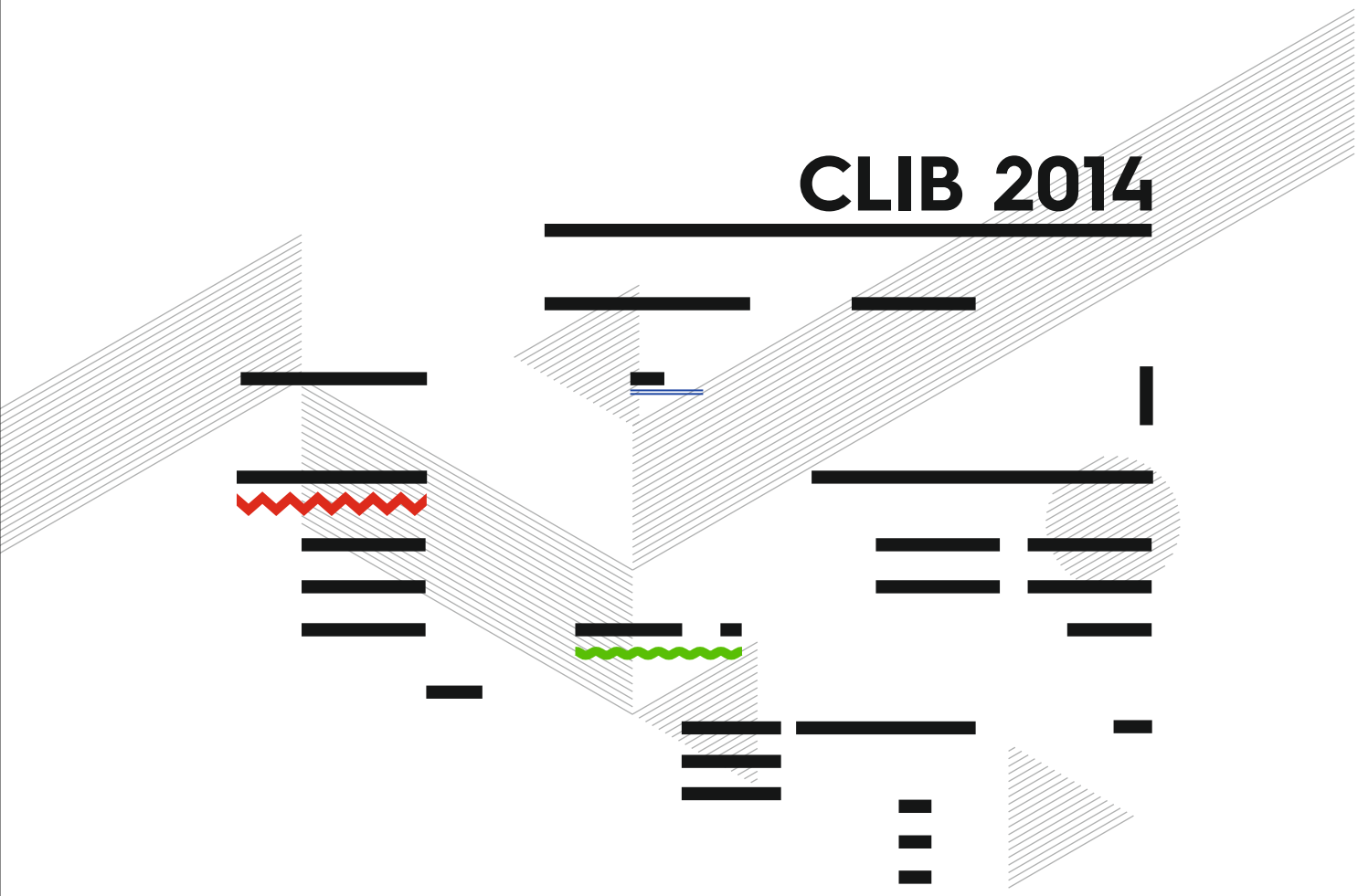
Computational Linguistics in Bulgaria

'14

CLIB 2014



4 September, 2014
Sofia, Bulgaria

CLIB 2014 is organised within the project
Integrating New Practices and Knowledge in Undergraduate and Graduate Courses in Computational Linguistics
Grant № BG051PO001-3.3.06-0022.

European Union

European Social Fund

The project is implemented with the financial support of the *Human Resources Development Operational Programme* 2007-2013 co-financed by the European Social Fund of the European Union.

CLIB 2014 is organised by:

The Department of Computational Linguistics,
Institute for Bulgarian Language,
Bulgarian Academy of Sciences

The Faculty of Slavic Studies,

The Faculty of Mathematics and Informatics,
Sofia University

Publication and cataloguing information

Title: **Proceedings of the First International Conference *Computational Linguistics in Bulgaria***

ISSN: **2367-5578**

Edited by: **Svetlozara Leseva, Tsvetana Dimitrova, Ivelina Stoyanova, Ekaterina Tarpomanova, Rositsa Dekova**

Published and distributed by: **The Institute for Bulgarian Language *Prof. Lyubomir Andreychin* – Bulgarian Academy of Sciences**

Editorial address: **Institute for Bulgarian Language Prof. Lyubomir Andreychin, Bulgarian Academy of Sciences**

52 Shipchenski prohod blvd, bldg. 17
Sofia 1113, Bulgaria
+359 2/ 872 23 02

Proceedings of the First International Conference

Computational Linguistics in Bulgaria

CLIB 2014

4 September 2014
Sofia, Bulgaria

PREFACE

We are excited to welcome you to the inaugural edition of the international conference *Computational Linguistics in Bulgaria* (CLIB 2014) in Sofia, Bulgaria!

CLIB is a joint effort launched by the Department of Computational Linguistics (DCL) at the Institute for Bulgarian Language of the Bulgarian Academy of Sciences together with the Faculty of Slavic Studies and the Faculty of Mathematics and Informatics at Sofia University.

CLIB aspires to foster the NLP community in Bulgaria and further the cooperation among researchers working in NLP for Bulgarian around the world. The need for a conference dedicated to NLP research dealing with or applicable to Bulgarian has been felt for quite some time. We believe that building a strong community of researchers and teams who have chosen to work on Bulgarian is a key factor to meeting the challenges and requirements posed to computational linguistics and NLP in Bulgaria. We share the hope that CLIB will establish itself as an international forum for sharing high-quality scientific work in all areas of computational linguistics and NLP and will grow in scope and scale with each new edition. The CLIB community will be dedicated to supporting the creation and improvement of advanced NLP resources, tools and technologies for mono- and multilingual language processing, machine translation and translation aids, content creation, localisation and personalisation, speech recognition and generation, information retrieval and information extraction.

The Conference was made possible due to the hard work of many people. We would like to thank the authors who trusted us and submitted their contributions to CLIB 2014. Their efforts and high-quality research are the chief factor that enabled us to create an interesting and solid scientific programme. We would also like to thank our industrial participants for sharing their insights, ideas and know-how with the research community. Let us also express our sincere gratitude to the members of the Programme Committee, who accepted to join us and invested a lot of expertise to provide valuable feedback to the authors. Special thanks are due to Prof. Svetla Kœva, who is the person behind the whole CLIB concept.

We hope that CLIB 2014 will be a useful and productive experience that we all will enjoy!

CLIB Organising Committee

PROGRAMMING COMMITTEE

Galia Angelova – Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Dan Cristea – Alexandru Ioan Cuza University of Iaşi
Radovan Garabík – Ľudovít Štúr Institute of Linguistics
Mariana Damova – Mozaika Ltd., Bulgaria
Ivan Derzhanski – Institute of Mathematics and Computer Science, Bulgarian Academy of Sciences
Kjetil Rå Hauge – University of Oslo
Verginica Barbu Mititelu – Research Institute for Artificial Intelligence, Romanian Academy
Stoyan Mihov – Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Preslav Nakov – Qatar Computing Research Institute
Svetla Koeva – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Cvetana Krstev – University of Belgrade
Éric Laporte – University of Paris-Est Marne-la-Vallée
Denis Maurel – François-Rabelais University of Tours
Maciej Ogrodniczuk – Institute of Computer Science, Polish Academy of Sciences
Karel Oliva – Institute of the Czech Language, Academy of Sciences of the Czech Republic
Maciej Piasecki – Wrocław University of Technology
Agata Savary – François-Rabelais University of Tours
Jan Šnajder – University of Zagreb
Ivelina Stoyanova – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Marko Tadić – University of Zagreb
Hristo Tanev – Joint Research Centre of the European Commission in Ispra, Italy
Tinko Tinchev – Sofia University St. Kliment Ohridski
Dan Tufiş – Research Institute for Artificial Intelligence, Romanian Academy
Duško Vitas – University of Belgrade
Radka Vlahova – Sofia University St. Kliment Ohridski

ORGANISING COMMITTEE

Rositsa Dekova – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Tsvetana Dimitrova – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Svetlozara Leseva – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Vladislav Nenchev – Sofia University St. Kliment Ohridski
Bilyana Radeva – Sofia University St. Kliment Ohridski
Andrey Sariev – Sofia University St. Kliment Ohridski
Ekaterina Tarpomanova – Sofia University St. Kliment Ohridski
Stefan Vatev – Sofia University St. Kliment Ohridski

INVITED SPEAKERS

Prof. Cvetana Krstev. *Developing Resources for the Culinary Domain.*
Prof. Eric Laporte. *Interaction between Linguists and Machine Learning.*

Table of Contents

Electronic Language Resources in Teaching Mathematical Linguistics

Ivan Derzhanski
Department for Mathematical Linguistics
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
iad58g@gmail.com

Rositsa Dekova
Department for Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences
rosdek@dcl.bas.bg

Abstract

The central role of electronic language resources in education is widely recognised (cf. Brinkley et al, 1999; Bennett, 2010; Derzhanski et al., 2007, among others). The variety and ease of access of such resources predetermines their extensive use in both research and education. With regard to teaching mathematical linguistics, electronic dictionaries and annotated corpora play a particularly important part, being an essential source of information for composing linguistic problems and presenting linguistic knowledge.

This paper discusses the need for electronic resources, especially for less studied or low-resource languages, their creation and various uses in teaching linguistics to secondary school students, with examples mostly drawn from our practical work.

1. Introduction

The mid-1960s saw the birth of the idea of presenting contemporary linguistics to secondary school students through a variety of entertaining extracurricular activities. The most prominent of those activities is the Linguistics Olympiad – contest in solving self-sufficient linguistic problems. Such problems present interesting linguistic phenomena in an enigmatic form and invite their discovery. The phenomena are presumed to be unfamiliar to the solver and may be facts of one or several languages or of language in general, or they may be ideas or concepts of linguistic science. Self-sufficiency means that a linguistic problem must be solvable using only logical thought and the information it contains, possibly supplemented by general knowledge and such concepts of linguistics, mathematics, etc., that are part of the regular school curriculum.

The First Linguistics Olympiad for secondary school students was held in 1965 in Moscow, which was the only venue of such events for 17 years. Then the linguistics competitions were launched in Bulgaria, mostly through the efforts of mathematicians, as accompanying events to contests in mathematics (Derzhanski, 2007). For these and other organisational reasons, and also because in the early years most problems that were composed in Bulgaria were on topics from mathematical or computational linguistics, linguistics as a subject of extracurricular activities for secondary school students is called 'mathematical linguistics' in Bulgaria to this day, even though the focus of the contests has shifted away from mathematical linguistics as a field of research and towards descriptive linguistics and typology. (This imprecision is tolerable, especially since, whatever topics the problems feature, the main asset for their solving is analytical thinking, which is generally associated with mathematics.)

Following similar efforts in the Netherlands and USA, in 2003 the International Olympiad in Linguistics (IOL) was launched, and has grown from 33 contestants from participating 6 countries at the first instalment to 152 contestants from 28 countries at the 12th (in 2014) and stimulated the setting up of numerous new regional and national olympiads and competitions in linguistics for secondary school students.

Thus all these countries introduced teaching contemporary linguistics (a field of study that tends to be absent from regular curricula) to secondary school students, on a narrower or broader scale, in the form of theory and practice of solving self-sufficient problems covering a wide variety of linguistic phenomena. When we refer to teaching linguistics (or mathematical linguistics) in schools in this paper, we have in mind mainly (though not exclusively) training in solving linguistic problems.

Naturally, '[a] steady supply of original, thoughtfully created and intriguing problems is absolutely necessary for the success of any ongoing [linguistic olympiad] programme' (Derzhanski and Payne, 2010). The efficiency of the problem composition and problem verification process is therefore critical. And it depends directly on the kind, size and quality of the resources available to authors and editors, especially dictionaries and corpora.

2. Language Resources for Creating Linguistic Problems

The variety and flexibility of the language resources for creating linguistic problems has to match the variety of problems, which is immense. There are monolingual problems, often on the solver's native language, focusing on little-known linguistic phenomena within the fields of grammar, semantics, or pragmatics; bilingual problems treating correspondences (regular but usually non-trivial ones) between two linguistic systems, which may be the solver's native tongue and an unfamiliar language, or the sound of a language and its written representation, or two cognate languages or dialects; and even multilingual problems, in which several such systems are compared. All levels of the language code can be involved—orthography, phonology, morphology, syntax, semantics, and discourse structure.

2.1. Use of Electronic Dictionaries

Electronic dictionaries (e-dictionaries), both monolingual and bilingual, are available now for many languages. With respect to their type and functionality, however, e-dictionaries vary widely — from a simple digital image of a printed dictionary to a digital dictionary which includes additional information (such as pronunciation or spelling in an alternative orthography, noun declension and verb conjugation, stemming and/or lemmatisation, links to derived words, sense-linked thesaurus, etc.), allows browsing, and features a powerful search engine. It is namely the latter type which serves best in composing linguistic problems, an activity in which advanced search using wildcards and/or regular expressions is especially useful.

A problem on morphology, for instance, typically illustrates some interesting rule of derivation or inflexion that makes the construction of a word or form depend on the phonological shape of the stem, the word class or some other category in a non-obvious way. To compose such a problem, one needs a significant amount of candidate data and test examples, and such can be found easily in a dictionary with adequate search tools. For example, a sizable class of Estonian nouns have single-vowel partitive plural endings, which correlate with the partitive singular ending and the stem-internal vowel. This phenomenon was demonstrated by a problem which was created using several resources: an electronic dictionary (an Estonian-Russian one) that allowed wildcard search for headwords but offered no grammatical information, the online tool Estonian Language Synthesiser[1] to verify whether the candidate words formed their partitive singular and plural forms in the required way, and a paper dictionary to resolve homonymy, which the Synthesiser doesn't do. A digital dictionary with the respective partitive singular and plural forms for every noun and an option to search for them would have made the task far easier.

Another reason to look for words of a certain morphological type may be to reduce morphological variety in a problem whose weight lies elsewhere, usually in syntax. For a problem which featured switch reference marking in Alabama the author needed to choose several verbs that would take the same set of subject and (if transitive) object person/number markers, so that the diversity of conjugation types, which is very large in this language, wouldn't obscure the main syntactic phenomenon. The verbs were collected by regular expression search in the text of an electronic edition of a paper dictionary (Sylestine et al., 1993), taking advantage of the fact that in the entries the headword was followed by grammatical

[1]Available at `http://www.filosoft.ee/gene_et/`.

information. In such cases, too, a more sophisticated structure of the dictionary can make the search significantly more efficient.

2.2. Use of Electronic Corpora

Besides dictionaries, a problem composer can use corpora as well, as tools for studying linguistic structure and as sources of naturally occurring examples of language use. Some problems are constructed entirely using material from a corpus. This is particularly desirable when the language is extinct (New Testament Greek, Middle Dutch, Tocharian, etc.) or the phenomenon calls for authentic material, as when composing problems on the structure of classical poetic forms or on word usage that occurs chiefly in literature, such as the sailors' manner of time-telling exemplified by the phrase *from about noon observation to about six bells* (Robert Louis Stevenson, *Treasure Island*). Or it may be the author's choice, aimed at making the problem more interesting. For example, a problem which presents a number of sentences in the working language which all contain the sole pronoun *we* and states that if the sentences were translated into (say) Tok Pisin, different pronouns would be used for reasons which the solver must discover, may be made more attractive if the sentences were taken from novels that the solver may know of (note that in this case it doesn't matter if the books exist in Tok Pisin at all). The quality of a corpus-based problem depends directly on the size, structure and search facilities of the corpus.

Most contemporary electronic corpora are annotated at various levels. Part-of-speech tagging is nearly ubiquitous; morphosyntactic annotation and lemmatisation is included with increasing frequency, and some corpora provide semantic and/or syntactic annotation. Furthermore, most electronic corpora are also equipped with a web search interface that allows searches for exact words or phrases, regular expressions, part of speech information, lemma, collocations, frequency and distribution of synonyms, syntactic and semantic features. These functionalities of annotated corpora and the diversity of possible queries play an essential part in contemporary problem making for the purposes of teaching mathematical linguistics.

The existence and the availability of national corpora for closely related languages, corpora of dialects or historical corpora is a useful asset for finding data for problems on phonology or morphology which draw on theoretical aspects from diachronic and comparative linguistics. Such is, for instance, a problem consisting of sentences in a regional dialect of South Bulgaria and their counterparts in contemporary standard Bulgarian where specific words are omitted so that solvers can discover a linguistic phenomenon which is present in the dialect but not in the standard (namely a distinction of proximity in demonstrative and relative pronouns and the definite article).

The availability of parallel and aligned corpora also greatly facilitates the finding of applicable excerpts of texts, as well as in the search for proper sample sentences of cognate words in unrelated languages. For example, a problem may focus on the change of meaning of cognates which could be reconstructed by students given suitable examples of natural language sentences; or students may be provided with a carefully selected coherent text and its translation and asked to discover grammar rules (a process which resembles a lot human-aided machine learning).

Problems may also comprise a set of words from two or more dialects (or closely related languages) focusing on a specific sound shift (e.g., Grimm's Law, Ruki sound law, palatalisation).

3. Task-driven Compilation of Electronic Resources

Both electronic dictionaries and corpora are often hard to come by, especially when working with exotic (or other low-resource) languages, but sometimes this difficulty can be circumvented. On one occasion, when creating a problem on Maori syntax, the author wished to have a corpus of Maori sentences in order to choose several syntactic constructions for inclusion into the problem. Since no such resource was available, a small working corpus was composed from examples given in an English–Maori dictionary (Ngata, 1993) and used successfully. Again, a large ready-made corpus with adequate search tools would have sped up the task.

Of course, not even the most sophisticated electronic dictionary or corpus can foresee all kinds of search that a user may need to perform, and the needs of authors of linguistic problems are among the most unforeseeable. It is unlikely, for example, that a dictionary will help to find anagrams, palindromic

headwords, or words which are cognate in the source and (related) target language. In such cases the problem composer (with a heart for programming) will want to download the dictionary and write his own programs to process it. Even a plain computer-readable word list is preferable to no resource at all.

4. Electronic Resources in Use by Teachers and Students

Being large and principled collections of naturally occurring language samples, corpora are used not only for composing and testing linguistic problems, but also for extracting examples to illustrate various linguistic phenomena in classroom teaching of mathematical linguistics.

