson and Ahmet Engin Ural. 2010. Reranking the berkeley and brown parsers. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 665–668. Association for Computational Linguistics.

[Lavelli and Corazza2009] Alberto Lavelli and Anna Corazza. 2009. The berkeley parser at the evalita 2009 constituency parsing task. In *EVALITA 2009 Workshop on Evaluation of NLP Tools for Italian*.

[Lavelli2011] Alberto Lavelli. 2011. The berkeley parser at the evalita 2011 constituency parsing task. In *Working Notes of EVALITA*.

[Marcus et al.1993] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

[McClosky et al.2006] David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics*, pages 152–159. Association for Computational Linguistics.

[Moschitti2006] Alessandro Moschitti. 2006. Efficient convolution kernels for dependency and constituent syntactic trees. In *European Conference on Machine Learning*, pages 318–329. Springer.

[Petrov and Klein2007] Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *HLT-NAACL*, volume 7, pages 404–411.

[Shen and Joshi2003] Libin Shen and Aravind K Joshi. 2003. An svm based voting algorithm with application to parse reranking. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 9–16. Association for Computational Linguistics.

Comparing State-of-the-art Dependency Parsers on the Italian Stanford Dependency Treebank

Alberto Lavelli
FBK-irst
via Sommarive, 18 - Povo
I-38123 Trento (TN) - ITALY
lavelli@fbk.eu

Abstract

English. In the last decade, many accurate dependency parsers have been made publicly available. It can be difficult for non-experts to select a good off-the-shelf parser among those available. This is even more true when working on languages different from English, because parsers have been tested mainly on English treebanks. Our analysis is focused on Italian and relies on the Italian Stanford Dependency Treebank (ISDT). This work is a contribution to help non-experts understand how difficult it is to apply a specific dependency parser to a new language/treebank and choose a parser that meets their needs.

Italiano. *Nell'ultimo decennio sono stati resi disponibili molti analizzatori sintattici a dipendenza. Per i non esperti può essere difficile sceglierne uno pronto all'uso tra quelli disponibili. A maggior ragione se si lavora su lingue diverse dall'inglese, perché gli analizzatori sono stati applicati soprattutto su treebank inglesi. La nostra analisi è dedicata all'italiano e si basa sull'Italian Stanford Dependency Treebank (ISDT). Questo articolo è un contributo per aiutare i non esperti a capire quanto è difficile applicare un analizzatore a una nuova lingua/treebank e a scegliere quello più adatto.*

1 Introduction

In the last decade, there has been an increasing interest in dependency parsing, witnessed by the organisation of various shared tasks, e.g. Buchholz and Marsi (2006), Nivre et al. (2007), Seddah et al. (2013), Seddah et al. (2014). Concerning Italian, there have been tasks on dependency parsing in the first four editions of the EVALITA evaluation campaign (Bosco et al., 2008; Bosco et al., 2009; Bosco and Mazzei, 2011; Bosco et al., 2014). In the 2014 edition, the task on dependency parsing exploited the Italian Stanford Dependency Treebank (ISDT), a treebank featuring an annotation based on Stanford Dependencies (de Marneffe and Manning, 2008).

This paper is a follow-up of Lavelli (2014b) and reports the experience in applying an updated list of state-of-the-art dependency parsers on ISDT. It can be difficult for non-experts to select a good off-the-shelf parser among those available. This is even more true when working on languages different from English, given that parsers have been tested mainly on English treebanks (and in particular on the WSJ portion of the PennTreebank). This work is a contribution to help practitioners understand how difficult it is to apply a specific dependency parser to a new language/treebank and choose a parser to optimize their desired speed/accuracy trade-off.

As in many other NLP fields, there are very few comparative articles where different parsers are directly run by the authors and their performance compared (Daelemans and Hoste, 2002; Hoste et al., 2002; Daelemans et al., 2003). Most of the papers simply present the results of a newly proposed approach and compare them with the results reported in previous articles. In other cases, the papers are devoted to the application of the same tool to different languages/treebanks. A notable exception is the study reported in Choi et al. (2015), where the authors present a comparative analysis of ten leading statistical dependency parsers on a multi-genre corpus of English.

It is important to stress that the comparison presented in this paper concerns tools used more or less out of the box and that the results cannot be used to compare specific characteristics like: parsing algorithms, learning systems, ...

2 Parsers

The choice of the parsers used in this study started from the ones already applied in a previous study (Lavelli, 2014b), i.e. MaltParser, the MATE dependency parsers, TurboParser, and ZPar. We then identified a few other freely available dependency parsers that have shown state-of-the-art performance. Some of such parsers are included in the study in Choi et al. (2015) and others have been made publicly available more recently. The additional parsers included in this paper are DeSR, the Stanford Neural Network dependency parser, EmoryNLP, RBG, YARA Parser, and LSTM parser.

Differently from what was done in the previous study, this time we have not included approaches based on combination of parsers' results, such as ensemble or stacking. They usually obtain top performance (see e.g. Attardi and Simi (2014) at EVALITA 2014) but in this case we focus on simplicity and ease of use rather than on absolute performance. Below you may find short descriptions of the parsers reported in the paper.

MaltParser (Nivre et al., 2006) (version 1.8) implements the transition-based approach to dependency parsing, which has two essential components: (i) a nondeterministic transition system for mapping sentences to dependency trees; (ii) a classifier that predicts the next transition for every possible system configuration. MaltParser includes different built-in transition systems, different classifiers and techniques for recovering non-projective dependencies with strictly projective parsers.

The MATE tools[1] include both a graph-based parser (Bohnet, 2010) and a transition-based parser (Bohnet and Nivre, 2012; Bohnet and Kuhn, 2012). For the languages of the 2009 CoNLL Shared Task, the graph-based MATE parser reached accuracy scores similar or above the top performing systems with fast processing (obtained with the use of Hash Kernels and parallel algorithms). The transition-based MATE parser is a model that takes into account complete structures as they become available to rescore the elements of a beam, combining the advantages of transition-based and graph-based approaches.

TurboParser (Martins et al., 2013)[2] (version 2.3) is a C++ package that implements non-projective graph-based dependency parsing exploiting third-order features. The approach uses AD^3, an accelerated dual decomposition algorithm extended to handle specialized head automata and sequential head bigram models.

ZPar (Zhang and Nivre, 2011) (version 0.75) is a transition-based parser implemented in C++. ZPar supports multiple languages and multiple grammar formalisms. ZPar has been most heavily developed for Chinese and English, while it provides generic support for other languages. It leverages a global discriminative training and beam-search framework.

DeSR (Attardi and Dell'Orletta, 2009) version 1.4.3 is a shift-reduce dependency parser, which uses a variant of the approach of Yamada and Matsumoto (2003). It is capable of dealing directly with non-projective parsing, by means of specific non-projective transition rules (Attardi, 2006). It is highly configurable: one can choose which classifier (e.g. SVM or Multi-Layer Perceptron) and which feature templates to use, and the format of the input, just by editing a configuration file.

EmoryNLP (Choi and McCallum, 2013)[3] (previously ClearNLP) dependency parser (version 1.1.1) uses a transition-based, non-projective parsing algorithm showing a linear-time speed for both projective and non-projective parsing.

The Stanford neural network dependency parser (Chen and Manning, 2014)[4] is a transition-based parser which produces typed dependency parses using a neural network which uses word embeddings as features besides forms and POS tags. It also uses no beam.

RBG (Lei et al., 2014; Zhang et al., 2014b; Zhang et al., 2014a)[5] is based on a low-rank factorization method that enables to map high dimensional feature vectors into low dimensional representations. The method maintains the parameters as a low-rank tensor to obtain low dimensional representations of words in their syntactic roles, and to leverage modularity in the tensor for easy training with online algorithms.

YARA Parser (Rasooli and Tetreault, 2015)[6] is an implementation of the arc-eager dependency model. It uses an average structured perceptron

[1] https://code.google.com/p/mate-tools/

[2] http://www.ark.cs.cmu.edu/TurboParser/

[3] http://nlp.mathcs.emory.edu/

[4] http://nlp.stanford.edu/software/nndep.shtml

[5] https://github.com/taolei87/RBGParser

[6] https://github.com/yahoo/YaraParser

as classifier and a beam size of 64. The feature setting is from Zhang and Nivre (2011) with additional Brown cluster features.

LSTM parser (Dyer et al., 2015; Ballesteros et al., 2015)[7] is a transition based dependency parser with state embeddings computed by LSTM RNNs and an alternative char-based model exploiting character embeddings as features. Both the models are applied in the experiments.

The list of parsers is still in progress because the field of dependency parsing is in constant evolution. In mid-May, SyntaxNet, the dependency parser by Google, was made publicly available; a few days later BIST parser (that claims to be "A faster and more accurate parser than Google's McParseface") was announced to become public.

SyntaxNet (Andor et al., 2016)[8], BIST parser (Kiperwasser and Goldberg, 2016)[9], and spaCy[10] are not yet included in our study because we are still trying to make them working in a satisfactory way.

3 Data Set

The experiments reported in the paper are performed on the Italian Stanford Dependency Treebank (ISDT) (Bosco et al., 2013) version 2.0 released in the context of the EVALITA 2014 evaluation campaign on Dependency Parsing for Information Extraction (Bosco et al., 2014)[11]. There are three main novelties with respect to the previously available Italian treebanks: (i) the size of the dataset, much bigger than the resources used in the previous EVALITA campaigns; (ii) the annotation scheme, compliant to *de facto* standards at the level of both representation format (CoNLL) and adopted tagset (Stanford Dependency Scheme); (iii) its being defined with a specific view to supporting information extraction tasks, a feature inherited from the Stanford Dependency scheme.

The training set contains 7,414 sentences (158,561 tokens), the development set 564 sentences (12,014 tokens), and the test set 376 sentences (9,066 tokens).

[7] https://github.com/clab/lstm-parser

[8] https://github.com/tensorflow/models/tree/master/syntaxnet

[9] https://github.com/elikip/bist-parser

[10] https://spacy.io/, https://github.com/spacy-io/spaCy

[11] http://www.evalita.it/2014/tasks/dep_par4IE.