When presented with a problem outside contest situations, students are usually left alone to solve the problem and thus to discover some underlying theoretical facts. Then the teacher's job is to deliver additional information on the newly discovered linguistic phenomenon and to supply examples for clarification. This is where electronic dictionaries and corpora play an essential part and help teachers provide the necessary linguistic data.

Electronic resources may be so used by students in their independent work as well. It is a recent policy of the Bulgarian Olympiad in Mathematical Linguistics that leading participants are advised to write a short research paper on a language phenomenon of their choice and to compose a sample linguistics problem (a good performance in this increases their chances to get on the national team for the International Linguistics Olympiad). And in this task students are strongly encouraged to use examples from corpora when providing linguistic evidence. Whilst originality is not expected at this stage, it is expected that the students can benefit from a small-scale first-hand encounter with linguistic research, including all stages of work with language resources (locating the resources themselves, finding the necessary information, formatting and citation). The higher accessibility of the Net, as compared to a traditional research library, means that electronic resources available online are especially well suited for this.

5. Conclusions and Future Work

In light of the rapid growth of the International Linguistics Olympiad (39 teams from 28 countries as of Edition 2014) and its national tributaries, the teaching society faces an increasing need of electronic language resources, especially on exotic and other low-resource languages, which allow for browsing and advanced searches. Although some small-size resources may be compiled in situ for a given task, the existence and the availability of large and searchable dictionaries and corpora is becoming an invaluable resource in teaching mathematical linguistics.

In the future it will be useful to establish a database with a list of available resources, as well as provide wider online access to resources created for specific teaching purposes.

Acknowledgements

The present paper was prepared within the project *Integrating New Practices and Knowledge in Undergraduate and Graduate Courses in Computational Linguistics* (BG051PO001-3.3.06-0022) implemented with the financial support of the Human Resources Development Operational Programme 2007 - 2013 co-financed by the European Social Fund of the European Union. The authors take full responsibility for the content of the present paper.

References

Bennett, G.R. (2010). *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers.* University of Michigan Press/ELT, 144 p.

Brinkley, A., Dessants, B., Flamm, M., Fleming, C., Forcey, C. and Rothschild, E. (1999). *The Chicago Handbook for Teachers: A Practical Guide to the College Classroom, University of Chicago Press,* pages 143–67. `http://www.press.uchicago.edu/ucp/books/book/chicago/C/bo3633179.html`

Derzhanski, I.A. (2007). Extracurricular Activities in Linguistics for Secondary School Students. In Dimitrova, L. and Pavlov, R. (eds.), Mathematical and Computational Linguistics. Jubilee International Conference, 6 July 2007, Sofia, pages 125–128. https://www.academia.edu/5680539/Extracurricular_activities_in_linguistics_for_secondary_school_students

Derzhanski, I.A. and Payne, T. (2010). The Linguistics Olympiads: Academic Competitions in Linguistics for Secondary School Students. In Denham, K. and Lobeck, A. (eds.), *Linguistics at School: Language Awareness in Primary and Secondary Education*, Cambridge University Press, pages 213–226. https://www.academia.edu/5680870/The_Linguistics_Olym piads-Academic_competitions_in_linguistics_for_secondary_school _students

Derzhanski, I.A., Dimitrova, L. and Sendova, E. (2007). Electronic Lexicography and Its Applications: The Bulgarian Experience. In Широков, В.А. (відп. ред.), *Прикладна лінгвістика та лінгвістичні технології. Megaling-2006: Збірник наукових праць*, Київ: «Довіра», стр. 111–118.

Ngata, H.M. (1993). *English–Maori Dictionary*. Learning Media Ltd. Online version available: http://www.learningmedia.co.nz/ngata

Sylestine, C., Hardy, H.K., and Montler, T. (1993). *Dictionary of the Alabama Language.* Austin: University of Texas Press. Online version available: http://www.ling.unt.edu/~montler/Alabama/Dictionary/

Harnessing Language Technologies in Multilingual Information Channeling Services

Diman Karagiozov
Tetracom IS Ltd.
diman@tetracom.com

Abstract

Scientists and industry have put significant efforts in creating suitable tools to analyze information flows. However, up to now there are no successful solutions for 1) dynamic modeling of the user-defined interests and further personalization of the results, 2) effective cross-language information retrieval, and 3) processing of multilingual content. As a consequence, much of the potentially relevant and otherwise accessible data from the media stream may elude users' grasp.

We present a multilingual information channeling system, MediaTalk, which offers broad integration between language technologies and advanced data processing algorithms for annotation, analysis and classification of multilingual content. As a result, the system not only provides an all-in-one monitoring service that covers both traditional and social media, but also offers dynamic modeling of user profiles, personalization of obtained data and cross-language information retrieval. Bulgarian and English press clipping services relying on this system implement advanced functionalities such as identification of emerging topics, forecasting and trend prediction, all of which allow the users to monitor their standing reputation, events and relations. The architecture of the system is robust, extensible and adheres to the Big Data paradigm.

1. Introduction

In the last decade, the information available on the Internet has grown significantly[1] and has increased the demand for efficient monitoring and information extraction for the purposes of industry and research, including publishing, marketing, advertising, social research, etc. The problem is related not only to the vast volume of information but also to the dynamic nature of the information flow, the variety of sources, data formats, media types (printed, electronic, audio, multimedia) and languages. Monitoring applications for this purpose have complex structures implementing advanced data processing and language technologies.

Moreover, the processing and information extraction from multilingual and multimodal content is still an area of active research with no established solutions. The efficiency of methods is also essential since they need to have close to real-time performance and give high-quality results.

Section 2 describes existing media monitoring services which provide similar functionalities to the system presented in this paper. The key design and functional decisions, implementation and integration approaches are described in Section 3. Section 4 outlines the functionalities of an intelligent web application built with our system and the benefits of using it. Section 5 summarizes the main achievements of the system and suggests improvements and extensions.

[1] Digital Universe studies series, John F. Gantz et al., IDC 2007, 2008, 2009, 2010
http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm

2. Overview of Related Service

Media monitoring services have gained popularity in recent years and have focused on monolingual electronic content: web pages of broadcasting agencies, TV and radio channels, newspapers, governmental and non-governmental organizations. Some of the most popular companies offering these services are WebClipping[2], PressClipping[3], eReleases[4]. Typically, monitoring applications process monolingual information in structured form and employ methods based on keywords, text categorization, named entity recognition and monolingual information extraction.

Another branch of information services reside in the Google ecosystem. Public services like Google Reader[5] and Google Alerts allow the users to describe their fields of interests by providing search queries. Users are then notified of recent query results for the same terms via RSS feeds and emails.

The service provided by Prismatic[6] integrates with the social profiles of the users and tries to present a user-focused English language information stream. Companies such as CyberAlert[7] and CustomScoop[8] offer search queries in multiple languages and instant machine translation of extracted clippings along with comprehensive news coverage and minimization of irrelevant information. The service Mention.net[9] is able to process multilingual content; the user profile information is described as a set of keywords.

There are specialized broadcast monitoring services, such as Critical Mention[10] and TV Eyes[11], that combine real-time TV and radio broadcast monitoring with online and social media coverage. Such services capture text, audio and video content, analyze it to some extent and distribute it.

To the best of our knowledge, there is no system applying cross-language information retrieval in news monitoring and no existing unified approach with respect to dynamic user profiling and multilingual and multimodal content monitoring.

3. A Multilingual Information Channeling System

We present a system for multilingual information channeling which aims for the following key objectives: (a) relevant news coverage and adequate data analysis (implemented); (b) efficient dynamic modeling of user profiles (implemented); (c) a uniform approach to user profiling and multilingual content monitoring (implemented); (d) efficient cross-language information extraction (under development); and (e) a uniform approach to processing multimodal content (a future task).

To achieve these objectives, we integrate the appropriate language technologies with advanced data processing algorithms for annotation, analysis and classification of multilingual content. More particularly, semantic entities[12] are extracted, represented as time series and classified in order to obtain relevant news coverage and to provide adequate data analysis. Relations between semantic entities are represented as a semantic graph that enables users to track the related persons, dates, locations and the like. Data (both target information flow and user information) are provided with internal semantic links that preserve content integrity and allow information tracking. Our experiment towards crossing the language barrier utilizes a hybrid (example-based, statistical and dictionary-based) machine translation

[2] `http://www.webclipping.com`

[3] `http://www.pressclip.net`

[4] `http://www.ereleases.com`

[5] Google Reader service has been discontinued since 1st July 2013. Google Alerts service provides the relevant to the query information in email format which is not convenient for integration in 3rd party software systems.

[6] `http://getprismatic.com/`

[7] `http://www.cyberalert.com`

[8] `http://www.customscoop.com`

[9] `https://en.mention.net/`

[10] `http://www.criticalmention.com`

[11] `http://tveyes.com`

[12] We will refer to both named entities and noun phrases as semantic entities further on in this paper.

engine and a language-independent presentation of the semantic entities. The processing of multimodal content will be implemented as extensions to existing text and metadata extraction systems, such as Apache TIKA[13].

3.1. System Workflow

First, we present the processes of data acquisition, normalization, processing, indexing, analysis and visualization. After that we describe the novel approach, called *Lambda architecture*, adopted in solving the inevitable Big Data problem (Marz and Warren, 2013).

The general workflow in the system can be generalized in the following subsequent processes:

1. Harvesting the data – based on a wide collection of RSS feeds and user defined queries and resources from the social networks, a pool of information items is created.
2. Text extraction – each information item is transformed to a list of textual fragments.
3. Semantic excerpts – each textual fragment is processed so that the most "interesting" semantic excerpts are indexed.
4. Graph of semantically related excerpts – a graph of interrelated information items and semantic excerpts is created.
5. User perspective – the user "interests", described as another set of information items, are processed in a uniform way, and thus the harvested information items are "contextualized" for the user.
6. Statistical methods are applied on contextualized items to identify the emerging topics and trends.

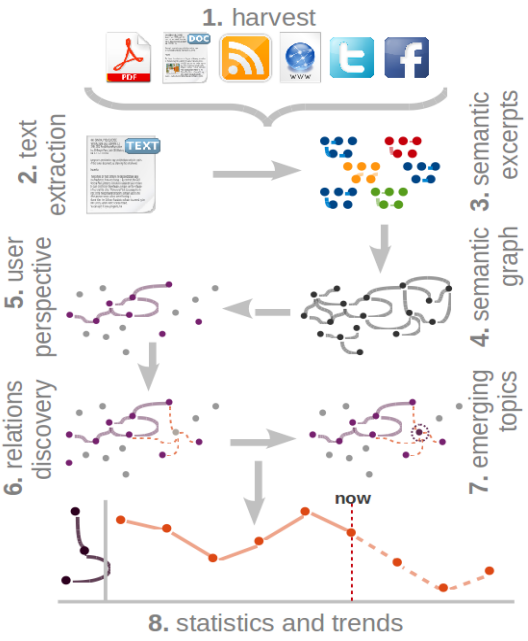


Diagram 1: Core system processes

Each of these processes are described in the following subsections.

3.1.1. Harvesting and Text Extraction

A customizable data harvesting engine is designed and implemented to deal successfully with the various formats of data on the Internet and the almost random intervals of update, as well as to track simultaneously news and social media. Data providers (RSS feeds, Facebook pages and groups, search queries results, queries on Twitter, feeds, websites and document libraries) follow standard schema definitions (Ronallo, 2012), and their properties are stored in a content management system. The object-oriented description of the data provider types creates an abstraction layer between the actual provider's content and the harvesting and storage processes. New data providers can easily be added, and new data medium types, such as audio and video podcasts, can be supported.

The harvester automatically collects structured and unstructured information from the Internet and stores it in an easy-to-process format while omitting irrelevant information such as navigational elements, templates, advertisements, etc. The collected raw data are converted to textual content and metadata using either Apache TIKA, when the content appears to be in a binary format, or a boilerpipe detection library (Kohlschütter et al., 2010), when the content is an (x)HTML page. The text content and the extracted metadata are stored in the content management system as immutable objects.

[13] http://tika.apache.org/

3.1.2. Linguistic Annotation

The system is designed for the processing of multilingual content. At the moment Bulgarian and English are implemented and the technology can be applied to other languages as well. We utilized the ATLAS linguistic framework (Ogrodniczuk and Karagiozov, 2011) as it provides multilingual light-weight automatic text annotation functionalities through a multilingual UIMA-based (Ferrucci and Lally, 2004) framework. The ATLAS framework is capable of segmenting the text and extracting named entities and noun phrases. Furthermore, ATLAS provides text extractive summarization, automatic categorization and cross-language information retrieval (CLIR) modules. The framework is extendable in terms of new annotations (e.g., semantic relations) and new languages.

Language-independent names similarity measure (Steinberger et al., 2011) is implemented in order to automatically link translation equivalents of named entities and noun phrases in a multilingual content, thus facilitating the CLIR-based analysis.

3.1.3. Integrating Language Technologies in Data Processing Algorithms

The dimensionality of the data targeted for analysis is reduced by tf.idf weighting as the top-ranked semantic entities are represented as time series. We use classification algorithms (Ratanamahatana et al., 2010) on these series in order to find cylinders (sharp raise, plateau and sharp drop), funnels (sudden increase and gradual decrease) and bells patterns (gradual increase and sharp drop). As such patterns indicate a significant change, they are used for the identification of emerging topics. Signal analysis of all possible bonds between the identified patterns reveals hidden semantic relations. A coherent signal alerts the user of themes and topics that constitute trends and provides knowledge for further actions. The signal route represents the evolution of processes and events within the series, allowing for identification of relevant data and generation of recommendations and conclusions.

The most relevant content items encompassing the identified semantic entities are clustered. Most text clustering algorithms (Aggarwal and Zhai, 2012) can easily group texts into clusters but provide synthetic labels (if any labels at all) which are far from meaningful. Instead, we have adopted the Lingo3 clustering algorithm (Osinski et al., 2004), which decides on cluster labels prior to the clustering. Such clusters are used for showing the user what is happening at the moment, or what has happened several hours ago, yesterday, last week or last month.

Furthermore, the user is able to track the relations between different semantic entities. Performing deep semantic analysis in a multilingual environment is a complex task and requires a lot of language-specific resources (Navigli and Ponzetto, 2012). Thus we assume that two semantic entities (concepts, people, locations, etc.) are related if they both appear in a sentence. Additionally, we implement anaphora resolution in order to replace pronouns with their antecedents if the latter are recognized as semantic entities. After that the semantic entities are indexed as bi- and tri- grams in a language-independent SOLR[14] core. Each n-gram is considered to represent a semantic relation, and a semantic relations graph is built on top of a SOLR index.

3.1.4. User Profiles

The system provides only those fragments of the information flow that are most relevant to the users' interests. To achieve this, the user profile is dynamically built using a daily analysis of the user's sources. The profile serves as a pattern against which multilingual textual content from digital media sources and social networks is screened and rendered.

User interests are not described with static keywords but are derived from data provided by or related to the user – websites, documents, news items, etc. The same process of harvesting and linguistic processing is applied to the user data, after which we cluster the user content and formulate the user's interests. Further on, we create supervised models which are later used for automatic categorization (channeling) of items in the information stream towards the user profile. We apply chi2 feature reduction (Manning et al., 2008), subsequently building a smoothed naïve Bayesian model (Chen and Goodman, 1996).

[14] `http://lucene.apache.org/solr/`

The user profile, represented as a graph of semantic relations, factors the full semantic graph using a graph pattern-matching algorithm (Gallagher, 2006). All other analysis – summarization, relations tracking, identification of emerging topics, suggestions for further evolution of the user profile – are based on the user-factored semantic graph.

3.2. Lambda Architecture

There are several important non-functional requirements behind the system that are necessitated by the complexity of the technologies: (a) robustness and fault tolerance in distributed environment, which address the random changes in machine and human behavior; (b) low latency in reads and updates needed for modeling near-real-time data analysis; (c) ad-hoc queries that support business optimization and new applications of the data and the system; (d) scalability, needed to address the increasing data volume and system load.

Traditional software architectures only partially provide the above-listed non-functional requirements. Consequently, we have adopted a novel approach called *Lambda architecture* (Marz and Warren, 2013). Lambda architecture enables the execution of arbitrary functions on arbitrary data in real-time by decomposing the problem into three layers: batch layer, serving layer and speed layer.

Our implementation of the Lambda architecture is as follows: (a) the master data store is maintained through the content management system ATLAS CMS,[15] which seamlessly integrates a linguistic processing framework into the processes of content management; the semantic excerpts – people, places, organizations, salient noun phrases, relations between them, summary of information items and categorization labels – are appended to the master datastore; (b) the batch layer and views are based on a relational database (PostrgeSQL[16]) and a set of SOLR[17] cores: both solutions provide horizontal scalability by replication and shading, which guarantees optimal performance and stability; (c) the serving layer is integrated in the front-end component; as it facilitates the merging process, the same data model is used in the master datastore and the batch.

4. Press-Clipping Case Study

The described system provides an all-in-one monitoring service that simultaneously tracks traditional and social media. In this way, essential information encompassing what is coming from the official media and what is said by people is captured. Hence, a press clipping service implemented on top of our system allows the users to effectively monitor their reputation, particular products or practices and to make timely and well-informed decisions. The service can successfully meet the needs of large international corporations and organizations, publishing houses, news and PR agencies, political entities, etc. Two press-clipping services have been built as a proof-of-concept:

- Press clippings in English:[18] – the service monitors 30 international news agencies and an average of 1,500 news items are processed daily.
- Press clippings in Bulgarian:[19] – the service monitors 44 Bulgarian news sources and blogging sites, and an average of 4,000 news items are processed daily.

Both services feature 12 public profiles mapping the top-level news topics such as World, Business, Entertainment, Sports, Technologies, Health, etc. The English press-clipping service provides special topic-oriented profiles for Terrorism and Security. The Bulgarian press-clipping service provides special topic-oriented profiles for emerging local events, such as Elections 2013, Economical crisis, Judicial system, Crime rate, Caretaker government, etc.

[15] `http://www.atlasproject.eu/`

[16] `http://www.postgresql.org/`

[17] `http://lucene.apache.org/solr/`

[18] The service address is `http://en.mtalk.eu/login`.

[19] The service address is `http://mtalk.eu/login`.

The general functionalities of the services can be accessed through Google or Facebook accounts or by registering as a news user with username and password.

In addition we have created several enterprise-oriented profiles in order to test the user-focused capabilities of the system. As an example we have set up four profiles of interest – gas & oil drilling, biofuels and renewable energy, Arctic shelf and Gulf of Mexico, for the British Petroleum company[20].

Figure 1 shows the main screen of the English-language press-clipping service. The screenshot shows the list of latest news, the news clusters (highlights), people, place and organizations which are currently in the news focus, and the merging topics.