4 Experiments

The level of interaction with the authors of the parsers varied. For MaltParser, MATE parsers, TurboParser, and ZPar we have mainly exploited the experience gained in the context of EVALITA 2014 (Lavelli, 2014a).

Concerning MaltParser, in addition to using the best performing configuration at EVALITA 2014 (Nivre's arc-eager, PP-head), we have used MaltOptimizer[12] (Ballesteros and Nivre, 2014) to identify the best configuration. This was done to be fair to the other parsers, given that MaltParser's best configuration was the result of extensive feature selection at the CoNLL 2006 shared task. According to MaltOptimizer, the best configuration is Nivre's arc-standard.

As for the MATE parsers, we have applied both the graph-based and the transition-based parser.

TurboParser was applied using the three standard configurations (basic, standard, full).

Concerning ZPar, the main difficulty emerged in 2014 (i.e., the fact that sentences with more than 100 tokens needed 70 GB of RAM) is no longer present and so its use is rather straightforward.

As for the new parsers, the only problems during installation concerned an issue with the version of the C++ compiler needed for successfully compiling LSTM parser.

For some of the parsers there is the possibility of exploiting word embeddings (RBG, Stanford parser, LSTM, EmoryNLP) or Brown clustering (YARA parser). As for word embeddings (WEs), we exploited the following (both built using `word2vec`):

- word embeddings of size 300 learned on WackyPedia/itWaC (a corpus of more than 1 billion tokens)[13];
- word embeddings of size 50 produced in the project PAISÀ (Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati)[14] on a corpus of 250 million tokens.

In general, WEs of size 300 produced an increase in performance, while those of size 50 produced a decrease in performance (with the excep-

[12] http://nil.fdi.ucm.es/maltoptimizer/

[13] http://clic.cimec.unitn.it/~georgiana.dinu/down/

[14] http://www.corpusitaliano.it/en/index.html

	LAS	UAS	LA
RBG (full, w/ WEs - size=300)	87.72	90.00	93.03
RBG (standard, w/ WEs - size=300)	87.63	89.91	93.03
RBG (full, w/o WEs)	87.33	89.94	92.41
RBG (standard, w/o WEs)	87.33	89.86	92.43
MATE transition-based	87.07	89.69	92.30
MATE graph-based	86.91	89.53	92.67
TurboParser (model_type=full)	86.53	89.20	92.22
TurboParser (model_type=standard)	86.45	88.96	92.29
ZPar	86.32	88.65	92.40
LSTM (EMNLP 2015, char-based w/ WEs - size=300)	86.07	88.96	92.15
RBG (basic, w/o WEs)	85.99	88.53	91.71
MaltParser (Nivre eager -PP head)	85.82	88.29	91.62
EmoryNLP (w/o WEs)	85.30	87.68	91.51
TurboParser (model_type=basic)	84.90	87.28	91.26
DeSR (MLP)	84.61	87.18	90.79
MaltParser (Nivre standard - MaltOptimizer)	84.44	87.17	90.94
LSTM (ACL 2015, w/ WEs - size=300)	84.20	87.13	90.80
LSTM (EMNLP 2015, char-based w/o WEs)	84.13	87.32	90.75
YARA parser (w/o BCs)	83.87	86.79	90.34
LSTM (ACL 2015, w/o WEs)	83.86	86.95	90.56
Stanford NN dependency parser (w/ WEs - size=50)	83.68	86.50	90.85

Table 1: Results on the EVALITA 2014 test set without considering punctuation, in terms of Labeled Attachment Score (LAS), Unlabeled Attachment Score (UAS) and Label Accuracy (LA).

tion of the Stanford NN dependency parser, which produced results comparable to other parsers with WEs of size 50 and absurdly low results with those of size 300). We were not able to successfully run the EmoryNLP parser with WEs. The use of WEs needs further investigation.

As for the use of Brown clusters (BCs), we are still working to build suitable resources for Italian, so the YARA Parser was used with standard settings and without Brown clusters.

The experiments were performed using the splits provided by the EVALITA 2014 organisers: training on the training set, tuning (if any) using the development set and final test on the test set.

In Table 1 we report the parser results ranked according to decreasing Labeled Accuracy Score (LAS), not considering punctuation. We have grouped together the parsers if the differences between their results (in terms of LAS) are not statistically significant (computation performed using DEPENDABLE (Choi et al., 2015)).

The results obtained by the best system submitted to the official evaluation at EVALITA 2014 (Attardi and Simi, 2014) are: 87.89 (LAS), 90.16 (UAS). More details about the task and the results obtained by the participants are available in Bosco et al. (2014).

5 Conclusions

In the paper we have reported on work in progress on the comparison between several state-of-the-art dependency parsers on the Italian Stanford Dependency Treebank (ISDT).

We are already working to widen the scope of the comparison including more parsers and to perform an analysis of the results obtained by the different parsers considering not only their performance but also their behaviour in terms of speed, CPU load at training and parsing time, ease of use, licence agreement, ...

The next step would be to apply the parsers in a multilingual setting, exploiting the availability of treebanks based on Universal Dependencies in many languages (Nivre et al., 2016)[15].