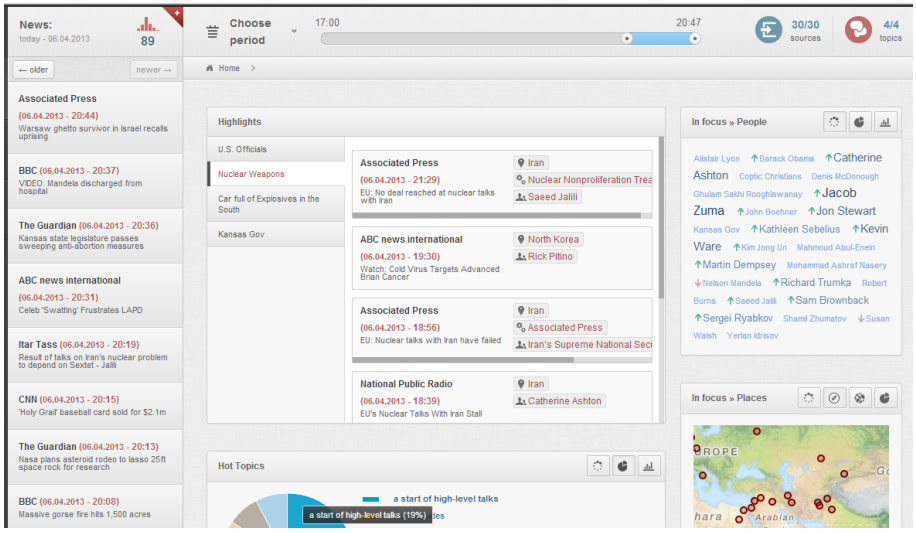


Figure 1: English press-clipping main screen

Figure 2 shows the news filtered through a user profile. In addition to the widgets shown in Figure 1, the screenshot includes statistics of news related to the particular user profile.

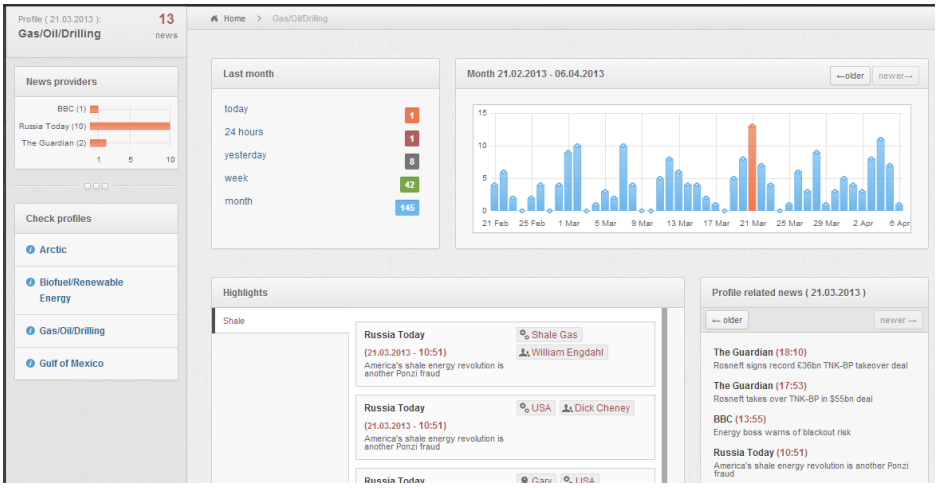


Figure 2: A user profile screen

[20] The press-clipping service for BP can be accessed at: `http://en.mtalk.eu/login` using `bp@en.malk.eu` as username and "bptest" (without the quote marks) as password.

5. Conclusion and Further Work

We have presented a multilingual information channeling system that can analyze formally structured and unstructured data representing users' interests, build user profiles dynamically and channel the information flows through these user profiles. The integration of language technologies and algorithms for analysis of time series utilizes functionalities such as identification of emerging topics, forecasting, and trend predictions. The fusion between an extensible linguistic processing framework and a multilingual content management system allows deeper semantic analysis and support of further languages. The architecture of the system is robust and adheres to the Big Data paradigm.

The system will be gradually extended in the directions of:

- Quality evaluation: Although techniques for evaluation of each individual component in the system (e.g. text categorization, text summarization, machine translation, information extraction) do exist, they are not applicable to the system as a whole. Furthermore, there is no annotated corpus suitable for the evaluation purposes of the multilingual information channeling system. Thus, a set of criteria will be developed in order to evaluate the quality of the implemented workflows and the usability of the system. The envisioned approach consists, not exclusively, of gathering a focus test group of individuals, developing test scenarios and questionnaires and building a manually annotated test corpus.
- Competitive intelligence: In addition to the news streams we will focus on defining, gathering, analyzing and distributing intelligence about products, customers and competitors. A foreseen challenge is the vast volume of potentially interesting information as well as the variety of source media types and data formats (printed, electronic, audio, multimedia) and languages.
- Cross-lingual information retrieval: The system processes information in six languages (English, Bulgarian, Greek, Polish, Romanian and German) and can easily be extended to support other languages. Future research and development will be focused on heavier utilization of CLIR in order to increase the added value of the provided analysis.
- Multi-document summarization: A logical extension of the current topical clustering mechanism is to create a summary report on the set of information items in each cluster. Algorithms as Lexrank (Erkan and Radev, 2004) and semantic graphs (Plaza and Díaz, 2011) can be effectively employed. The harmonization of the CLIR, MT and multi-document summarization components is still an open field for research and development.

References

Aggarwal, C. and Zhai, C. (2012). A Survey of Text Clustering Algorithms. In Aggarwal, C. and Zhai, C. (eds.), *Mining Text Data*, pages 77–128. Springer.

Chen, S. F. and Goodman, J. (1996). *An Empirical Study of Smoothing Techniques for Language Modeling*, pages 310–318. Association for Computational Linguistics.

Erkan, G. and Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence (JAIR)*, 22:457–479.

Ferrucci, D. and Lally, A. (2004). UIMA: an Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10:327–348.

Gallagher, B. (2006). Matching Structure and Semantics: A Survey on Graph-Based Pattern Matching. *In AAAI FS '06: Papers from the 2006 AAAI Fall Symposium on Capturing and Using Patterns for Evidence Detection*, pages 45–53.

Kohlschütter, C., Fankhauser, P., and Nejdl, W. (2010). Boilerplate Detection Using Shallow Text Features. *In Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 441–450. ACM.

Manning, C. D., Raghavan, P. and Schutze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Marz, N. and Warren, J. (2013). *Big Data: Principles and Best Practices of Scalable Realtime Data Systems.* Manning Publications Company.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250.

Ogrodniczuk, M. and Karagiozov, D. (2011). ATLAS Multilingual *Language* Processing Platform. *Procesamiento del Lenguaje Natural*, 47:241–248.

Osinski, S., Stefanowski, J., and Weiss, D. (2004). Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. *Intelligent Information Processing and Web Mining Advances in Soft Computing, 25*:359–368.

Plaza, L. and Díaz, A. (2011). Using Semantic Graphs and Word Sense Disambiguation Techniques to Improve Text Summarization. *Procesamiento del Lenguaje Natural*, 47:97–105.

Ratanamahatana, C. A., Lin, J., Gunopulos, D., Keogh, E. J., Vlachos, M., and Das, G. (2010). *Mining Time Series Data*, chapter Data Mining and Knowledge Discovery Handbook, pages 1049–1077. Springer.

Ronallo, J. (2012). HTML5 Microdata and Schema.org. *The Code4Lib Journal*, 16 (2012-02-03).

Steinberger, R., Pouliquen, B., Kabadjov, M. A., Belyaeva, J., and der Goot, E. V. (2011). JRC-NAMES: A Freely Available, Highly Multilingual Named Entity Resource. *In Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, Hissar, Bulgaria,* pages 104–110.

Automatic Semantic Filtering of Morphosemantic Relations in WordNet

Svetlozara Leseva, Ivelina Stoyanova, Borislav Rizov,
Maria Todorova, Ekaterina Tarpomanova
Institute for Bulgarian Language – Bulgarian Academy of Sciences
{zarka,iva,boby,maria,katja}@dcl.bas.bg

Abstract

In this paper we present a method for automatic assignment of morphosemantic relations between derivationally related verb–noun pairs of synsets in the Bulgarian WordNet (BulNet) and for semantic filtering of those relations. The filtering process relies on the meaning of noun suffixes and the semantic compatibility of verb and noun taxonomic classes. We use the taxonomic labels assigned to all the synsets in the Princeton WordNet (PWN) – one label per synset – which denote their general semantic class.

In the first iteration we employ the pairs *<noun suffix : noun label>* to filter out part of the relations. In the second iteration, which uses as input the output of the first one, we apply a stronger semantic filter. It makes use of the taxonomic labels of the noun-verb synset pairs observed for a given morphosemantic relation. In this way we manage to reliably filter out impossible or unlikely combinations. The results of the performed experiment may be applied to enrich BulNet with morphosemantic relations and new synsets semi-automatically, while facilitating the manual work and reducing its cost.

1. The Morphosemantic Relations in WordNet

Morphosemantic relations are a type of semantic relations which have morphological expression in at least one language (Koeva, 2008), for instance through derivational means. Since these relations link concepts, they are universal and transferable across languages, as has been demonstrated successfully in the context of different initiatives within the WordNet community (Bilgin et al., 2004; Pala and Hlaváčková, 2007; Koeva, 2008; Koeva et al., 2008; Fellbaum et al., 2009; Barbu Mititelu, 2012; Piasecki et al., 2012a; Piasecki et al., 2012b; Dimitrova et al., 2014). The typology and the specifics of a language determine whether the lexemes that lexicalise the respective concepts will be derivationally related. The morphosemantic relations we deal with are the morphosemantic links encoded between derivationally related literals in verb–noun pairs of synsets in the Princeton WordNet – PWN (Fellbaum et al., 2009).

Currently a relatively small portion of the derivationally related synsets in the PWN are supplied with a semantic label. In the PWN 3.0 version used in this paper there are 36,142 pairs of derivationally related verb–noun synsets, with at least one pair of derivationally related literals in each pair of synsets, while morphosemantic links have been assigned to 17,740 pairs of literals.

These relations have been mapped from the stand-off file distributed with the PWN[1] to the corresponding synsets in the Bulgarian WordNet (Koeva, 2010) using the cross-language relation of equivalence between synsets (Vossen, 2004).

13 out of the 14 types of morphosemantic relations encoded in the PWN denote a relation between a predicate and a participant in its semantic representation and hence correspond to thematic roles. Those are: Agent, By-means-of (corresponding to inanimate Agents or Causes but also to Means), Instrument,

[1] http://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls

Material, Body-part, Uses (function of purpose) Vehicle (means of transportation), Location, Result, State, Undergoer, Destination, Property[2]. The only exception is the relation Event, which links a verb to a deverbal noun denoting the same event.

Derivationally related verb–noun pairs may be obtained through direct or non-direct derivational paths. Moreover, the direction of the derivation is usually not taken into account. In the pair *programiram* ("to program") – *programa* ("computer program"), which is assigned the morphosemantic relation Result, the verb is produced from the noun via direct derivation. However, the relation (Agent) between *programiram* and *programist* ("computer programmer") results from non-direct derivation, since both words are derived independently from the noun *programa*.

The derivation between a pair of literals may involve one or more derivational steps. For example, the place noun *kovachnitsa* ("forge, smithy") is produced from the verb *kova* ("forge, hammer") in two steps: first an agentive noun *kovach* ("(black)smith") is formed with the suffix *-ach*, and then the place suffix *-nitsa* is attached to the agentive noun base.

We identified the following cases of derivationally related pairs in Bulgarian that remain unconnected by means of a morphosemantic relation after the automatic transfer from the PWN. For a derivationally related pair of synsets in Bulgarian: (i) the corresponding synsets in the PWN may not be derivationally related, e.g. *kova* ("hammer") – *kovach* ("blacksmith"); (ii) the English corresponding noun and/or verb may be compounds, and therefore – unrelated – e.g. *chakam* ("wait") – *chakalnya* ("waiting room"); (iii) there may be a derivational relation in the PWN but it is not assigned a morphosemantic link, e.g. *izvarsha* ("perpetrate") – *izvarshitel* ("perpetrator").

Our goal is to discover derivationally related literals in verb–noun pairs of synsets in BulNet, such as the ones in (1-3), and to assign these pairs one or more morphosemantic relations using the semantics of the derivational means (focusing on suffixes). We assume that each morphosemantic relation corresponds to a distinct sense of a given suffix. Many suffixes express more than one morphosemantic relation. Usually, the knowledge about the semantics of the suffix is not sufficient alone to predict the morphosemantic relation unambiguously. We try to disambiguate fully or partially the possible morphosemantic relations for a given suffix by applying further semantic filtering. In this way we aim to facilitate the manual work on encoding and/or validating new instances of the morphosemantic relations. Once validated, they may be transferred to other languages.

The method uses a language-independent module – an inventory of morphosemantic relations obtained from the PWN automatically, and two language-dependent modules: (i) an inventory of suffixes and suffix variants; and (ii) mapping between suffix variants and suffix canonical forms. The former of the language-dependent modules is acquired automatically while the latter involves manual work. The method can be adapted relatively effortlessly to other sets of morphosemantic relations implemented in other wordnets, and to other languages. In the first case it would require an extension of the language-independent module (by transferring relations from another wordnet), and in the second case – an adaptation of the affix recognition algorithm and subsequent mapping to canonical forms.

2. Establishing Derivationally Related *Verb–Noun* Pairs and an Inventory of Affixes

After the assignment of the morphosemantic relations, an algorithm for recognising derivationally related pairs of verb–noun literals (Lv–Ln) was implemented (Dimitrova et al., 2014). The algorithm relies on string similarity and heuristic procedures. Similarity is established if at least one of the following conditions are met: (i) one of the literals is a substring of the other; (ii) the two literals have a common beginning (estimated to be at least half the length of the shorter literal); (iii) the two literals have a Levenshtein distance smaller than a given empirically determined value. This procedure resulted in linking 6,135 verb–noun pairs, each of which was validated manually.

2.1. Establishing an Inventory of Affixes

In order to establish the inventory of derivational patterns and the morphosemantic relations expressed by each of them, we extracted those 6,135 Lv–Ln pairs and identified the substrings which we assumed

[2] http://wordnetcode.princeton.edu/standoff-files/morphosemantic-links-README.txt

contained the affixes involved in the derivation. An expert linguist inspected the unique verb and noun beginnings and endings and associated each of them with a canonical form of the respective affix(es), suffixes in particular. This process was required because unlike prefixes, which usually concatenate with the stem, suffixes are realised by a number of morphophonemic variants due to the fact that their attachment to the stem may be accompanied by vowel and consonant changes in both the stem and the suffix. For example, the suffix *-nie* is also realised by the following variants: *-anie, -enie, -zhdenie, -zhenie, -zanie, -lenie, -sanie, -ovanie, -ovlenie, -shenie, -shlenie, -ovenie, -iyanie*. We identified 62 verb and 228 noun patterns of the type *canonical suffix* > *suffix variant*. Regular sound alternations resulting from assimilation and dissimilation processes, such as *k* > *ts, k* > *ch, g* > *h, sh* > *s*, etc. were also taken into consideration.

2.2. Suffix Normalisation in Detail

1. Given a pair of derivationally related literals Lv–Ln belonging to the synsets Sv and Sn, which are linked via a morphosemantic relation, we remove the vowels and find the longest similar substrings so that one of the substrings can be produced from the other. This is achieved using a dynamic programming algorithm. The vowel removal aims at reducing the phonetic alternations in the stems of the related forms caused by different linguistic phenomena, such as metathesis, e.g. *krav* ("blood") – *okarvavya* ("blood"); vowel mutation, e.g. *izbor* ("choice") – *izbera* ("choose"); elision, e.g. *bera* ("pick, pluck") – *brane* ("picking, plucking"), etc. In the examples the algorithm identifies the common strings – *krv*, *izbr* and *br* respectively. The common substring expanded by the vowels between the consonants is considered to be an approximation of the stem. The stem variants are generated by including/excluding the bordering vowels. For each stem, the remaining substring(s) that either precede(s) it (conditionally called a *prefix*), in the first example *o-*, or follow(s) it (conditionally called a *suffix*) – *-avya* in the same example), are established. They are subsequently checked against a list of prefix and suffix variants and the longest matches are selected.

2. We map the *suffix* substrings found in the noun literals to the list of canonical noun suffixes on the basis of the patterns *canonical suffix* > *suffix variant*, looking for the longest match. We are interested in noun suffixes since they express the morphosemantic relations under consideration, while verb suffixes have mainly a grammatical meaning. Finally, the results were post-edited manually. 83 canonical noun suffixes were established.

The normalisation of affixes helps in two ways. It allows us (i) to identify more reliably the morphosemantic relations expressed by each affix; and (ii) to reduce data sparsity that arises from the morphophonemic variants.

3. Establishing an Inventory of Pairs <*Affix : Relations*>

For the literals containing a given canonical noun suffix, we calculated the types of morphosemantic relations with which it is associated and the number of instances for each relation. Out of the 83 noun suffixes, 32 are unambiguous (one morphosemantic relation per suffix). The largest portion of the unambiguous suffixes denote the relation Agent (13 suffixes), followed by Event (7), and the remaining 12 are distributed among several relations – Material, Result, Undergoer, Property, State, Instrument. The rest 51 suffixes are ambiguous. The number of senses for all the noun suffixes is 252. Not all the predictions are accurate since some <*affix : relations*> pairs are attested in few instances or not attested at all.

The senses expressed by a suffix are not arbitrary but clustered around a given relation which is the preferred reading for this suffix. Table 1 shows that for the most productive suffixes that express the relation Agent the majority of instances are cases of default reading, and the rest of the relations are represented by much fewer examples.

The other senses of the suffixes given in Table 1 also have agentive properties since they denote inanimate agents and causes, such as Instrument, Material, By-means-of, Vehicle. Certain agentive suffixes can also express the relation Undergoer when the verb is unergative, e.g., *rabotnik* ("worker" – a person who works at a particular occupation).

	Agent	Instrument	Material	Undergoer	Vehicle	By-means-of	other
-tel	169	13	17	1	-	6	1(Event),1(Uses)
-(y)ach	128	2	-	2	2	1	-
-(n)ik	87	2	1	4	-	-	3(Event)
-sht	83	-	-	4	-	-	-
-tor	42	15	12	-	-	8	3(Result),1(Uses)

Table 1: Distribution of senses of the top 5 agentive suffixes

So even though many of the suffixes are ambiguous, at least for a part of them the ambiguity is very predictable. The examination of the data shows that the different senses of a given suffix are to a great extent taxonomically distinct. For instance, nouns with the suffix *-(n)ik* which are Agents, are persons, Instruments are artifacts, Materials are substances. This works also for untypical suffix senses. For instance, the suffixes *-ne, -stvo, -tsiya* may express the relation Agent due to a metaphorical extension of the meaning of some eventive deverbal nouns to denote Agents. Since persons cannot be Events, these suffixes may be disambiguated on the basis of the noun semantics alone; in these cases the semantic (taxonomic) class of the noun is a very strong indicator for the relation.

4. Semantic Filtering

In order to (partially) filter out the possible combinations <*suffix : morphosemantic relation*> we explore the possibility of using the taxonomic restrictions imposed by each suffix as a semantic filter.

The taxonomic distinctions between the different senses of the suffixes largely correspond to natural semantic classes, such as persons, artifacts, locations, acts, etc. Being a linguistic taxonomy, WordNet distinguishes these classes.

The PWN synsets are organised in 45 lexicographer files (26 – nouns, 15 – verbs, 4 – for the other parts of speech) based on the syntactic category and the taxonomic class of a synset[3]. Nouns denoting people are found in the file *noun.person*, nouns denoting feelings and emotions – in the file *noun.feeling*, etc. This allows us to use the file names as taxonomic labels for the noun and verb synsets.

Given that (i) there is an algorithm that recognises the suffix of a word and associates it with its canonical form, and (ii) the taxonomic label and the morphosemantic relations associated with a canonical suffix can be obtained from the synsets, we can use those labels to filter the morphosemantic relations associated with a given suffix.

1. From the already validated noun literals and the synsets to which they belong we extract the pairs <*suffix: morphosemantic relation*>. For example, for the suffix *-(n)ik*, the following morphosemantic relations are licensed:

   -(n)ik: agent, *-(n)ik: undergoer*, *-(n)ik: instrument*, *-(n)ik: material*, *-(n)ik: event*

2. Given the pair <*suffix: morphosemantic relation*>, we rule out the taxonomically incompatible morphosemantic relations, that is, those relations that have not been attested for the pair <*suffix : taxonomic label*> in BulNet and obtain triples of the type <*suffix : taxonomic label : morphosemantic relation*>. For example, after applying this semantic filter, for the suffix *-(n)ik* we acquire the following triples:

   <*-(n)ik : noun.person : Agent, Undergoer*>

   <*-(n)ik : noun.artifact : Instrument*>

   <*-(n)ik : noun.substance : Material*>

   <*-(n)ik : noun.act : Event*>

[3] http://wordnet.princeton.edu/wordnet/man/lexnames.5WN.html

The triples represent the linguistic generalisations for the semantic restrictions on the senses of the suffixes. Those predictions are based on and therefore limited by the observed instances.

Since the algorithm which discovers derivationally related verbs and nouns links all the pairs that may be mapped by it, two types of problems arise: (i) erroneously linked unrelated words due to coincidental string similarity, such as in *slon* ("elephant") and *podslonya* ("to shelter"); (ii) overgeneration due to the lack of semantic restrictions on the verbs: for example, the noun *zaemane* ("loan") is connected not only to the verb *zaema* ("to loan"), but also to homonyms, such as *zaema* ("to assume a pose").

The first issue requires further improvement of the recognition algorithm, which we leave for future research. Overgeneration can be at least partially resolved by introducing additional semantic filters. To this end, we decided to explore further the potential of the taxonomic labels. For each instance of a morphosemantic relation we retrieve the taxonomic labels of the respective noun and verb synsets and calculate the frequency of occurrence of each triple *<morphosemantic relation : verb.label : noun.label>* in the PWN. For example, *<Agent : verb.communication : noun.person>* has 411 instances, or 15.83% of the instances of Agent, followed by *<Agent : verb.social : noun.person>* – 337 instances, or 12.98%. Certain patterns such as *<Agent : verb.change : noun.plant>* have few occurrences (1 instance or 0.01% of the occurrences of Agent). The low frequency of a pattern may indicate a specific or semantically restricted relation, compare *author* ("be the author of") – *author* ("writer"), as opposed to *tense* ("become tense, nervous, or uneasy") – *tensor* ("any of several muscles that cause an attached structure to become tense or firm"). The first pair illustrates a typical Agent relation between a verb of creation and a noun person, and the second one exemplifies a more specialised Agent-like relation (Body-part), which involves verbs and nouns from semantically restricted classes (bodily functions, movements, etc. and a part of the body that performs them respectively). Low frequency may also indicate semantically dubious or unlikely relations, such as the Agent relation between *titter* ("laugh nervously") and *titter* ("a nervous restrained laugh") assigned in the PWN. Although we filter out the combinations with low frequency, we consider including rarely seen legitimate patterns manually at a later stage (see Section 7.).

In order to test the application of semantic patterns to the task of semantic filtering, we set up an experiment, which we describe in the following Section.

5. Experimental Method

The experiment consists in: (1) identifying derivational pairs in BulNet that have not been assigned a morphosemantic relation, and predicting the probable morphosemantic relations for each of the pairs on the basis of information about the suffix senses and taxonomic classes; and (2) filtering out a part of these relations using semantic criteria. The main purpose of the method is to facilitate the manual validation of automatically assigned morphosemantic relations. Manual inspection is nevertheless necessary in order to ensure high-quality data that can be used for training various linguistic models and applications.

1. **Identification of potential derivational pairs Ln–Lv in BulNet.** This step requires two distinct procedures: (i) recognition of derivational pairs, and (ii) identification of the canonical suffix of the noun literal in the pairs.
   (a) **Recognition:**
      i. Given a noun in BulNet – Ln, look up its ending in the list of morphophonemic variants of the noun suffixes.
      ii. If the ending is found in the list, remove it from the word.
      iii. If the remaining string is at least 4 characters long, attach to it a verb suffix from the list of the morphophonemic variants of verb suffixes.
      iv. If the resulting word is a legitimate verb in BulNet – Lv, find all the verb synsets in which Lv occurs.
   (b) **Mapping:**
      i. Given a pair Ln–Lv recognised at the previous stage, map the morphophonemic variant of the suffix of Ln to its canonical form. In this way we acquire all the instances of a given suffix, regardless of the morphophonemic environment.

ii. For a given pair Ln–Lv, retrieve all the synsets Sn and Sv in which they are found.

2. **Semantic filtering.** The semantic filtering is performed in two steps. The output of Step 1 serves as input for Step 2.

   (a) **Step 1.**
      i. For each Ln, retrieve all the morphosemantic relations licensed by the combination of the suffix and the taxonomic label of the synset in which Ln is found by intersecting the possible pairs <*canonical suffix : noun.label*> with the possible pairs <*canonical suffix : morphosemantic relation*>.
      ii. Assign all the morphosemantic relations licensed by Ln to the pairs Sn–Sv, such that Sn contains Ln and Sv contains Lv.

   (b) **Step 2.**
      i. Given the frequency of occurrence of a triple <*morphosemantic relation : verb label : noun label*> in the PWN, estimate the probability of each triple in the output of Step 1.
      ii. For a given probability threshold, filter out the Ln–Lv pairs that are below the threshold. By varying the threshold we can obtain balance between precision and coverage in accordance with the particular purposes – a lower threshold means a larger number of assigned relations and more manual work on their validation but higher recall, and vice versa. We determine the threshold empirically on randomly selected samples of the data (see Section 6.).

6. Results

57,771 derivationally related literal pairs were identified at Step 1, out of which 7,601 pairs could not be assigned a morphosemantic relation because the particular semantic pattern <*canonical suffix : taxonomic label*> had not been observed previously in the manually validated literals. These pairs need to be examined systematically so that we can extend the already discovered combinations with new attested patterns, such as <*-iya : noun.body*>, which was found in the pair *anatomiya* ("anatomy, a human body") – *anatomiziram* ("anatomize"). The remaining pairs were assigned one or more relations (up to 8) out of the 14 morphosemantic relations, which amounted to a total of 219,597 relations assigned.

At Step 2, in order to determine the threshold, we experimented with several values from 0.1 to 0.9 set apart by 0.1, by observing the proportion of assigned relations, on the one hand (Table 2), and by evaluating the precision and recall on random samples, on the other. The samples included (i) 100 automatically assigned relations, and (ii) 100 discarded relations for each threshold value[4] (Figure 1). Each threshold is evaluated using $F_{0.5}$ measure, where precision is given twice as much weight as recall, although results were consistent for other F_β measures for $0 < \beta < 1$. The highest $F_{0.5}$ measure of 0.882 was achieved for a threshold of 0.7.

The performance of the method with the selected threshold of 0.7 was evaluated using the following set of criteria:

- Efficiency – it was evaluated in terms of the reduction in the number of assigned morphosemantic relations as a measure of the feasibility of further manual validation. The total number of 219,597 relations was reduced to 26,766 (12.19% of the total). Moreover, the number of highly ambiguous cases of initial relation assignment was markedly decreased by an average factor of 6.76. As a result, the manual validation of the semantic filtering is rendered much more tractable.

- Precision and recall – the precision and recall were estimated based on a different set of samples of 100 assigned relations (above the threshold) and 100 discarded relations (below the threshold), using the formulae:

$$\text{Precision} = \frac{\text{correctly assigned}}{\text{all assigned}}, \qquad \text{Recall} = \frac{\text{correctly assigned}}{\text{correctly assigned} + \text{incorrectly discarded}}$$

[4]Assuming that all the possible morphosemantic relations were identified in advance, precision can be calculated as the percentage of the correctly assigned relations out of all the assigned relations, and recall – as the percentage of all the correctly assigned relations out of all the correct relations (assigned and discarded).

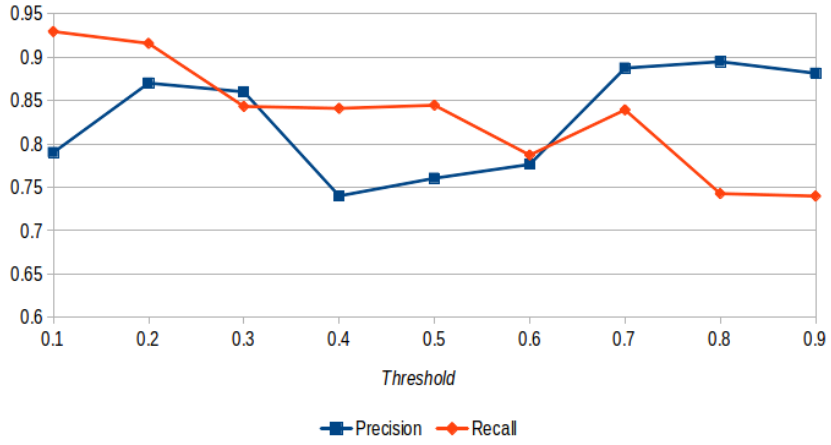


Figure 1: Precision and recall for various thresholds

Threshold	Assigned,#	Assigned,%
0	219,597	100.00
0.1	77,987	35.51
0.2	54,409	24.78
0.3	46,673	21.25
0.4	40,495	18.44
0.5	34,338	15.64
0.6	30,376	13.83
0.7	26,766	12.19
0.8	26,195	11.93
0.9	24,461	11.14

Table 2: Decrease in the number of relations using various thresholds

With a precision of 0.90, the results are promising and justify the application of the method for semi-automatic expansion of WordNet with morphosemantic relations. The relations that were filtered out were predominantly invalid, resulting in recall of 0.84. This result leads us to the conclusion that the implemented semantic filtering largely preserves the coverage of the morphosemantic relations.

7. Discussion

In order to improve the method, we performed a preliminary error analysis, focusing on 2,000 pairs of literals that had been assigned a single morphosemantic relation at Step 1. Three types of errors were identified: (i) the pair of words is wrongly recognised due to coincidence of symbol strings; (ii) the words in a pair are derivationally related but none of the defined morphosemantic relations is appropriate; (iii) the words have a derivational relation but the assigned morphosemantic label is wrong. With respect to the third type of errors we draw two directions for further improvement.

1. **Enriching the semantic description of suffixes.** In the collection of automatically assigned morphosemantic relations we observed valid suffix senses unattested in the synsets related through morphosemantic relations in the PWN. For instance, the suffixes *-er/-ier/-ur* and *-in* had not been attested with the meaning of Undergoer and when such cases were discovered in the automatically assigned pairs: *pensioner* ("pensioner, retired person") – *pensioniram* ("superannuate, retire") and *grazhdanin* ("citizen") – *pograzhdanyavam* ("urbanise"), the nouns were incorrectly recognised as Agents. The systematic exploration of falsely assigned relations will make it possible for us to draw a full description of the semantics of the suffixes and to make more precise predictions.

2. **Enriching the semantic restrictions imposed by the taxonomic labels.** In analysing the pairs

that were falsely assigned a morphosemantic relation, we observed new semantic restrictions. For instance, the agentive suffix *-ach/-yach* had been found in the synset {povdigach:1; levator:1} ("a muscle that serves to lift some body part") which has the taxonomic label *noun.body*. We looked up for other synsets with the same taxonomic tag that also contain a noun with the suffix *-ach/ -yach* but were assigned a morphosemantic relation. We found such an example – {obtegach:1; tensor:1} ("any of several muscles that cause an attached structure to become tense or firm") – which was assigned the relation Body-part. As a result, we acquired the following generalisation *<-ach/ -yach : noun.body : Body-part>*. Although this pattern has a low frequency, it shows very distinct semantic properties so it can be safely included in the list of semantic restrictions. Respectively, the suffix senses also need to be updated. This line of research is directed towards increasing the coverage of morphosemantic relations.

8. Related Work

The task of recognising and/or generating derivatives from existing words in a wordnet is explicitly or implicitly directed towards the expansion of a wordnet with new synsets and relations, and/or the transfer of those synsets and relations to other wordnets (Bilgin et al., 2004; Pala and Hlaváčková, 2007; Koeva, 2008; Koeva et al., 2008; Piasecki et al., 2012a; Stoyanova et al., 2013).

We focus on the task of assigning new instances of the morphosemantic relations and proposing an algorithm for (partial) disambiguation by means of semantic filters. In a similar vein, Piasecki et al. (2012a) use a bigger inventory of relations which include the morphosemantic relations in the PWN to the end of training a tool to discover derivational pairs of words and to suggest derived words missing in the Polish WordNet. The authors discuss the possibility of using semantic information obtained from WordNet, such as upper-level hypernyms and semantic domains, to filter erroneous pairs. Piasecki et al. (2012b) propose a method for semantic classification of verb–noun derivational relations using supervised machine learning. Their approach uses context features of the derivationally related pairs observed in a huge corpus to disambiguate the derivational relations, whereas our method employs semantic patterns observed in the Princeton WordNet. Our proposal is closest in spirit to the work of Stoyanova et al. (2013), who suggest filtering morphosemantic relations assigned automatically to derivationally related pairs of synsets by means of a semantic filter based on the taxonomic labels in WordNet. The results of their experiment have not been reported in detail. Drawing on their idea, we further expand on and test the hypothesis that together with the semantics of suffixes verb–noun taxonomic labels are a reliable semantic filter for morphosemantic relations of the type discussed herein.

9. Future Directions

The methodology reported in this paper gives promising results. Future work will be focused on exploring the possibilities of mutually disambiguating the suffixes of words from the same synset on the basis of their senses and the semantic restrictions imposed by them both in a monolingual and in a multilingual setting. As suggested in the previous Section, the analysis of the errors and the cases where no relation is assigned will be further employed to identify and collect new semantic restrictions imposed by suffixes and possibly new suffix senses. The application of additional semantic filters, such as upper-level hypernyms, will also be explored. Another line of research that is worth investigating is the application of the method to enriching WordNet with new synsets on the basis of morphosemantic relations.

Acknowledgements

The present paper was prepared within the project *Integrating New Practices and Knowledge in Undergraduate and Graduate Courses in Computational Linguistics* (BG051PO001-3.3.06-0022) implemented with the financial support of the Human Resources Development Operational Programme 2007-2013 co-financed by the European Social Fund of the European Union. The authors take full responsibility for the content of the present paper.

References

Barbu Mititelu, V. (2012). Adding Morpho-semantic Relations to the Romanian Wordnet. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2596–2601.

Bilgin, O., Cetinoglu, O., and Oflazer, K. (2004). Morphosemantic Relations in and across Wordnets – A Study Based on Turkish. In *Proceedings of the Second Global Wordnet Conference (GWC 2004)*, pages 60–66.

Dimitrova, T., Tarpomanova, E., and Rizov, B. (2014). Coping with Derivation in the Bulgarian WordNet. In *Proceedings of the Seventh Global Wordnet Conference (GWC 2014)*, pages 109–117.

Fellbaum, C., Osherson, A., and Clark, P. (2009). Putting Semantics into WordNet's "Morphosemantic" Links. In *Proceedings of the Third Language and Technology Conference, Poznan, Poland. [Reprinted in: Responding to Information Society Challenges: New Advances in Human Language Technologies. Springer Lecture Notes in Informatics]*, volume 5603, pages 350–358.

Koeva, S., Krstev, C., and Vitas, D. (2008). Morpho-semantic Relations in Wordnet – A Case Study for two Slavic Languages. In *Proceedings of the Fourth Global WordNet Conference (GWC 2008)*, pages 239–254.

Koeva, S. (2008). Derivational and Morphosemantic Relations in Bulgarian Wordnet. *Intelligent Information Systems*, pages 359–368.

Koeva, S. (2010). Bulgarian Wordnet - Current State, Applications and Prospects. In *Bulgarian-American Dialogues*, pages 120–132. Sofia: Prof. M. Drinov Academic Publishing House.

Pala, K. and Hlaváčková, D. (2007). Derivational Relations in Czech Wordnet. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, pages 75–81.

Piasecki, M., Ramocki, R., and Maziarz, M. (2012a). Automated Generation of Derivative Relations in the Wordnet Expansion Perspective. In *Proceedings of the 6th Global Wordnet Conference (GWC 2012)*, pages 273–280.

Piasecki, M., Ramocki, R., and Minda, P. (2012b). Corpus-based Semantic Filtering in Discovering Derivational Relations. In Ramsay, A. and Agre, G. (eds.), *Applications – 15th International Conference, AIMSA 2012, Varna, Bulgaria, September 12-15, 2012. Proceedings. LNCS 7557*, pages 14–22. Springer.

Stoyanova, I., Koeva, S., and Leseva, S. (2013). WordNet-based Cross-language Identification of Semantic Relations. In *Proceedings of the 4th Biennal International Workshop on Balto-Slavonic Natural Language Processing*, pages 119–128.

Vossen, P. (2004). EuroWordNet: A Multilingual Database of Autonomous and Language-Specific Wordnets Connected via an Inter-Lingual Index. *International Journal of Lexicography*, 17(3):161–173.

Noun-Verb Derivation in the Bulgarian and the Romanian WordNet – A Comparative Approach

Ekaterina Tarpomanova, Svetlozara Leseva, Maria Todorova,
Tsvetana Dimitrova, Borislav Rizov
Institute for Bulgarian Language – Bulgarian Academy of Sciences
{katja,zarka,maria,cvetana,boby}@dcl.bas.bg

Verginica Barbu Mititelu, Elena Irimia
Research Institute for Artificial Intelligence – Romanian Academy
{vergi,elena}@racai.ro

Abstract

Romanian and Bulgarian are Balkan languages with rich derivational morphology that, if introduced into their respective wordnets, can aid broadening of the wordnet content and the possible NLP applications. In this paper we present a joint work on introducing derivation into the Bulgarian and the Romanian WordNets, BulNet and RoWordNet, respectively, by identifying and subsequently labelling the derivationally and semantically related noun-verb pairs. Our research aims at providing a framework for a comparative study on derivation in the two languages and offering training material for the automatic identification and assignment of derivational and morphosemantic relations needed in various applications.