Acknowledgments

We thank the authors of the parsers for making them freely available, for kindly answering our questions and for providing useful suggestions. We thank the reviewers for valuable suggestions to improve this article.

[15] http://universaldependencies.org/

References

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. *CoRR*, abs/1603.06042.

Giuseppe Attardi and Felice Dell'Orletta. 2009. Reverse revision and linear tree combination for dependency parsing. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 261–264, Boulder, Colorado, June. Association for Computational Linguistics.

Giuseppe Attardi and Maria Simi. 2014. Dependency parsing techniques for information extraction. In *Proceedings of EVALITA 2014*.

Giuseppe Attardi. 2006. Experiments with a multilanguage non-projective dependency parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 166–170, New York City, June. Association for Computational Linguistics.

Miguel Ballesteros and Joakim Nivre. 2014. MaltOptimizer: Fast and effective parser optimization. *Natural Language Engineering*, FirstView:1–27, 10.

Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015. Improved transition-based parsing by modeling characters instead of words with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 349–359, Lisbon, Portugal, September. Association for Computational Linguistics.

Bernd Bohnet and Jonas Kuhn. 2012. The best of both worlds – a graph-based completion model for transition-based parsers. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 77–87, Avignon, France, April. Association for Computational Linguistics.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1455–1465, Jeju Island, Korea, July. Association for Computational Linguistics.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China, August. Coling 2010 Organizing Committee.

Cristina Bosco and Alessandro Mazzei. 2011. The EVALITA 2011 parsing task: the dependency track. In *Working Notes of EVALITA 2011*, pages 24–25.

Cristina Bosco, Alessandro Mazzei, Vincenzo Lombardo, Giuseppe Attardi, Anna Corazza, Alberto Lavelli, Leonardo Lesmo, Giorgio Satta, and Maria Simi. 2008. Comparing Italian parsers on a common treebank: the EVALITA experience. In *Proceedings of LREC 2008*.

Cristina Bosco, Simonetta Montemagni, Alessandro Mazzei, Vincenzo Lombardo, Felice DellOrletta, and Alessandro Lenci. 2009. Evalita09 parsing task: comparing dependency parsers and treebanks. In *Proceedings of EVALITA 2009*.

Cristina Bosco, Simonetta Montemagni, and Maria Simi. 2013. Converting Italian treebanks: Towards an Italian Stanford Dependency Treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69, Sofia, Bulgaria, August. Association for Computational Linguistics.

Cristina Bosco, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The EVALITA 2014 dependency parsing task. In *Proceedings of EVALITA 2014*.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City, June. Association for Computational Linguistics.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar, October. Association for Computational Linguistics.

Jinho D. Choi and Andrew McCallum. 2013. Transition-based dependency parsing with selectional branching. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1052–1062, Sofia, Bulgaria, August. Association for Computational Linguistics.

Jinho D. Choi, Joel Tetreault, and Amanda Stent. 2015. It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 387–396, Beijing, China, July. Association for Computational Linguistics.

Walter Daelemans and Véronique Hoste. 2002. Evaluation of machine learning methods for natural language processing tasks. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain.

Walter Daelemans, Véronique Hoste, Fien De Meulder, and Bart Naudts. 2003. Combined optimization of feature selection and algorithm parameters in machine learning of language. In *Proceedings of the 14th European Conference on Machine Learning (ECML 2003)*, Cavtat-Dubronik, Croatia.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Manchester, UK, August. Coling 2008 Organizing Committee.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343, Beijing, China, July. Association for Computational Linguistics.

Véronique Hoste, Iris Hendrickx, Walter Daelemans, and Antal van den Bosch. 2002. Parameter optimization for machine-learning of word sense disambiguation. *Natural Language Engineering*, 8(4):311–325.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *CoRR*, abs/1603.04351.

Alberto Lavelli. 2014a. Comparing state-of-the-art dependency parsers for the EVALITA 2014 dependency parsing task. In *Proceedings of EVALITA 2014*.

Alberto Lavelli. 2014b. A preliminary comparison of state-of-the-art dependency parsers on the Italian Stanford Dependency Treebank. In *Proceedings of the first Italian Computational Linguistics Conference (CLiC-it 2014)*.

Tao Lei, Yu Xin, Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. 2014. Low-rank tensors for scoring dependency structures. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1381–1391, Baltimore, Maryland, June. Association for Computational Linguistics.

Andre Martins, Miguel Almeida, and Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 617–622, Sofia, Bulgaria, August. Association for Computational Linguistics.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 2216–2219.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague, Czech Republic, June. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, et al. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *CoRR*, abs/1503.06733.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, et al. 2013. Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 146–182, Seattle, Washington, USA, October. Association for Computational Linguistics.

Djamé Seddah, Sandra Kübler, and Reut Tsarfaty. 2014. Introducing the SPMRL 2014 shared task on parsing morphologically-rich languages. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 103–109, Dublin, Ireland, August. Dublin City University.

Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 188–193, Portland, Oregon, USA, June. Association for Computational Linguistics.