1. Introduction

Wordnet enrichment by linking synsets via semantically labelled derivational relations (called morphosemantic relations) has been reported for Turkish (Bilgin et al., 2004), Czech (Pala and Hlaváčková, 2007), Serbian (Koeva et al., 2008), Polish (Piasecki et al., 2009), Romanian (Barbu Mititelu, 2012) and Bulgarian (Dimitrova et al., 2014; Koeva, 2008; Stoyanova et al., 2013), among others. Depending on the derivational specificities of the language and / or the methodology adopted, different, possibly overlapping sets of morphosemantic relations have been identified and implemented in the different wordnets.

In this work we consider the morphosemantic relations encoded in the Princeton WordNet (PWN) (Fellbaum et al., 2009) as a stand-off file, which have been transferred automatically in BulNet and RoWordNet. These semantic links are established between literals "that are similar in meaning and where one word is derived from the other by means of a morphological affix" (Fellbaum et al., 2009). Although these relations are morphologically expressed on particular pairs of lexemes (literals) in English (and possibly in other languages), they also hold between the synsets to which these literals belong, given the semantic dimension of the relation.

The PWN morphosemantic links were automatically transferred to the Bulgarian and the Romanian WordNet provided that both synsets that were members of a relation were present. Afterwards, the teams working on the two wordnets performed automatic extraction of literal pairs and derivational models from the morphosemantically related synsets, followed by manual validation of the pair members.

The goal of this paper is to summarise the findings of our joint work with a view to proposing a framework for the automatic discovery of derivational relations and the automatic assignment of morphosemantic relations which makes use of the rich inventory of derivational patterns of the languages under study. A further objective is to implement these linguistic generalisations in applications that benefit from the existence of such wordnet relations.

2. Derivational Morphology of Bulgarian and Romanian

Bulgarian and Romanian show great similarities in derivational morphology due to the common Indo-European inheritance, and to the interaction of the two languages in the Balkan Sprachbund. In both languages suffixation is the most productive means of word formation, but also the most complicated: one or more suffixes may be added to a stem, or a suffix may be substituted for another; suffixation may or may not change the part of speech of a word, while prefixation (usually) does not change the part of speech. In Bulgarian prefixes have an important role for verb-to-verb derivation as they may involve change of verbal aspect. The two derivation processes, suffixation and prefixation, may occur simultaneously to form a new word (parasynthetic derivation) in both languages.

Conversion is a disputable notion in the traditional linguistic descriptions of both Romanian and Bulgarian. According to the Romanian tradition it is distinct from derivation and always implies homonymy. In the Bulgarian literature conversion is usually interpreted in a broader sense as a process of word formation in which the written forms of two words in a derivational pair differ only by their inflectional markers: (*rabotya* ("to work") – *rabota* ("work"). Formation of deverbal nouns by removing the thematic vowel and the inflection of a verb without adding a suffix to the noun, such as in (*nanizha* ("to string") – *naniz* ("string"), are called zero suffixation. In the Bulgarian data discussed below zero suffixation is subsumed under conversion and labelled accordingly. Cases of conversion in Romanian are not discussed in this paper, as it does not serve creating verbs (in their infinitive form) from nouns or, vice versa, nouns from infinitives.

Word formation in Bulgarian and Romanian often involves vowel or consonant alternations, or both. Most of the alternations are phonetically motivated (metaphony, palatalisation), others mark grammatical forms (apophony in Bulgarian). Because of their irregular behaviour, while phonetic alternations often impede the automatic detection of derivational pairs.

3. The Nature of Morphosemantic Relations

The morphosemantic relations encoded in the Princeton WordNet and transferred to BulNet and RoWordNet, usually denote a relation between a predicate and a participant in its semantic representation. In consequence, most of the relations correspond to thematic roles in the representation of the respective derivationally related predicates: Agent, Event, State, Result, Undergoer, Property, Vehicle, Destination, Material, Body-part, Cause, Instrument, Location, By-means-of, (cf. Fellbaum et al. (2009). The only exception is the relation Event, which links verbs to deverbal nouns denoting the same event.

Given the semantic dimension of morphosemantic relations, the semantic label associated with such a relation holds between synsets and is transferable across languages, even though the morphological relation (between literals) needs not be expressed (Koeva, 2008). Besides, even in a language in which a morphosemantic relation or an instance of such a relation has morphological expression, its specific semantics may be derivationally expressed only by certain literal pairs in the respective synsets. For example, consider the synsets *write, compose, pen, indite – writer, author*, where only *write* and *writer* are derivationally related, as opposed to *cry, weep – weeper, crier*, where both *cry* and *crier* and *weep* and *weeper* are morphologically related.

The morphosemantic relations may be expressed through direct or non-direct derivation (obtained through different derivation paths). Consider the Bulgarian verb *analiziram* ("to analyse") and the agentive noun *analizator* ("analyser"), each derived independently from the noun *analiz* ("analysis"). In this case, two pairs are linked via morphosemantic relations – *analizator* ("analyser") – *analiziram* ("to analyse") (Agent) and *analiz* ("analysis") – *analiziram* ("to analyse") (Event). Also, the derivation path may involve more than one operation, such as the two-step derivation of the nouns from the corresponding verbs in: Bulgarian *kova* ("to forge") > *kovach* ("blacksmith") > *kovachnitsa* ("a forge"), and Romanian *topi* ("to melt") > *topitor* ("melter") > *topitorie* ("foundry").

4. Methods of Assigning Morphosemantic Relations

In the framework of our research we assigned the morphosemantic relations from the PWN stand-off file to the lexicalised synsets in BulNet and RoWordNet and checked the literals in the relevant pairs of

synsets to establish the derivationally related pairs of literals. We used different approaches for identification of the derivational subsets in BulNet and Ro-WordNet as each team had started working independently, with different resources at their disposal and with different aims. The particularities of the methods used for each language are described in Dimitrova et al. (2014) and Barbu Mititelu (2012).

The derivationally related pairs of literals were verified manually. In Romanian 2,767 pairs were found altogether, with the following distribution: 2,429 cases of suffixation, 318 cases of verbal suffixation equivalent to the Bulgarian conversion and 20 cases of parasynthetic derivation. In Bulgarian 6,135 pairs were found, as follows: 4,590 cases of suffixation, 930 cases of substitution of a noun suffix for a verb suffix or vice versa, 433 cases of conversion, 139 cases of prefixation, 12 cases of parasynthetic derivation, 31 cases of non-transparent derivation.

5. Expression of Morphosemantic Relations through Derivational Patterns

Morphosemantic relations are expressed both at the form level and at the meaning level, which makes their cross-lingual analyses informative and useful in different NLP tasks. Table 1 shows the derivational patterns associated with each morphosemantic relation in Bulgarian (BG) and Romanian (RO), with the number of occurrences found in the respective database in brackets. The verbal suffixes are written in parentheses and preceded by a dash: for example, – (-a) among the Romanian suffixes. "Total" refers to the total number of affixes/derivational patterns with the respective semantic label.

Semantic label	**BG Affixes (number of occurrences)**	**RO Affixes (number of occurrences)**
Agent	-tel (169), -ach/-yach (128), -(n)ik (87), -sht (83), -or/-yor (44), -tor (-tor/-tyor/-ator/-itor/-ityor) (42), -ets (33), – (-iram/-iziram) (22), -ar/-yar (19), – (-stvam) (19), -ist (14), -ant/-ent (13), -ne (13), -dzhiya/-chiya (11), -er/-ier/-ăr (11), conversion (10), -in (7), -l (7), -chik (7), -n (6), -nie (6), -ovach (6), -ko (5), -ak (2), -at (4), -tsiya (-tsiya/-atsiya/-itsiya/-ziya/-siya) (5), -stvo (4), (-uvam) (4), -entsiya (2), -ir (2), -itsa (2), -telka (2), -lo (2), -ba (1), -ezh (1), -ek (1), -ik (1), -ka (1), -lyo (1), -(n)itsa (1), -m (1), – -ot (1), -t (1), -yaga (1), -yay (1) **Total 45**	-(ă)tor (176), -t/-s (31), – (-i) (20), – (-a) (10), – (-iza) (4), -re (7), – (în- + -a) (1), -ar (5), -ant (7), -or (5), -ier (3), -t + -el (1), -[ăi]cios (3), -ist (2), -ăros (1), -u (1), -ici (1), -ăreţ (5), -aş (5), -(ă)toare (4), -ură (1), -(ă)tor+-easă (2), -aci (1), -angiu (1), -nic (1), – (în- + -i) (1), – (-ui) (1), -aţie (2), -aţiune (1), -aş (1) **Total 31**
Body-part	conversion (4), -ka (1), -nie (1) **Total 3**	(-a) (1) **Total 1**
By-means-of	-ne (64), -nie (53), conversion (45), -ka (33), – (-iram, -iziram) (29), -tsiya (-tsiya/-atsiya/-itsiya/-ziya/-siya) (28), -(n)ost/-est (8), -tor (-tor/-tyor/-ator/-itor/-ityor) (8), -ina (8), – (-vam/-avam/uvam) (7), -tel (6), -lo (5), -no (5), -ets (3), -ie (3), -iya (3), -lka (3), -ovka (3), – (-(n)icha) (3), – (-na) (3), -stvo (3), -at (2), -entsiya (2), -izăm (2), -achka (1), -ba (1), -er/-ier/-ăr (1), -or/ -yor (1), – (-(y)asam/-(y)osam) (1), – (-stvam) (1) **Total 30**	-re (98), – (-a) (37), – (-i) (12), -(ă)tor (10), – (-ifica) (6), -aţie (8), -t/-s (6), -(ă)tură (6), -eală (4), -ant (2), -or (3), -(ă)toare (4), – (-î) (2), – (în- + -a) (6), -ăciune (2), -tor + -ie (1), -t + -ie (1), – (-ui) (3), – (-iza) (3), -ment (3), -ie (2), -ală (1), – (-ia) (2), – (-(ur)i) (1) **Total 25**
Destination	– (-ifitsiram) (2) **Total 1**	– (-a) (1), -ar (1), -ant (1) **Total 3**

Continued

Table 1 – Continued

Semantic label	BG Affixes (number of occurrences)	RO Affixes (number of occurrences)
Event	-ne (2372), conversion (418), -nie (353), -tsiya (-tsiya/-atsiya/-itsiya/-ziya/-siya) (325), -ka (83), – (-iram) (66), – (-vam/-avam/uvam) (51), -ie (48), -stvo (42), -(n)ost/-est (29), -iya (23), -ăk (21), -ezh (18), -ba (15), -(n)itsa (17), – (-na) (9), -ek (8), -ovka (7), – (-(y)asam/-(y)osam) (7), -n (6), -azh (5), -tva (5), -ina (5), -at (4), -nya (4), -entsiya (4), -(n)ik (3), -zăm (3), -ar (3), -da (2), -na (2), -ot (2), -s (2), -t (2), -al (1), -ans (1), -et (1), -ishte (1), -ota (1), -otevitsa (1), -tba (1), -tel (1), -ulka (1), -h (1) **Total 45**	-re (1174), -t/-s (112), -aţie (110), –(-a) (97), -(ă)tură (48), -(e)ală (44), – (-i) (20), -ment (9), -or (1), -ător/ătoare (4), -ăt (1), -et (5), – (-î) (1), -ă (1), – (-ări) (1), -aci (1), -(ă)ciune (3), – (-ui) (6), -ie (8), – (-ifica) (1), -aţiune (1), -iş (1), -uş (1), – (-ia) (1), – (-ua) (1), – (-âi) (1), -e (4), -aj (3), -[a/e/i]nţă (4), – (-iza) (3), -iune (1), – (-(i)ona) (5), -erie (1), – (-ăi) (1) **Total 34**
Instrument	conversion (26), -tor (-tor/ -tyor/ -ator/ -itor/ -ityor) (15), – (-iram/-iziram) (15), -tel (13), -er/-ier/-ăr (6), -ka (9), -ie (4), -or/ -yor (3), -(n)ik (2), -ach/-yach (2), -lka (2), -l (2), -nie (2), -(n)itsa (1), -ik (1), -la (1), -nya (1), -ovach (1) **Total 18**	-(ă)tor (21), – (-a) (10), (-iza) (2), -re (1), -t/-s (1), – (-i) (2), -(ă/i)toare (6), -tură (1), – (-î) (1), – (în- + -a) (1) , pre- + -ător (1), – (-ui) (2), – (-ia) (1) **Total 13**
Location	conversion (25), – (-iram) (9), – (-vam/-avam/uvam) (7), -ishte (6), -iya (2), -ne (2), -(n)itsa (1), -ing (1), -ka (1) **Total 9**	-re (6), – (-iza) (1), – (în- + -i) (1), -ment (1) **Total 4**
Material	-tel (17), -tor (-tor/ -tyor/ -ator/ -itor/ -ityor) (12), – (iram/-iziram) (12), -tsiya (-tsiya/-atsiya/-itsiya/-ziya/-siya) (6), – (-ifitsiram) (5), conversion (2), -ka (2), -lka (2), -nie (2), -ant/-ent (2), -at (2), -(n)ik (1), -atoar (1), -ezh (1), -er/-ier/-ăr (1), -ivo (1), – (-osam) (1) **Total 17**	– (-iza) (2), – (-a) (5), -ant (1), -(ă)tor (3), – (-i) (1), -tură (1), – (-ona) (1) **Total 7**
Property	-nie (28), -(n)ost/-est (24), conversion (23), -ne (15), -tsiya (-tsiya/-atsiya/-itsiya/-ziya/-siya) (4), -ie (4), -ba (3), -entnost (2), -entsiya (2), -iya (2), – (-iram) (4), – (-(u/o)vam) (4), -ka (1), -ota (1), -stvo (1) **Total 15**	-re (32), – (-a) (6), – (-i) (2), – (în- + -a) (1), – (-ui) (2), -ment (2) **Total 6**
Result	conversion (126), – (-iram/-iziram) (53), -ne (46), -tsiya (-tsiya/-atsiya/-itsiya/-ziya/-siya) (40), -nie (32), -ka (19), -(n)ost/-est (18), -at (10), -no (9), – (-asam/yasam) (4), -ets (4), -iya (4), -n (4), -ie (3), -ina (3), -tor (-tor/ -tyor/ -ator/ -itor/ -ityor) (3), -ezh (2), -(n)itsa (2), – (-vam) (2), -azh (1), -al (1), -ar (1), -ba (1), -e (1), -ek (1), -eriya (1), -iy (1), -ing (1), -l (1), -ma (1), -ment (1), -ovka (1), -ura (1), -ăk (1) **Total 34**	-re (83), – (-a) (26), – (-iza) (13), – (-ifica) (15), -t/-s (11), – (-i) (6), -eală (4), – (în- + -a) (2), -tură (10), -et (2), – (în- + -i) (1), – (-ui) (2), – (-ifia) (2), -aţie (1), -ere (1), – (-ona) (1), -ment (3), – (-ua) (1) **Total 18**
State	-ne (68), -nie (47), conversion (37), -(n)ost/-est (30), -tsiya (-tsiya/-atsiya/-itsiya/-ziya/-siya) (15), -ie (7), -stvo (6), – (-iram/-iziram) (5), -ist (2), -iya (2), –(-osam) (2), – (-uvam) (2), -ka (2), -ika (1), -ota (1), -ăk (1) **Total 16**	-re (94), – (-a) (5), -(e)ală (3), supra- + -re (1), – (-i) (2), -t/-s (1), – (în- + -a) (1), – (în- + -i) (1), -ie (1), -ment (1), -e (1) **Total 10**

Continued

Table 1 – Continued

Semantic label	BG Affixes (number of occurrences)	RO Affixes (number of occurrences)
Undergoer	conversion (65), -ne (28), -nie (23), – (-iram/-iziram) (18), -n (11), -tsiya (-tsiya/-atsiya/-itsiya/-ziya/-siya) (10), -(n)ost/-est (9), -at (9), -ka (8), -ba (5), – (-vam/-avam/uvam) (5), -(n)ik (4), -ie (4), -ina (4), -sht (4), -ach/-yach (2), -ek (2), -da (2), -m (2), – (-ifitsiram) (2), – (-(y)asam/-(y)osam/-(d)isam) (2), -ant/-ent (1), -el (1), -entsiya (1), -ivo (1), -iya (1), -ma (1), -n (1), -nya (1), -och (1), -tva (1), -tel (1), -t (1), -ăk (1), – (-stvam) (1) **Total 35**	-re (27), – (-a) (23), -t/-s (15), -ant (1), – (-i) (2), – (-ui) (4), – (-iza) (2), – (-ifica) (1), -aţie (1), -ment (1) **Total 10**
Uses	– (-iram/-iziram) (45), -ne (25), conversion (22), -nie (20), -tsiya (-tsiya/-atsiya/-itsiya/-ziya/-siya) (13), -ka (10), -lo (7), -stvo (7), -iya (5), – (-vam/-avam/uvam) (5), -at (4), -ie (3), – (-ifitsiram) (3), -et (2), -iy (2), -ina (2), -lka (2), -ovka (2), -ura(2), – (-(y)asam/-(y)osam/-(d)isam) (4), -(n)ost/-est (1), -ant/-ent (1), -ezh (1), -er/-ier/-ăr (1), -tel (1), -tor (-tor/-tyor/-ator/-itor/-ityor) (1) **Total 26**	-re (24), – (-a) (23), – (-iza) (4), -t/-s (1), -eală (1), -tor (1), – (-i) (2), -tură (1), – (în- + -a) (1), – (în- + -i) (1), – (-ui) (2), – (-ifica) (1), -aţie (1), -ment (3) **Total 14**
Vehicle	-ach/-yach (1), -er/-ier/-ăr (1), -ovach (1) **Total 3**	– (-a) (1), -or (1), -er (1) **Total 3**

Table 1: Derivational affixes in Bulgarian and Romanian wordnets associated with semantic labels

As a consequence of the fact that prefixes normally do not change the part of speech of the stems they are attached to and that we focus on noun-verb pairs, the affixes discussed here are almost exclusively suffixes (with the exception of parasynthetic derivational patterns).

The statistics (see Table 1) show that more affixes are found in the Bulgarian data – 252 noun suffixes (in each of their senses), 38 verbal ones, and 12 cases of conversion. In Romanian there are 91 noun suffixes and 45 verbal ones (plus 26 cases of verbal derivation that are equivalent to conversion in Bulgarian). Besides the quantitative difference between the pairs subject to analysis here (4,590 pairs in Bulgarian and 2,429 pairs in Romanian), the difference in the number of the suffix senses can also be explained in terms of the specifics of the derivational morphology of the two languages. As a Slavic language, Bulgarian has a rich inventory of noun suffixes that outnumber considerably the corresponding Romanian suffixes: compare the three most productive Bulgarian suffixes with a primary agentive reading (*-tel*, *-ach/-yach*, and *-(n)ik*) vs. one such suffix in Romanian (*-(ă)tor*). Additionally, Bulgarian has adopted many Romance suffixes through the active borrowing of Romance words, so that the Romanian *-(ă)tor* has an exact equivalent in Bulgarian, the suffix *-tor*. The verbal aspect in Bulgarian is another linguistic reason for the greater diversity of patterns as both imperfective and perfective stems may be productive in verb–noun derivation and some noun suffixes may attach preferentially or exclusively to either an imperfective or a perfective verb stem, giving rise to different derived words; for example both *-ne*, which combines only with imperfective stems, and *-nie*, which usually selects perfective stems, correspond to *-re* in Romanian.