Yuan Zhang, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2014a. Greed is good if randomized: New inference for dependency parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1013–1024, Doha, Qatar, October. Association for Computational Linguistics.

Yuan Zhang, Tao Lei, Regina Barzilay, Tommi Jaakkola, and Amir Globerson. 2014b. Steps to excellence: Simple inference with refined scoring of dependency trees. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Baltimore, Maryland, June. Association for Computational Linguistics.

May the Goddess of Hope Help Us.

Homonymy in Latin Lexicon and Onomasticon

Marco Passarotti
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo Gemelli, 1 – 20123 Milan, Italy
marco.passarotti@unicatt.it

Marco Budassi
Università di Pavia
Corso Strada Nuova, 65
27100 Pavia, Italy
marcobudassi@hotmail.it

Abstract

English. We present a study on the degree of homonymy between the lexicon of a morphological analyser for Latin and an Onomasticon. To understand the impact of homonymy, we discuss an experiment on four Latin texts of different era and genre.

Italiano. *L'articolo presenta uno studio sul grado di omonimia tra il lessico di un analizzatore morfologico per il latino e un Onomasticon. Al fine di comprendere l'impatto dell'omonimia, viene descritto un esperimento condotto su quattro testi latini di diversa epoca e genere.*

1 Introduction

Ambiguity affects linguistic analysis at various levels. In particular, homonymy plays a substantial role in the analysis of single words. Indeed, when considered out of context, one same word can be assigned different Parts of Speech (PoS), morphological features, lemmas and meanings. Contextual disambiguation is the task of Natural Language Processing (NLP) tools like PoS-taggers, morphological analysers, lemmatisers and word-sense disambiguators.

The problem of ambiguity is particularly remarkable for NLP when Named Entity Recognition (NER) is concerned. In order to automatically classify the textual occurrences of (multi)words into categories such as names of persons, locations and organisations, NER faces that specific kind of ambiguity consisting in the homonymy between proper names and other words in the lexicon (Nadeau and Sekine, 2007). For instance, the word *mark* in English can be a proper name, a noun or a verb. Although such homonymy is often tackled by using the upper/lowercase distinction for the initial letter of words, this solution is neither decisive (as uppercase letters can also be motivated by punctuation) nor always available. The latter is especially true for historical languages, as a large amount of texts in such languages comes with no upper/lowercase distinction and it may follow different editorial criteria.

The recent extension of the lexical basis of the morphological analyser and lemmatiser for Latin Lemlat with an Onomasticon (i.e. a list of proper names) makes it possible to evaluate the degree of homonymy of proper names in Latin and, thus, to understand the extent of the disambiguation task (Passarotti and Ruffolo, 2004). To this aim, in this paper we explore the lexical basis of Lemlat as providing the empirical evidence supporting our analysis on homonymy between names in the Onomasticon and words in the Latin lexicon.

2 Lemlat

Together with *Morpheus* (Crane, 1991) and Whitaker's *Words*, Lemlat (Passarotti, 2004) is one of the most widespread tools for automatic analysis of Latin morphology available. The original lexical basis of Lemlat (L) results from the collation of three Latin dictionaries (Georges and Georges, 1913-1918; Glare, 1982; Gradenwitz, 1904). It counts 40,014 lexical entries and 43,432 lemmas (as more than one lemma can be included into the same lexical entry). Such lexical basis was recently merged with most of the Onomasticon (O) (26,250 lemmas out of 28,178) provided by the 5th edition of *Lexicon Totius Latinitatis* (Forcellini, 1940) (Budassi and Passarotti, 2016).

Since the large majority of lemmas in O are nouns (19,599 out of 26,250), we will focus on them here, first by comparing their distribution in L and O. Table 1 shows the number of nouns and their percentage (on the total of nouns) in L and O by inflectional category.

Infl. Cat.	Lemlat	Onomasticon
I decl.	5,009 (22.26%)	6,651 (33.94%)
II decl.	7,466 (33.17%)	7,235 (36.92%)
III decl.	8,677 (38.54%)	4,464 (22.77%)
IV decl.	980 (4.35%)	58 (0.29%)
V decl.	101 (0.45%)	6 (0.03%)
Uninfl.	278 (1.23%)	1,185 (6.05%)
TOTAL	22,511	19,599

Table 1. Nouns in L and O.

While third declension nouns are more frequent in L than in O, the opposite holds for first declension and (to a lesser extent) second declension nouns. The main difference between L and O concerns uninflected nouns, which are much more in O than in L because of the large number of loans recorded in O.

Also gender-based distribution of nouns by inflectional category shows substantial differences between L and O. Among the most relevant is that O includes more first declension masculine nouns than L (1,626 vs. 562). Instead, the number of second declension neuter nouns is larger in L than in O (4,005 vs. 1,523), because O tends to include more proper names of persons than of places, the latter being often assigned the neuter gender. As for third declension, feminine nouns are more than masculine in L (5,112 vs. 2,590), while the opposite holds in O (2,847 masculine vs. 1,185 feminine).

3 Mining Nominal Homonymy

To categorise nominal homonymy in L and O, we defined three kinds of homonymy: (a) Full Homonymy (FH): words with the same lemma, PoS, inflectional category and gender in L and O; (b) Partial Homonymy (PH): words with the same lemma in L and O, but with different PoS, inflectional category or gender (the last for nouns only); (c) Mixed Homonymy (MH): words with the same lemma in L and O and with more than one PoS, inflectional category or gender, thus resulting partly into FH and partly into PH.

An example of FH in our data is the word *spes*, which means "hope" in L and "the Goddess of Hope" in O. PH is represented, for instance, by the word *augustus*, which is an adjective in L ("majestic") and a noun in O (a cognomen given to Octavius Caesar as emperor). The word *spina* is a case of MH, being a first declension feminine noun in L ("thorn") and both a first declension feminine noun (an old town in Aemilia) and a third declension masculine noun with genitive in *–anis* (a river God) in O, the former thus showing FH and the latter PH.

Table 2 presents the rates of homonymy in L and O by each kind per inflectional category. The total number of homonyms is provided as well (column "H"). This corresponds to the number of nouns of an inflectional category that are graphically identical in L and O. For instance, the first row of table 2 shows that there are 556 lemmas recorded as first declension nouns (in L or O) that are identical to a lemma occurring in the other section of the lexical basis of Lemlat. 383 lemmas out of these show FH, i.e. they share not only the same graphical form but also the same PoS, inflectional category and gender in L and O (column "FH"). Instead, 163 lemmas occur as graphically identical in L and O but do not have in common at least one among PoS, inflectional category or gender (column PH"). Finally, 10 lemmas show MH.

Infl. Cat.	H	FH	PH	MH
I decl.	556	383	163	10
II decl.	752	307	389	56
III decl.	584	334	226	24
IV decl.	85	9	73	3
V decl.	6	5	1	0
Uninfl.	60	0	60	0
TOTAL	2,043	1,038	912	93

Table 2. Kinds of homonymy.

Most of the PH instances for first declension lemmas are due to different gender. An example is the first declension noun *caligula*, which is feminine in L ("a small military boot") while it is masculine in O (a cognomen). Second declension shows several cases of PoS change, like in the case of *severus*, which is an adjective in L ("serious") and a noun in O (a proper name). Instead, a large number of verb-noun changes holds for third declension. This mostly occurs for imparisyllable nouns ending in *–o*, like *cato*, which is a first conjugation verb in L ("to see") and a noun in O (a proper name).

PH does not raise any tricky issue for NLP, the task of PoS/morphological taggers being just that of disambiguating contextually PoS and morphological features. Conversely, FH (including the FH-like part of MH) represents a challenging question for NLP. Indeed, if upper/lowercase distinction is not available in input data, only context-based semantic properties can disambiguate between candidate lemmas affected by FH. For instance, in the clause "spes est expectatio boni" ("hope is expectation of good", Cicero, *Tusculanae*, 4, 37, 80) there is nothing but semantics to help us to understand that the word *spes* is an occurrence of the noun from L instead of the proper name from O. In order to evaluate the extent of homonymy in real texts and to understand how much big the impact of FH is, we performed the experiment discussed in the next section.

4 Homonymy in Texts. A Case-study

We run Lemlat on four Latin texts of similar size and different genre and era.[1] Table 3 shows the number of distinct words out of the total (column "Types") analysed by the original version of Lemlat (column "Lemlat") and by the one enhanced with the Onomasticon (column "LemlatON").

Text	Types	Lemlat	LemlatON	Improv.
(1)	3,092	2,888	3,039	+151
(2)	5,057	4,717	5,005	+288
(3)	3,542	3,357	3,487	+130
(4)	4,589	4,292	4,537	+245

Table 3. Results of Lemlat(ON) on four texts.

Beside the words analysed by LemlatOn only (column "Improv."), there is a certain degree of overlapping between Lemlat and LemlatOn. The words falling in this 'grey zone' are those that are analysed both by Lemlat and by LemlatOn, as they are lemmatised both under a lemma from L and under one from O. Among these words, those affected by homonymy are to be found.

[1] (1) Caesar, De Bello Gallico, 1 (Classical Lat., prose); (2) Virgil, Aeneid, 1 & 2 (Classical Lat., poetry); (3) Tertullian, Apologeticum (Late Lat., prose); (4) Claudian, De Raptu Proserpinae (Late Lat., poetry). All the texts were downloaded from the Perseus Digital Library (www.perseus.tufts.edu).

Text	L/O	H	FH	PH	MH
(1)	618	405	303	88	14
(2)	1,207	799	546	186	67
(3)	686	486	330	120	36
(4)	1,062	706	469	177	60

Table 4. Overlapping and homonymy rates.