Verbal patterns involve the attachment or removal of one verbal suffix. The noun suffix representative for the respective relation may remain 'hidden' as it is present both in the noun and the verb: see the Romanian pair *călători/călător* ("to travel/traveller"), which involves the attachment of the verbal suffix *-i* to the base noun (the agentive suffix *-tor* is considered part of the base noun). There are examples, such as *ucenici/ucenic* ("to apprentice/apprentice") or *grădinări/grădinar* ("to garden/gardener") in which the verbs are formed from suffixed Slavic (Bulgarian) loan nouns following a Romanian verbal pattern.

Despite the difference in the number of suffixes, the two languages show similarity in the derivational productivity of the morphosemantic relations. The relations with the highest diversity of derivational patterns are Agent (expressed by 45 derivational patterns in Bulgarian and 31 in Romanian) and Event (45 patterns in Bulgarian, 37 in Romanian). They also cumulate the greatest number of occurrences.

Agentive suffixes in Bulgarian are both domestic and loaned, with prevalence of the former, such as *-tel*, *-ach/-yach*, *-nik*, which are also the most productive. Another productive pattern is represented by the suffix for the present active participle (*-sht*) substantivised to express Agent. Another frequent derivational pattern is formed by a noun, usually loaned, that cannot be morphologically segmented and is verbified by adding the suffix *-iram/-iziram* (*tip* ("type") > *tipiziram* ("to type, to typecast"). To express Agent, Bulgarian uses also suffixes loaned from the Romance languages, mainly French (*-tyor*), from Turkish (*-dzhiya/-chiya*), Russian (*-chik*), and other languages. The Romanian agentive affixes are of various origins (Latin, Slavic, Romance, Hungarian and domestic), with a prevalence of Latin and Romance affixes, among which the most productive one can also be found: *-tor*. Quite frequently the participle (or even the gerund) is used to denote an agent. In both languages agentive nouns are usually derived from verbs, but there is a considerable number of instances (one fourth of the total number of affixes for Romanian) where verbs are derived from agentive nouns.

In Bulgarian the relation Event is most typically expressed by the suffix *-ne*, whose occurrences outnumber the sum of the occurrences of all the other suffixes. The suffix *-nie*, traditionally associated with a resultative meaning, was found to be very productive in expressing Event, too. Conversion is used to form 418 derivational verb–noun pairs in Bulgarian. Event is expressed mostly by domestic suffixes, except for *-tsiya* and its variants, which is among the productive patterns. The suffixes *-ie*, *-stvo*, *-(n)ost/-est*, whose typical meaning is associated with Event, are ranked among patterns with medium productivity. The prevalent eventive suffixes in Romanian are of Latin or Romance origin. The most productive one is the old infinitive formant *-re* reinterpreted as a suffix for deverbal nouns. In Romanian the participle used as a noun is also a productive means of denoting events. The cases where the verb is derived from the noun denoting Event are quite numerous (one third of the affixes).

Other relations expressed by a variety of derivational patterns in Bulgarian and Romanian are By-means-of, Instrument, Result, State, Undergoer, and Uses. Disparity between the two languages is observed in the relations Material and Property. The former is expressed by 17 derivational patterns in Bulgarian vs. 7 in Romanian, and the latter by 15 patterns in Bulgarian vs. 6 in Romanian.

The relations represented by a small number of occurrences and derivational patterns are Location, Destination, Body-part, and Vehicle. Due to the lack of evidence, they are of little importance for the general analysis, yet some observations can be made. In Bulgarian Vehicle is expressed by the loaned suffix *-er/-ier/-ăr*, which is the equivalent of *-er* in Romanian, and the Slavic suffix *-ach/-yach* and its variant *-ovach*. Two typical Location suffixes occur in the Bulgarian data as well – *-ishte* and *-nitsa*.

6. Derivational Patterns. The Nature and Properties of Derivational Relations

In this study we analyse mainly noun suffixes as they are the bearers of the semantics of the relations under discussion. Verb suffixes in both languages have mostly grammatical functions. In Bulgarian, typically, they either imperfectivise a perfective verb, or perfectivise an imperfective verb, or are used to derive a verb from a word pertaining to a different part of speech. The verb suffixes occurring in the Romanian data set always create verbs from other parts of speech. Besides the established noun suffixes, we look at certain participial and adjectival suffixes as participles and adjectives can be substantivised.

The data below are based on noun and verb synsets with equivalents in the PWN. Therefore, the results are not conclusive either with respect to the language system or to the parts of speech involved.

In the data we have analysed (Table 2) there is a large number of monosemous affixes: 32 for Bulgarian and 45 for Romanian, associated mostly with the labels Event (18 in Romanian: *-erie*, *-anţă*, *-aj*, etc., and 7 in Bulgarian: *-ulka*, *-tba*, *-otevitsa*, etc.), and Agent (18 in Romanian: *-aci*, *-angiu*, *-nic*, etc., and 13 in Bulgarian: *-chik*, *-ar/-yar*, *-chiya/-dhziya*, *-in*, *-lyo*, etc.). Several other relations – Material, Result, Undergoer, Property, State and Instrument in Bulgarian, and By-means-of, Instrument, State, Result and Vehicle in Romanian are represented by one or a couple of unambiguous suffixes.

Number of relations	1	2	3	4	5	6	7	8	9	10	11	12	13
Number of Bg suffixes	32	19	6	7	6	4	3	2	2	-	2	1	1
Number of Ro suffixes	45	7	5	2	4	4	1	3	1	2	-	-	-

Table 2: Number of Relations across suffixes

Polysemous suffixes are usually associated with clusters of relations with one of them being the default reading (estimated in terms of number of instances). For example, suffixes which primarily express the relation Agent can also express relations denoting inanimate agents and causes, such as Instrument, Material, By-means-of. The relations Vehicle and Body-part typically should also be included in this group, but the number of instances is too small so we defer judgement. The relation Uses, which denotes a function or a purpose, is also often expressed by agentive suffixes. In certain cases the same suffix denotes both Agent and Undergoer depending on whether the verb is unergative or unaccusative (Fellbaum et al. 2009), e.g., *demonstrant* ("demonstrator"), and *mutant* ("mutant"), respectively.

Another large part of the suffixes typically express relations such as Event, Result and / or other relations involving the process or result of an action, state or another kind of situation, such as State and Property. A relatively frequent relation associated with this type of suffixes is Undergoer, which in this case denotes patients. The relation Uses and By-means-of can also be expressed by event-like suffixes.

In Table 3 below, we show Bulgarian and Romanian suffixes primarily associated with the relations Agent and Event and their other senses expressed by the respective relations.

Language	**Suffix**	**Default semantic value**	**Other semantic values**
Bg	-tel	Agent (169)	Material (17), Instrument (13), By-means-of (6), Undergoer (1), Uses (1)
Bg	-tor	Agent (42)	Instrument (15), Material (12), By-means-of (8), Result (3), Uses (1)
Bg	-tsiya	Event (325)	Result (40), By-means-of (28), State (15), Uses (13), Undergoer (10), Material (6), Agent (5), Property (4)
Bg	-ne	Event (2372)	State (68), By-means-of (64), Result (46), Undergoer (28), Uses (25), Property (15), Agent (13), Location (2)
Bg	-nie	Event (353)	By-means-of (53), State (47), Result (32), Property (28), Undergoer (23), Uses (20), Agent (6), Instrument (2), Material (2), Body-part (1)
Ro	-tor	Agent (180)	Instrument (27), By-means-of (14), Event (3), Material (3), Uses (1)
Ro	-re	Event (1173)	By-means-of (98), State (94), Result (84), Property (32), Undergoer (27), Uses (24), Agent (7), Location (6), Instrument (1)
Ro	-ţi(un)e	Event (111)	By-means-of (8), Agent (3), Undergoer (1), Result (1), Uses (1)
Ro	-t/s	Event (112)	Agent (30), Undergoer (15), Result (11), By-means-of (6), Instrument (1), State (1), Uses (1)
Ro	-(ă)tură	Event (48)	Result (10), By-means-of (6), Instrument (1), Material (1), Uses (1)

Table 3: Semantic labels for corresponding productive suffixes

The Agent reading is the default one for the Bulgarian suffixes *-tel*, *-tor*, *-(n)ik*, *-ant*/*ent*, *-ar*/*-yar*, *-ach*/*-yach*, *-er*/*-ier*/*-ăr*, *-ets*, *-ist*, *-or*/*-yor*, *-dzhiya*, *-ak*, and the Romanian suffixes *-tor*, *-ant*, *-ar*, *-or*. The Bulgarian suffixes *-ne*, *-nie*, *-ba*, *-ezh*, *-ie*, *-iya*, *-ka*, *-stvo*, *-tsiya*, *-ăk*, and the Romanian *-re*, *-aţie*, *-tură*, *-eală*, *-t*/*s* are most often associated with Event.

Event-type suffixes may adopt agentive meanings through metaphorical extension of an activity to a

body of people who are responsible for or carry out the activity, e.g., *-stvo*, *-nie* and *-tsiya* in *răkovodstvo*, *upravlenie*, and *administratsiya* (all meaning "administration" in the sense of a governing body). The Romanian *administraţie* ("administration") and *organizaţie* ("organization") are similar examples. The reverse process – the extension of the meaning of agentive suffixes to event-type meanings – is rarely observed in Bulgarian (*-ik* in *plesnik* "a smack, smacking").

A productive pattern in Bulgarian involves participle substantivisation. Active participles usually express Agent (*pregovaryasht* ("negotiator"), *otselyal* ("survivor")), while passive participles are mostly associated with Undergoer (*intervyuiran* ("interviewee")), Result (*izgoreno* ("a burn, burn wound")), Event (*razlyano* ("a spill, spilled liquid") and By-means-of (*zakărpeno* ("mend, patch, darn")). The agentive reading of the passive participle is quite untypical for Bulgarian and the 6 instances registered in our data are due either to reflexive/middle interpretation, as in *pristrastya se*/*pristrasten* ("to addict/an addict"), *uhilya se*/*uhilen* ("grin/grinner"), or to incorrect relation assignment in PWN (Agent instead of Undergoer), as in *zapodozra*/*zapodozryan* ("to suspect/a suspect"). Since Romanian does not distinguish between active and passive participles, the participle (ending in *-t*/*s*) descending from the Latin perfect passive participle has assumed functions typical for both the active and the passive participles in Bulgarian. This is reflected in the morphosemantic relations denoted by Romanian participles, which to the exception of several single instances cover the same relations as in Bulgarian: Event (*plagiat* ("plagiarization")), Agent (*conjurat* ("conspirator"), Undergoer (*intervievat* ("interviewee"), Result (*bubuit* ("thunder")) and By-means-of (*certificat* ("a certificate")).

Table 3 shows the distribution of several suffixes related by origin or function in the two languages. For example, the deverbal suffixes *-ne* in Bulgarian and *-re* in Romanian are functional equivalents and express very similar sets of morphosemantic relations. Their distribution in terms of frequency across relations differ, with *-ne* showing a stronger preference for the Event reading than *-re*. The Event-type suffix *-tsiya* in Bulgarian, which has common origin and meaning with *-ţi(un)e* in Romanian, expresses a broader range of relations and has more even distribution across relations than its Romanian counterpart. The Latin/Romance agentive suffix *-tor* has developed identical meanings in Bulgarian and Romanian. The only differences that we found are 3 instances of Result in Bulgarian expressed by the literals *emulgator* and *emulsifikator* ("emulsifier") whose Romanian counterpart is formed by the suffix *-ant* (*emulsifiant*), and 3 instances of Event in Romanian.

7. Applications and Further Work

We plan to expand our work by further identifying derivationally related literals and semantically related synsets that have not been discovered so far due to imperfections in the recognition algorithms or because the derivationally related pairs are not morphologically related in English.

The benefit of adding new relations is two-fold – it will enable us to increase the connectivity of the wordnet synsets on the one hand, and to establish procedures for semi-automatic expansion with new synsets, on the other. A possible source of new relations that is worth exploring are other wordnets that have implemented (possibly other sets of) morphosemantic relations. The ones implemented by us in the Bulgarian and the Romanian WordNet can also be transferred to wordnets for other languages.

In the context of automatic labelling of morphosemantic relations this study can be helpful in the task of disambiguating the suffixes of words from the same synset (in a monolingual setting) or from corresponding synsets (in a multilingual setting) on the basis of their senses and semantic restrictions.

From the applications perspective, marking morphosemantic relations explicitly in individual and aligned wordnets can prove useful in text processing and information retrieval both in a monolingual and in a multilingual context. For instance, Barbu Mititelu (2013) showed how marking derivational relations in RoWordNet can help improve a task of Question Answering that makes use of lexical links. Extending that experiment and imagining a cross-language Question Answering system, the resource created by us can help identify and subsequently transfer relations between words that are morphosemantically unrelated in one language, but are related in another.

Acknowledgements

The present paper was prepared within the project *Integrating New Practices and Knowledge in Undergraduate and Graduate Courses in Computational Linguistics* (BG051PO001-3.3.06-0022) implemented with the financial support of the Human Resources Development Operational Programme 2007 - 2013 co-financed by the European Social Fund of the European Union. The authors take full responsibility for the content of the present paper.

References

Barbu Mititelu, V. (2012). Adding Morpho-semantic Relations to the Romanian Wordnet. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2596–2601.

Barbu Mititelu, V. (2013). Increasing the Effectiveness of the Romanian Wordnet in NLP Applications. *Computer Science Journal of Moldova*, 21(3):320–331.

Bilgin, O., Cetinoglu, O., and Oflazer, K. (2004). Morphosemantic Relations in and across Wordnets – A Study Based on Turkish. In *Proceedings of the Second Global Wordnet Conference (GWC 2004)*, pages 60–66.

Dimitrova, T., Tarpomanova, E., and Rizov, B. (2014). Coping with Derivation in the Bulgarian WordNet. In *Proceedings of the Seventh Global Wordnet Conference (GWC 2014)*, pages 109–117.

Fellbaum, C., Osherson, A., and Clark, P. (2009). Putting Semantics into WordNet's "Morphosemantic" Links. In *Proceedings of the Third Language and Technology Conference, Poznan, Poland. [Reprinted in: Responding to Information Society Challenges: New Advances in Human Language Technologies. Springer Lecture Notes in Informatics]*, volume 5603, pages 350–358.

Koeva, S., Krstev, C., and Vitas, D. (2008). Morpho-semantic Relations in Wordnet – A Case Study for Two Slavic Languages. In *Proceedings of the Fourth Global WordNet Conference (GWC 2008)*, pages 239–254.

Koeva, S. (2008). Derivational and Morphosemantic Relations in Bulgarian Wordnet. *Intelligent Information Systems*, pages 359–368.

Pala, K. and Hlaváčková, D. (2007). Derivational Relations in Czech Wordnet. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, pages 75–81.

Piasecki, M., Szpakowicz, S., and Broda, B. (2009). *A Wordnet from the Ground up*. Wroclaw: Oficyna Wydawnicza Politechniki Wroclawskiej.

Stoyanova, I., Koeva, S., and Leseva, S. (2013). WordNet-based Cross-language Identification of Semantic Relations. In *Proceedings of the 4th Biennal International Workshop on Balto-Slavonic Natural Language Processing*, pages 119–128.

Semi-automatic Detection of Multiword Expressions in the Slovak Dependency Treebank

Daniela Majchráková,* **Ondřej Dušek,‡** **Jan Hajič,‡** **Agáta Karčová*** and **Radovan Garabík***
*Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava
‡Charles Univ. in Prague, Fac. Math. & Phys., Inst. of Formal and Applied Linguistics
danam@korpus.sk, {odusek,hajic}@ufal.mff.cuni.cz,
agatak@korpus.sk, garabik@kassiopeia.juls.savba.sk

Abstract

We describe a method for semi-automatic extraction of Slovak multiword expressions (MWEs) from a dependency treebank. The process uses an automatic conversion from dependency syntactic trees to deep syntax and automatic tagging of verbal argument nodes based on a valency dictionary. Both the valency dictionary and the treebank conversion were adapted from the corresponding Czech versions; the automatically translated valency dictionary has been manually proofread and corrected. There are two main achievements – a valency dictionary of Slovak MWEs with direct links to corresponding expressions in the Czech dictionary, PDT-Vallex, and a method of extraction of MWEs from the Slovak Dependency Treebank. The extraction reached very high precision but lower recall in a manual evaluation. This is a work in progress, the overall goal of which is twofold: to create a Slovak language valency dictionary paralleling the Czech one, with bilingual links; and to use the extracted verbal frames in a collocation dictionary of Slovak verbs.

1. Introduction

This work is primarily aimed at building a Slovak valency lexicon interlinked with a dependency treebank, and in this paper we focus on multiword expressions (MWEs). The prospective valency lexicon is inspired by the Czech PDT-Vallex, a lexicon based on the Prague Dependency Treebank (PDT). We exploit here the fact that Czech and Slovak are very closely related, mutually intelligible languages that show a direct 1:1 relation in a greater part of their grammatical and lexical inventory, including MWEs.

Following the definitions of MWEs for PDT annotation, here we understand by MWEs those lexical combinations "that contain some idiosyncratic element that differentiates them from normal expressions" (Bejček et al., 2012: 234). There are two types of MWEs we focused on: *light verb constructions* and *verbal phrasemes*. The valency frames of both groups are marked with special semantic labels (functors) in the deep-syntax/semantic annotation of the PDT (*tectogrammatical layer*): Compound Phraseme (CPHR) for light verb phrases and Dependent Phraseme (DPHR) for phrasemes.

In the first stage of our work, PDT-Vallex was automatically translated into Slovak and valency frames for Slovak verbs were automatically created based on their Czech counterparts. Subsequently, the translations of verbs and their valency frames were manually proofread to ensure correctness, especially those related to MWEs. The result of this process is a preliminary version of the Slovak Valency Lexicon (SVL).

The second stage involves linking the SVL to the Slovak Dependency Treebank (SDT) (Šimková and Garabík, 2006). We developed an automatic procedure to convert the SVL to a deep-syntactic representation parallel to the PDT. Here we used a list of MWE candidates extracted from the SVL to automatically identify the individual occurrences of MWEs. We evaluated the precision and recall of the automatic MWE detection by manual assessment on a small part of the SDT.

The paper is structured as follows: In Section 2., we introduce the SDT. We describe the creation of the SVL in Section 3., contrasting MWE usage in Czech and Slovak. Section 4. details our auto-

matic procedure for the conversion of the SDT to a deep-syntactic representation. Section 5. presents the evaluation of the automatic MWE detection in the treebank and Section 6. concludes the paper.