Column "L/O" in table 4 reports the number of words for each text that are analysed both by Lemlat and by LemlatON. The other columns show the homonymy rates by the kinds described in Section 3. For instance, in the text of Caesar (1) there are 618 words analysed by both the versions of Lemlat (L/O). 405 out of them share the same lemma in at least one analysis (H). This is further detailed: 303 out of 405 show FH, 88 PH and 14 MH. An example of a word analysed by both the versions of the tool that does not share the same lemma in all analyses is *acie*, which is lemmatised under *acies* ("dagger") by Lemlat (fifth declension feminine noun) and also under the proper name *acius* by LemlatON (second declension masculine noun). The word *constantia* is an example of H: it is lemmatised as a form of both lemmas *consto* ("to agree"; first conjugation verb) and *constantia* ("steadiness"; second declension feminine noun) by Lemlat, and also as a form of both proper names *constantius* (second declension masculine noun) and *constantia* (second declension feminine noun) by LemlatON. The word *constantia* is also an example of FH, as the analyses provided by the two versions of Lemlat that share the same lemma have in common even the same inflectional category and gender. PH is shown by the word *crassi*, which is assigned the same lemma (*crassus*) both by Lemlat and by LemlatON, but while it is a first class adjective in the former ("solid"), it is a second declension masculine noun in the latter (a proper name). An example of MH is the word *amico*, which is lemmatised under the lemma *amicus* ("friend") both by Lemlat and LemlatON. The lemma *amicus* is both an adjective and a second declension masculine noun in Lemlat, but only the latter analysis is shared with LemlatON, because the lemma *amicus* in the Onomasticon is recorded only as a noun and not also as an adjective. Thus, when the word *amico* is assigned PoS noun it shows FH, while when it is assigned PoS adjective it shows PH.

The proportions between the kinds of homonymy remain quite similar for all the texts.

Words affected by H tend to be more than half of L/O; among them the large majority is affected by FH. By comparing columns "FH" and "MH" in table 4 with column "Types" in table 3, one can see that slightly more than 10% of the words of all the texts is affected by FH. This is the percentage rate of words whose lemmatisation cannot be disambiguated by a PoS tagger, because semantic features only are here at work to choose between candidate lemmas. For instance, if a PoS tagger assigns to one occurrence of the word *constantia* PoS noun and gender feminine, it cannot disambiguate between the two (fully morphologically identical) lemmas *constantia* provided by LemlatON.

If we focus on textual occurrences (tokens) instead of distinct words (types), the rates of FH (+MH) range between 8.44% (Caesar) and 13.19% (Tertullian), as shown by table 5. This result represents the extent of the impact of FH on the texts that we used in the case-study.

Text	Tokens	FH+MH
(1)	8,171	690 (8.44%)
(2)	10,045	1,325 (13.19%)
(3)	7,317	668 (9.13%)
(4)	6,991	797 (11.4%)

Table 5. Token-based homonymy rates.

Most of the words showing FH can be easily disambiguated (at least, manually) according to peculiarities of single texts. For instance, the word *amicitiam* (from Caesar's text) belongs to lemma *amicitia* both in L ("friendship") and in O ("the Goddess of Friendship"), thus showing FH. However, it is more likely that the former is the one occurring in Caesar than the latter. Conversely, in the same text the word *galli* (lemma *gallus*) is more likely a proper name from O ("Gauls") than a noun from L ("cock").

5 Conclusion

We presented a study about the degree of homonymy between the lexical basis of a morphological analyser for Latin and an Onomasticon recently added in the tool. We have shown the impact of nominal homonymy on a number of Latin texts of different era and genre.

Since the analysis of many homonymous words can be disambiguated according to the features of single texts (and authors), in the near future we foresee to enhance such words in Lemlat with information about their distribution in a number of manually tagged reference texts.

References

Marco Budassi and Marco Passarotti. 2016. Nomen Omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon. *Proceedings of the 10th Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities (LaTeCH 2016)*. The Association for Computational Linguistics, Berlin, Germany, 90–94.

Gregory Crane. 1991. Generating and Parsing Classical Greek. *Literary and Linguistic Computing*, 6(4):243–245.

Egidio Forcellini. 1940. *Lexicon Totius Latinitatis / ad Aeg. Forcellini lucubratum, dein a Jos. Furlanetto emendatum et auctum; nunc demum Fr. Corradini et Jos. Perin curantibus emendatius et auctius meloremque in formam redactum adjecto altera quasi parte Onomastico totius latinitatis opera et studio ejusdem Jos. Perin*. Typis Seminarii, Padova.

Karl Ernst Georges and Heinrich Georges. 1913-1918. *Ausführliches Lateinisch-Deutsches Handwörterbuch*. Hahn, Hannover.

Peter G.W. Glare. 1982. *Oxford Latin Dictionary*. Oxford University Press, Oxford.

Otto Gradenwitz. 1904. *Laterculi Vocum Latinarum*. Hirzel, Leipzig.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Marco Passarotti. 2004. Development and perspectives of the Latin morphological analyser LEMLAT. *Linguistica Computazionale*, XX-XXI:397–414.

Marco Passarotti and Paolo Ruffolo. 2004. L'utilizzo del lemmatizzatore LEMLAT per una sistematizzazione dell'omografia in latino. *Euphrosyne*, XXXII:99–110.

Author Index

Abad, Azad, 76
Abel, Andrea, 110
Alfieri, Linda, 138
Alfter, David, 214
Alicante, Anita, 99, 278
Arcara, Giorgio, 247

Baiamonte, Daniela, 116
Barlacchi, Gianni, 76
Basile, Pierpaolo, 209, 257, 273
Basile, Valerio, 273
Basili, Roberto, 58, 267
Bellandi, Andrea, 225
Benjamin, Martin, 47
Benotto, Giulia, 204, 225
Bernardi, Raffaella, 70
Bizzoni, Yuri, 214
Bogers, Toine, 52
Bompolas, Stavros, 235
Bondielli, Alessandro, 177
Bordea, Georgeta, 52
Bosco, Cristina, 104
Bottini, Roberto, 262
Bracchi, Alice, 143
Brunato, Dominique, 18
Budassi, Marco, 310
Buitelaar, Paul, 52

Cabrio, Elena, 273
Caputo, Annalina, 209
Cardillo, Franco Alberto, 235, 241
Caruso, Valeria, 149
Casasanto, Daniel, 262
Caselli, Tommaso, 116, 143
Chatterjee, Rajen, 35
Cherchi, Manuela, 41
Chiusaroli, Francesca, 47
Corazza, Anna, 1, 64, 99, 187, 278
Corino, Elisa, 154
Crepaldi, Davide, 262
Croce, Danilo, 58, 267
Culy, Chris, 127
Curci, Antonietta, 257
Cutugno, Francesco, 64

De Martino, Maria, 252
De Meo, Anna, 149
Del Tredici, Marco, 192
Dell'Orletta, Felice, 18, 29, 219
Desantis, Anna, 41
Di Nunzio, Giorgio Maria, 81

Esposito, Fabrizio, 64

Fantini, Anna, 160
Feltracco, Anna, 122
Ferro, Marcello, 235, 241
Ferro, Nicola, 52
Filice, Simone, 58
Franzon, Francesca, 247
Frey, Jennifer-Carmen, 284

Gagné, Christina L., 241
Gebremelak, Gebremedhen, 35
Giovannetti, Emiliano, 204, 225
Glaznieks, Aivars, 110, 284
Gregori, Lorenzo, 93
Guglielmi, Francesca, 257

Herbelot, Aurelie, 70
Hernández Farías, Delia Irazú, 104

Iovino, Rossella, 219
Isgrò, Francesco, 99

Jezek, Elisabetta, 23, 122

Lai, Mirko, 104
Lapata, Mirella, 11
Laudanna, Alessandro, 252
Lavelli, Alberto, 304
Lebani, Gianluca E., 12
Lenci, Alessandro, 12, 177, 230
Lieto, Antonio, 166
Litta, Eleonora, 127
Logozzo, Felicia, 289
Luisi, Roberta, 209

Maggio, Valerio, 187
Magnini, Bernardo, 23, 87, 122
Maistro, Maria, 81
Mancuso, Azzurra, 252
Mansour, Sina, 47
Marchi, Simone, 204
Marzi, Claudia, 235
Mazzei, Alessandro, 172

Mencarini, Letizia, 104
Mensa, Enrico, 166
Minard, Anne-Lyse, 87
Mitkov, Ruslan, 41
Montemagni, Simonetta, 1, 29
Monti, Johanna, 41, 47
Moretti, Giovanni, 198
Moschitti, Alessandro, 76, 299
Mozzachiodi, Michele, 104

Nadalini, Andrea, 262
Nardi, Daniele, 267
Negri, Matteo, 35
Nicolas, Lionel, 110
Nissim, Malvina, 192
Nivre, Joakim, 10

Orletti, Franca, 219

Palmero Aprosio, Alessio, 132
Panunzi, Alessandro, 93
Passaro, Lucia C., 177
Passarotti, Marco, 127, 310
Patti, Viviana, 104
Pezzelle, Sandro, 70
Piccini, Silvia, 225
Pieri, Giulia, 18
Pironti, Antonio, 278
Pirrelli, Vito, 235, 241
Pisano, Simone, 182
Ponti, Edoardo Maria, 23
Prodanof, Irina, 116, 143

Qwaider, Mohammed R. H., 87

Radicioni, Daniele P., 166
Ravelli, Andrea Amelio, 93
Rodda, Martina A., 230
Rossinelli, Emanuele, 76
Russo, Claudio, 154
Russo, Irene, 182

Saltori, Francesca, 132
Sangati, Federico, 47
Scanniello, Giuseppe, 187
Semeraro, Giovanni, 1, 209, 257
Senaldi, Marco S. G., 12, 230
Silvello, Gianmaria, 52
Silvestri, Stefano, 99
Soria, Claudia, 182
Sorodoc, Ionut, 70
Spalding, Thomas L., 241
Speranza, Manuela, 87
Sprugnoli, Rachele, 198
Stede, Manfred, 122
Stemle, Egon W., 110, 284
Sulis, Emilio, 104

Tamburini, Fabio, 138, 294
Taslimipoor, Shiva, 41
Tonelli, Sara, 132, 198
Turchi, Marco, 35
Tusa, Erica, 29

Uva, Antonio, 299

Vanzo, Andrea, 267
Venturi, Giulia, 29
Vignoli, Daniele, 104
Villata, Serena, 273
Vitale, Vincenzo Norman, 149

Zaninello, Andrea, 192
Zanini, Chiara, 247
Zilio, Daniel, 81