2. Slovak Dependency Treebank

The Slovak Dependency Treebank (SDT) (Šimková and Garabík, 2006) is a manually annotated dependency treebank of contemporary written Slovak. The annotation follows the methodology of the Prague Dependency Treebank (PDT) (Hajič et al., 1999). However, the SDT contains only surface dependency (*analytical*) trees, it does not include the deep-syntax/semantic (*tectogrammatical*) layer (see Section 4.), where valency and MWEs are annotated in the PDT.

The SDT contains 1,159,462 tokens in 71,672 sentences, 50,313 sentences (846,967 tokens) out of which were annotated by two independent annotators. Most texts in the treebank include manual morphological annotation (lemmas and morphological tags) based on the Slovak National Corpus tagset (Garabík and Šimková, 2012).[1]

The selection of the texts aims at a somewhat balanced corpus – there are professional texts (scientific articles, theses), fiction, and journalistic texts.

3. Building the Slovak Valency Lexicon: PDT-Vallex Translation

The PDT-Vallex (Hajič et al., 2003; Urešová, 2011a; Urešová, 2011b) is a valency lexicon interlinked with the Prague Dependency Treebank. It consists of over 11 thousand valency frames for more than 7,000 verbs. The verbs, their senses, and their valency frames are collected from sentences in the PDT.

Although Czech and Slovak are close languages, the translation of PDT-Vallex was not straightforward. The automatic translation consists of simple lexical substitution of verbs and their complementations. We then manually checked all entries relevant to the MWE extraction (261 light verbs/CPHR nodes and 480 phrasemes/DPHR nodes). The manual proofreading of the automatic translation and contrastive analysis of equivalent Czech and Slovak MWEs proved that given the closeness of both languages, there was a huge overlap of MWEs in Czech and Slovak. However, we found several cases where identical semantic content was represented by very different lexical and/or syntactic means, mainly in phrasemes.

For the purpose of obtaining the list of Slovak MWEs for automatic annotation, we mention only briefly some similarities and differences between Czech and Slovak equivalent expressions we encountered in the translation of the valency dictionary.

3.1. Similarities of Czech and Slovak MWEs

The similarities of Czech and Slovak CPHR and DPHR structures can be summarized as follows:

- Most verbs and nouns from PDT-Vallex expressing the same semantic content are etymological cognates – e.g., *podat/podať*,[2] ("hand over"), *obracet/obracať* ("turn over"), *dojem/dojem* ("impression"), *zřetel/zreteľ* ("consideration").
- Slovak and Czech verbal aspects are identical in almost all cases[3] and reflexive verbs in Czech are also reflexive in Slovak – e.g., *dát se/dať sa* ("be possible"), *udělat si/urobiť si* ("make").
- The structure of light verbs and phrasemes is identical in both languages, with just a few exceptions.

3.2. Differences between Czech and Slovak MWE Equivalents

The differences between Czech and Slovak MWEs include grammatical and/or lexical distinctions, which are reflected in the component structure of some MWEs.

[1] There are some short texts in the treebank which were tagged automatically, but these were excluded for the purpose of this article.

[2] In these examples, the Czech word is displayed first, followed by the Slovak equivalent separated by a slash.

[3] Both Czech and Slovak verbs form aspectual pairs for incompletive/processual and completive aspect, e.g., *házat'/hodit'* ("be throwing"/"throw").

Grammatical differences. According to the grammatical features, some MWE equivalents vary in the noun case; this is usually connected to the absence of or the preference for a different preposition: *přicházet v úvahu/prichádzať do úvahy* ("come into consideration"; accusative vs. genitive), *zažít na vlastní kůži/prežiť na vlastnej koži* ("experience on one's own"; accusative vs. locative).

As PDT-Vallex consists only of MWEs occuring in PDT, some of the phrases were not covered by verbs in both verbal aspects. In some cases, the aspect variant included in PDT-Vallex is less frequent or outright rare in the Slovak equivalent. In order to obtain better coverage, we decided to use both verb aspects in the Slovak translation: *zavádět řeč na jiné téma → zavádzať reč na inú tému*, *zaviesť reč na inú tému* ("steer to another topic")

Differences in lexical component. Some MWEs differ in lexical components in the use of synonymic equivalent, e.g., *vzít nohy na ramena/vziať nohy na plecia* ("run away"), *shodit pod stůl/zmietnuť zo stola* ("drop from the table"). Significant differences are present in idioms like *vyšly navrch/vyšli na povrch* ("come out"), in which the Czech adverb corresponds to Slovak noun in accusative form. There were also differences in verbal components. In some cases we preferred more frequent and neutral synonyms instead of the equivalents perceived as marked (e.g., archaic, poetic etc.) *učinit/urobiť*, *náležet/prislúchať*.

Differences in component structure of MWE equivalents. There were not many structural differences between Czech and Slovak MWEs. They can be illustrated by the following schematics (with Czech MWE structures on the left and Slovak on the right):

- Adding/removal of a grammatical component (preposition):

V + S	V+ Prep + S
zírat údivem	*civieť súdivom* ("gape in awe")

V+ Prep + S	V+ S
dát za vyučenou	*dať príučku* ("give a lesson")

- Adverbs change to a prepositional phrase (petrified in the second example, cannot be split into separate components):

V+ Adv	V+ Prep + S
vyjít navrch	*vychádzať na povrch* ("come out")

V+ Adv	V+ [Prep + S]
vycházet vstříc	*vychádzať vústrety* ("to be acommodating")

- Absence of a Slovak equivalent for the Czech particle *co*:

V+ Part + [Prep + S]	V+ [Prep + S]
mít co do činění	*mať do činenia* ("have something to do with")

- Partial disagreement arising from the nature of the Slovak particle *treba* ("is needed"). The difference is only apparent in the present tense where the particle *treba* does not require the auxiliary verb *byť*[4]; this is different from past and future tense:

V+ Adv	Adv
je třeba	*treba* ("is needed"; present tense)

[4]The present tense also occurs with the auxiliary verb *byť (je treba)*; this is, however, considered colloquial. We still included this variant in the dictionary to increase coverage.

- the phrase *bůh vám zaplať/pánboh zaplať* ("God bless you") has a different structure and lexical components in Slovak:

S+ Pron + V	S + V
bůh vám zaplať	*pánboh zaplať*

In some cases, the translation of a Czech MWE is not possible at all; either it contains lexical lacunae or the phrase as a whole is not used in Slovak. Examples of light verb constructions without an equivalent in Slovak are: *dát/dávat preferenci* ("give preference"), examples of phrasemes are: *vydat všanc* ("submit to risk"), *vzít roha* ("run away"), *být na štíru* ("have a problem with").

4. Automatic Tectogrammatical Annotation

To link SDT to a valency lexicon paralleling PDT-Vallex, we created a procedure for the conversion of the SDT from surface dependency trees to tectogrammatial trees, a deep-syntactic/semantic representation based on the Functional Generative Description (Sgall, 1967; Sgall et al., 1986). The tectogrammatical representation of a sentence is a dependency tree which only consists of nodes that carry lexical meaning; auxiliary words are no longer included. Each tectogrammatical node is marked with a lemma, a *functor* (semantic role label) and a set of *grammatemes*, which carry grammatical meanings, such as number, tense, or modality.

The surface dependency trees are automatically converted into tectogrammatical trees by a set of small, rule-based modules implemented within the Treex NLP framework (Popel and Žabokrtský, 2010). Since the conversion makes heavy use of morphology information and was primarily developed with the Czech positional morphological tagset (Hajič, 2004) used in PDT in mind, it also includes a morphological tagset conversion step.

4.1. Morphological Tagset Conversion

For morphological tagset conversion, we make use of the Interset framework (Zeman, 2008). This framework contains a common list of various morphological properties across languages and their values to support conversion among different tagsets. One can either use directly the morphological information stored in Interset, or convert the source morphological tag into a different framework.

We have created an Interset driver (converter) for the Slovak National Treebank morphological tagset. We use both the information stored directly in Interset and a conversion to the PDT tagset. This allows us to reuse both language-independent and Czech-specific modules in the conversion process.

4.2. From Analytical to Tectogrammatical

The Treex modules for the conversion from analytical (surface dependencies) to tectogrammatical representation (deep syntax/semantics) closely follow the modules used for a similar conversion in Czech and English within the CzEng parallel corpus (Bojar et al., 2012) and the TectoMT machine translation system (Žabokrtský et al., 2008). However, unlike in CzEng and TectoMT, we apply the conversion to manually annotated analytical trees.

The conversion consists (roughly) of the following steps:

1. Auxiliary and grammatical words, such as prepositions and auxiliary verbs, are identified in the analytical tree. A new tectogrammatical tree is built that does not contain the auxiliary words as separate nodes, but retains links to the multiple analytical nodes for a single tectogrammatical node, including all auxiliaries.

2. Coordination and apposition functors (such as CONJ, DISJ, ADVS for conjunctive, adversative, and disjunctive relation) are identified.

3. Links to auxiliaries are distributed through coordination structures, i.e., if a preposition applies to multiple coordinated nouns, tectogrammatical nodes for all nouns will have a link to its analytical node.

4. Finite clause heads, relative clause heads, and relative clause co-reference are marked.

5. Tectogrammatical lemmas are normalized. In the current implementation for Slovak, this applies to personal and possessive pronouns, which all obtain a technical lemma *#PersPron*, and to reflexive tantum verbs, where the reflexive particle *sa/si* becomes part of the lemma (e.g., *smiať_sa* for "laugh").

6. All nodes are assigned grammatemes. In the current version, all nodes obtain semantic part-of-speech (noun, adjective, verb, adverb), and semantic verbs further obtain diathesis information (active, passive, reflexive diathesis).[5]

7. Functors are assigned to all nodes. We use rules based on lexical meaning, auxiliary words linked from a given node, and part-of-speech of the lexical word to estimate its semantic function.

   This step also includes detection of multiword expressions – light verb constructions and phrasemes, which are given functors *CPHR* and *DPHR*, respectively. These are detected based on candidate lists gathered from the Slovak Valency Lexicon (SVL, see Section 3.).[6]

8. Special tectogrammatical nodes are generated for actors not expressed on the surface — pro-dropped pronominal subjects and generic actors in reflexive passive constructions, such as *Dom sa stavia* (lex. *A-house itself builds*).

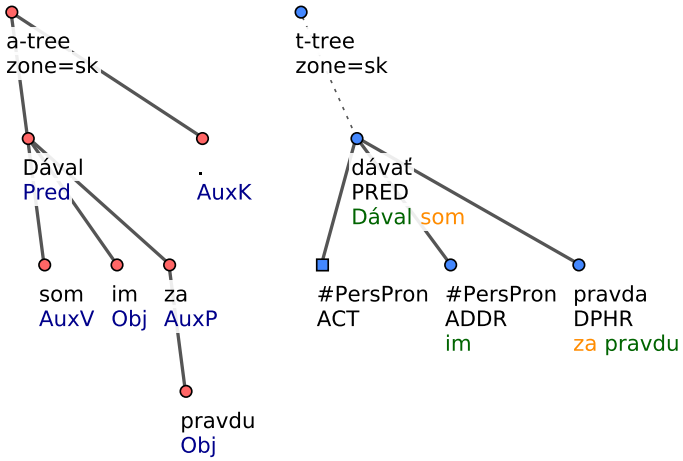


Figure 1: An original dependency tree from the Slovak Dependency Treebank (left, with dependency labels given in blue) and a tectogrammatical tree after conversion (right, with functors on the second line). The DPHR functor marks a dependent part of the phraseme *Dával som im za pravdu.* ("I agreed with them.").

Figure 1 shows a comparison of the original dependency tree with the result of the tectogrammatical conversion.

4.3. The Result of the Tectogrammatical Conversion

While the tectogrammatical layer conversion is almost equivalent to automatic tectogrammatical annotation used for English and Czech, it is missing some of the attributes present in the manual annotation of PDT:

- Generated nodes for other semantic participants than actors,

[5] Cf. Urešová and Pajas (2009) for more information on diathesis.

[6] The detection algorithm checks for the presence of all dependent parts of a MWE in the surface dependency subtree governed by its verb, then assigns MWE functors to corresponding tectogrammatical nodes. It abstracts from particular inflection forms by checking base word forms (lemmas) only. While such an abstraction may possibly result in lower precision, our experiments in Section 5. show that it is sufficient in practice.

- Full pronominal co-reference,
- Generated nodes for cases of ellipsis,
- Explicit valency frame assignment, i.e., sense disambiguation for verbs and some nouns,
- Focus-topic articulation and discourse structure.

However, even this level of annotation is suitable for linguistic inquiry and automated tasks such as machine translation, and can be used as a starting point for full manual tectogrammatic annotation.

5. Evaluation

In order to estimate the performance of the automatic MWE annotation, we randomly selected about one thousand sentences out of the tectogrammatical conversion of the Slovak Dependency Treebank,[7] where we annotated CPHR and DPHR nodes for light verbs and phrasemes manually. We then compared this sample to the result of the automatic conversion.

Table 1 shows estimates of precision and recall for three main types of text – newspaper texts, professional texts (i.e., scientific), and fiction. The ratio columns show the ratio of CPHR and DPHR nodes to the total (tectogrammatical) nodes of the sample. Given the rather small sample size, the number of these nodes is small. The precision and recall figures should therefore be considered with this in mind.

The manual proofreading of the sample of sentences showed that only 46 % of all MWEs were identified automatically. This is caused by the fact that only MWEs listed in the Slovak Valency Lexicon (SVL) are detected. As a translation of the original PDT-Vallex dictionary, which only includes MWEs present in the PDT data, SVL currently has a limited coverage of MWEs. As soon as more MWEs are added into SVL, the recall of our method will improve.

type	number CPHR	number DPHR	ratio CPHR [%]	ratio DPHR [%]	precision CPHR [%]	recall CPHR [%]	precision DPHR [%]	recall DPHR [%]
newspaper	14	15	0.36	0.39	89	53	100	33
professional	28	7	0.38	0.09	95	72	100	57
fiction	24	31	0.64	0.83	91	42	88	23
overall	66	53	0.44	0.35	93	57	94	30

Table 1: Precision and recall of automatic annotation of MWEs.

6. Conclusions and Future Work

We presented a work-in-progress report of the creation of the Slovak Valency Lexicon (SVL) interlinked with the Slovak Dependency Treebank (SDT), aimed at annotating multiword entities (MWEs).

The Slovak Valency Lexicon, created by a translation of the Czech PDT-Vallex lexicon and subsequent post-processing of multiword expression entries, is considered the first successful outcome of our experiments. It contains 10 038 verbs and 741 MWE entries (261 valency frames for light verbs and 480 frames for phrasemes).

The lexicon can be further used for the purpose of contrastive analysis of syntactic and semantic properties of Slovak and Czech. The list of multiword expressions can be used to examine syntactic patterns of multiword expressions and will be used for automatic verification of the forthcoming Lexicon of Slovak Verbal Collocations.

The other outcome of this paper is the method for automatic conversion of the SDT to a deep-syntactic/semantic representation following the annotation schema of the Prague Dependency Treebank, which is specifically aimed at annotating MWEs – light verb constructions and phrasemes – using a list

[7] Our subset preserved the genre balance described in Section 2.

of MWE candidates. Our results show that with this method we can identify MWEs with very good precision.

Our further immediate plans include work on improving MWE coverage in the SVL; in particular, extending the list of MWEs and adding further features that would help for their automatic identification in the syntactic treebank. A broader aim of our research is to create a full Slovak valency dictionary with links to the Czech PDT-Vallex lexicon and to use the extracted verbal frames in compiling a collocation dictionary of Slovak verbs.

Acknowledgements

We thank Mária Šimková and Katarína Gajdošová from the Slovak National Corpus for their work on Slovak language treebank. This paper was created within the research plan of the ICT COST Action IC1207 Parseme: Parsing and multi-word expressions. Towards linguistic precision and computational efficiency in natural language processing.

References

Bejček, E., Panevová, J., Popelka, J., Straňák, P., Ševčíková, M., Štěpánek, J., and Žabokrtský, Z. (2012). Prague Dependency Treebank 2.5 – a Revisited Version of PDT 2.0. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 231–246, Mumbai, India. Coling 2012 Organizing Committee.

Bojar, O., Žabokrtský, Z., Dušek, O., Galuščáková, P., Majliš, M., Mareček, D., Maršík, J., Novák, M., Popel, M., and Tamchyna, A. (2012). The Joy of Parallelism with CzEng 1.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC 2012)*, pages 3921–3928, Istanbul, Turkey, May. ELRA, European Language Resources Association.

Garabík, R. and Šimková, M. (2012). Slovak Morphosyntactic Tagset. *Journal of Language Modelling*, 0(1):41–63.

Hajič, J., Panevová, J., Buráňová, E., Urešová, Z., and Bémová, A. (1999). *Anotace Pražského závislostního korpusu na analytické rovině: pokyny pro anotátory*. Technical Report 28, ÚFAL MFF UK, Prague, Czech Republic.

Hajič, J., Panevová, J., Urešová, Z., Bémová, A., Kolárová, V., and Pajas, P. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, pages 57–68, Växjö, Sweden.

Hajič, J. (2004). *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Karolinum, Praha.

Popel, M. and Žabokrtský, Z. (2010). TectoMT: Modular NLP Framework. In *Proceedings of IceTAL, 7th International Conference on Natural Language Processing*, pages 293–304, Berlin - Heidelberg. Springer-Verlag.

Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Academia and Dordrecht: Reidel, Prague.

Sgall, P. (1967). *Generativní popis jazyka a česká deklinace*. Academia, Praha.

Šimková, M. and Garabík, R. (2006). Синтаксическая разметка в Словацком национальном корпусе. In *Труды международной конференции Корпусная лингвистика – 2006*, pages 389–394, Sankt-Petersburg. St. Petersburg University Press.

Urešová, Z. and Pajas, P. (2009). Diatheses in the Czech Valency Lexicon PDT-Vallex. In *Slovko 2009, NLP, Corpus Linguistics, Corpus Based Grammar Research*, pages 358–376, Brno. Tribun.

Urešová, Z. (2011a). *Valence sloves v Pražském závislostním korpusu. Studies in Computational and Theoretical Linguistics*. Ústav formální a aplikované lingvistiky, Praha.

Urešová, Z. (2011b). *Valenční slovník Pražského závislostního korpusu PDT-Vallex. Studies in Computational and Theoretical Linguistics*. Ústav formální a aplikované lingvistiky, Praha.

Žabokrtský, Z., Ptáček, J., and Pajas, P. (2008). TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 167–170, Ohio. Columbus.

Zeman, D. (2008). Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 213–218, Marrakech, Morocco.

Automatic Categorisation of Multiword Expressions and Named Entities in Bulgarian

Ivelina Stoyanova
DCL, IBL, BAS
Sofia, Bulgaria
iva@dcl.bas.bg

Abstract

This paper describes an approach for automatic categorisation of various types of multiword expressions (MWEs) with a focus on multiword named entities (MNEs), which compose a large portion of MWEs in general. The proposed algorithm is based on a refined classification of MWEs according to their idiomaticity.

While MWE categorisation can be considered as a separate and independent task, it complements the general task of MWE recognition. After outlining the method, we set up an experiment to demonstrate its performance. We use the corpus Wiki1000+ that comprises 6,311 annotated Wikipedia articles of 1,000 or more words each, amounting to 13.4 million words in total. The study also employs a large dictionary of 59,369 MWEs noun phrases (out of more than 85,000 MWEs), labelled with their respective types. The dictionary is compiled automatically and verified semi-automatically.

The research presented here is based on Bulgarian although most of the ideas, the methodology and the analysis are applicable to other Slavic and possibly other European languages.

1. Introduction

Statistical analyses show that multiword expressions (MWEs) comprise a significant part of the lexical system of a language. For instance, 24.49% of the Bulgarian WordNet and 22.5% of the Princeton WordNet 2.0 (Koeva, 2006) are MWEs. MWEs pose a complex set of problems to both theoretical linguistics and Natural Language Processing (NLP). Developing efficient methods for their automatic identification and categorisation will help improve results in Information Retrieval, Machine Translation, and other areas of Computational Linguistics.

A wide variety of approaches towards MWE recognition have been developed in recent years. Generally, they differ in the amount of linguistic information used and the particular statistical tools applied in the analysis. However, neither statistical methods nor methods heavily dependent on linguistic resources have proved successful for the general purpose of MWE recognition independently of each other, which has led to extensive exploration of hybrid methods.

Moreover, MWEs exhibit a wide variety of features and types, which additionally complicates their automatic processing. This paper presents an approach towards the automatic categorisation of MWEs following their automatic recognition. Multiword named entities (MNE) comprise a large portion of MWEs and are thus paid special attention here.

The research presented in this paper is based on Bulgarian although the methodology and analysis are largely applicable to other Slavic languages and possibly to other European languages as well.

2. Characteristics of MWEs

2.1. Main Features

The classification of MWEs we employ uses the feature *idiomaticity* in the sense of Nunberg et al. (1994), who consider this to be a chief characteristic of MWEs. It combines the degree of conventionality, understandability and compositionality of the MWE. Baldwin (2006) discusses a similar characteristic of MWEs and proposes a complex model for description of lexical units based on the following types of markedness:

- *Lexical markedness* – lexical and grammatical constraints on the realisation, such as paradigmatic constraints, e.g. *kick the bucket* but not *kick the buckets*, prosodic markedness, etc.;
- *Syntactic markedness* – syntactic irregularities in gender agreement or lack of agreement, or institutionalisation where lexemes preserve their historical characteristics regardless of the changes in the modern language, e.g. the preservation of the masculine gender of the noun *vecher* ('evening'), a feminine noun in modern Bulgarian, in the expression *Dobar vecher* ('Good evening').
- *Semantic markedness* – a relative (non-)compositionality of meaning, semantic relations (such as synonymy) with single words, e.g. *poshtenska stantsiya – postha*, both meaning 'a post office';
- *Pragmatic markedness* – in cases where the pragmatic features of the MWE components differ from those of the MWE as a whole, or the MWE is associated with a particular pragmatic reference point – consider the expression *Pusheneto zabraneno!*, literally *Smoking forbidden!* ('No smoking!') which is appropriate in certain communicative situations and not suitable in others;
- *Statistical markedness* – conventionality is reflected by high frequency of occurrence of particular collocations and markedly low or zero frequency of its synonymous counterparts, e.g. *strogo sekreten* ('strictly confidential') vs. the synonymous expression *striktno sekreten*.

Idiomaticity is a very broad concept. Here we use the term mainly with respect to the restrictions idiomaticity imposes on the morphosyntactic form, the semantics and the statistical frequency of MWEs. The degree of idiomaticity, or markedness, determines the way MWEs are treated in various NLP applications, such as, for example, Machine Translation. Compositionality represents the degree to which the complex meaning of the MWE is a combination of its components. After a MWE is formed, it enters into paradigmatic and syntagmatic relations in the lexical system. This means that in its context of use a MWE may change its compositionality and respectively – its level of idiomaticity.

For example, the phrase *poshtenska kutiya* ('post box') is formed as a regular decomposable combination where the adjective *post* (relating to a postal service) and *box* are realised with their usual lexical meanings. In recent years the phrase acquired an additional meaning – 'electronic post box, email', which is clearly idiomatic although the origins of the phrase and the relation between the components is still easily recoverable.

2.2. Classification of MWEs with respect to Idiomaticity

We adopt the general classification of MWEs presented by Baldwin et al. (2003). The authors distinguish between the following three categories: (a) non-decomposable MWEs for which a decompositional analysis of the meaning is not possible, e.g. *shepherd's purse*; (b) idiosyncratically decomposable MWEs for which some components of the phrase have a meaning not observed independently outside the MWE, e.g. *periodic table*; and (c) simple decomposable MWEs whose meaning can be decomposed to that of their constituents but nonetheless comprise a single lexical unit, e.g. *Bulgarian language*. For instance, due to institutionalisation simple decomposable MWEs often exhibit restrictions in the syntactic structure or synonym substitutions within the MWE. In these respects they differ from free phrases which are decomposable and are not considered lexical units, e.g. *important factor*.

For the purposes of some applications we may be interested simply in distinguishing between MWEs and free phrases in order to define separate methodologies for their treatment, e.g. keyword extraction,

while in other cases a more detailed categorisation may be required because the categories of MWEs differ with respect to their characteristic features and thus pose different problems, e.g. Machine Translation. On the one hand, the non-decomposable MWEs need to be defined in a dictionary so that they can be supplied with suitable translations. On the other hand, it is inefficient to add decomposable MWEs to the dictionary as their number is large and their meaning is defined as a function of their constituents. Different translation approaches may be adopted depending on the features of the different types of MWEs. Therefore, in many cases we are interested not only in recognising MWEs but also in discriminating between different categories of MWEs.

We divide simple decomposable MWEs into ten categories based on the following semantic and pragmatic factors: (1) Reference to NEs: (i) whether they contain a NE; and/or (ii) whether they constitute a NE; (2) Degree to which the connection between the components is explicit or can be restored. The classification is based on idiomaticity (Stoyanova, 2012):

(1) NEs without an (evident) connection between the elements – e.g., personal names *Ivan Petrov*. These are more often transliterated into other languages rather than translated, unless there is an established form for the NE in the target language.

(2) NEs with a meaningful element – e.g., *Stara Zagora* (literally, 'Old Zagora'), *North Korea*. The meaningful component is very often translated.

(3) Non-NEs with a vague connection between the components – e.g., *cave lion*. Most often these MWEs cannot be translated literally but have an established equivalent, e.g. *vodno konche* (literally, 'water horse') whose equivalent in English is 'dragonfly'.

(4) NEs containing meaningful components with difficult to restore connection – e.g., *Black Sea*. The approach to their rendition in other languages is mixed – some components may be translated and others transliterated, depending on how much of the linking information can be restored.

(5) NEs consisting of a descriptor and a NE, e.g. *Treaty of London*. These MWEs are usually translated, often rendered literally. Even if the translation of the NE is not fully equivalent to the original in meaning, the NE is still recognisable.

(6) Non-NEs which contain a NE as one of its components – *Down syndrome*. Similar to (5).

(7) Non-NEs with a standard, easy to restore connection between the components, e.g. *sea turtle* where the connection between the components is 'habitat' – 'turtle inhabiting the sea'. Categories 7-10 are very often translated literally since these are mostly descriptive decomposable MWEs.

(8) NEs with a standard, easy to restore connection between the components – *Association for Computational Linguistics*.

(9) Non-NEs with an explicit connection between the components – *self-retracting knife*. There is a subtle difference between categories (7) and (9) – in the latter the connection is explicit (e.g. 'retracts itself'), while in (7) it is not present in the MWE but is easy to recover (e.g. 'sea' is habitat). Same correspondence exists between categories (8) and (10). Explicit connection usually implies the presence of a verbal component – a participle or a verbal adjective or noun. The corresponding categories with explicit/easy to recover connection usually receive simillar treatment in automatic processing.

(10) NEs with an explicit connection between the components – *Center for the Treatment and Study of Anxiety*.

(*) Free collocations – *chist vazduh* ('fresh air'). Free collocations are free phrases (non-MWEs) which are statistically marked, i.e. they appear with high frequency compared to other synonymous candidates but are not linguistically (lexically, semantically or morphosyntactically) marked. Here they are included for completeness.

On the one hand, NEs are strongly institutionalised, which means that they may have an established translation different from the literal one, and the translation variants might be restricted. For example, the NE *Organizatsiya na obedinenite natsii* (literally 'Organisation of the United Nations') in Bulgarian differs from its English correspondence *United Nations*. On the other hand, MWEs which are not NEs are usually less restricted and allow certain variations.

The composition of MWEs often imposes different restrictions mainly on the subordinate components. Firstly, these are grammatical constraints – agreement between the subordinate part and the head (A N phrases). Some cases, however, require additional restrictions on the subordinate component which can further be used for the successful identification of MWEs. Prepositional phrases in MWEs usually express a class of objects but not a concrete object, for example *pasta za zabi (toothpaste)* – literally, 'paste for teeth', is a MWE, while *pasta za zabite na Ivan* ('paste for Ivan's teeth') is not a MWE, **pasta za zab* ('paste for a tooth'), **pasta za zabite* ('paste for the teeth') are unacceptable (their frequency in BNC is 0 compared to 417 occurrences of *pasta za zabi* ('toothpaste').

The modifications of decomposable MWE components are not always strictly restricted as in the other categories of MWEs. Although the MWE denotes a single concept, in some cases component modifications are allowed which leads to concept modification and a different meaning. It may result in the composition of a new lexical item – for example, *pasta za mlechni zabi* ('toothpaste for milk teeth') considered as a separate MWE, hyponym of *pasta za zabi* ('toothpaste'), or of a free phrase where the meaning of a component is concrete – for example, *torta s morkovi* ('carrot cake') → *torta s morkovite ot gradinata* ('cake with the carrots from the garden').

3. Method for Automatic Categorisation of MWEs Based on Idiomaticity

The method presented here is focused on MWE categorisation for the purposes of automatic text processing of Bulgarian. Different types of MWEs exhibit distinctive features and thus require specific treatment with regards to various applications (see section 2.2.).

The method is applied on annotated Bulgarian texts – sentence splitting, POS tagging, grammatical characteristics. The type of nouns – common or proper, has also been assigned. The method comprises the following rules:

1. Given that a MWE consists only of words recognised as proper nouns, classify it as a NE (category 1).

2. Given that a MWE consists of a proper noun and other elements and all the words begin with a capital letter, classify it as category 2.

3. Given that a MWE consists of a proper noun and other words and the first word of the MWE begins with a capital letter, classify it with the greatest probability as category 4 or 5.

4. Given that a MWE includes a proper noun and the first word of the MWE does not begin with a capital letter, classify it with the greatest probability as category 6.

5. Given that a MWE does not include a proper noun and the MWE begins with a capital letter, classify it with the greatest probability as category 8 or 10.

6. Given that a MWE does not include a proper noun and does not begin with a capital letter, classify it with the greatest probability as category 3, 7 or 9.

Figure 1 sketches the algorithm used for automatic detection of the MWE categories on the basis of the proposed rules.

More fine-grained categorisation might be achieved if we introduce some more specific rules incorporating semantic analysis such as Latent Semantic Analysis (Landauer et al., 2007), or lexical-semantic information from WordNet such as noun labels (e.g., noun.location) or semantic relations. In some cases it is sufficient to determine the group of categories the MWE belongs to, depending on the purposes of the study, and it may be inefficient to unambiguously assign a single category.

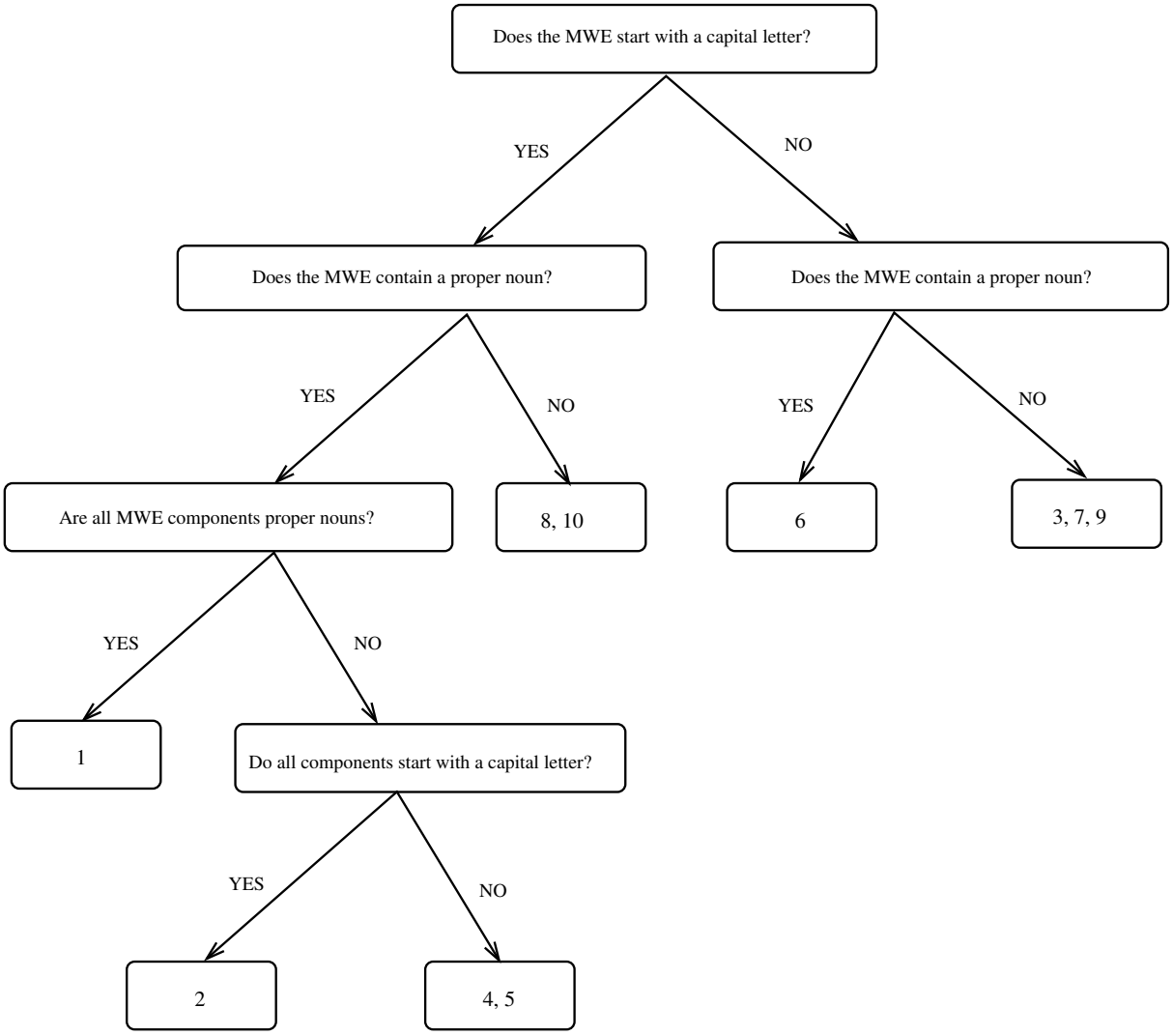


Figure 1: Algorithm for MWE category recognition.

The most problematic is the combined group of categories 3, on the one hand, and 7 and 9 on the other, since these can share the same form but have different semantic structure and thus may require different processing and analysis. Moreover, there are many MWEs which are on the boundary between categories and it can be difficult to distinguish between them.

It should be mentioned that the rules rely on some language-specific information such as the use of capital letters. However, there are many other Slavic and European languages which share these rules – capitalising first letter of names; common nouns are not capitalised (or only a limited numbers of categories are – months, days of the week); etc. The rules in this form have limited applicablity for German and other languages which capitalise all nouns, although they can be adapted and/or extended accordingly.

4. Experiment

4.1. Linguistic Resources

The experiments are based on the `Wiki1000+` corpus which comprises 6,311 Wikipedia articles, each of them containing at least 1000 words. The corpus amounts to 13.4 million words of running text distributed between 25 domains (Leseva and Stoyanova, 2014). The corpus has been supplied with linguistic annotation which includes several components – sentence segmentation, tokenisation, POS tagging and lemmatisation. The annotation is performed automatically using the set of tools of the Bulgarian

Language Processing Chain (Koeva and Genov, 2011). The POS tagger also assigns additional lexical information including the type of each noun – common or proper, and the grammatical characteristics of the word.

Syntactic type	# entries	% of all
(A) N	16,791	28.3
N N	35,314	59.5
N PP	4,424	7.5
(A) N PP	965	1.6
Other	1,875	3.1

Table 1: Syntactic types of MWEs in the dictionary (A=Adjective, N=noun, P=Preposition, PP=Prepositional phrase; brackets denote possible repetition, i.e. (A)N includes phrases of the form AN, AAN, etc.)

Idiomatic type	# entries	% of all
NE	39,982	67.3
non-NE	13,774	23.2
NE, contains-NE	3,339	5.6
non-NE, contains-NE	1,672	2.8
Unclassified	602	1.0

Table 2: Idiomatic types of MWEs in the dictionary

Additionally, noun phrases (NPs) in `Wiki1000+` have been identified using a list of possible syntactic constructions, and all MWEs have been annotated by applying a large dictionary containing over 85,000 MWEs, of which 59,369 NPs (Todorova and Stoyanova, 2014). The distribution of dictionary entries in terms of their syntactic structure is presented in Table 1, while their distribution with respect to references to NEs is shown in Table 2. Table 3 shows the result of the annotation of the different MWE categories in the corpus using the MWE dictionary.

Category	Label	#	% of all MWE
Non-decomposable	A	700	0.23
Idiosyncratically decomposable	B	3,156	1.02
Category	1	36,932	11.95
Category	2	11,248	3.64
Category	3	1,461	0.47
Category	4	1,086	0.35
Category	5	18,962	6.13
Category	6	27,373	8.86
Category	7	140,394	45.42
Category	8	16,653	5.39
Category	9	1,468	0.47
Category	10	0	0
”Free collocations”	X	49,651	16.06
Free phrases	Y	1,197,762	-

Table 3: Distribution of types of MWEs in Wiki1000+ corpus

The corpus `Wiki1000+` and the MWE dictionary are distributed as part of META-SHARE[1].

[1] `http://metashare.ibl.bas.bg/repository/search/`

4.2. Tasks

In order to observe the performance of the method, two distinct sets of tasks were defined.

1. Automatic MWE categorisation without prior MWE recognition – in this case the method for categorisation is applied on all NPs. It involves the following steps:
   - POS tagging and lemmatisation;
   - identification of NPs and syntactic filtering;
   - categorisation on all identified NPs.
2. Automatic MWE categorisation following MWE recognition – in this case categorisation is applied only on NPs identified as MWEs. It includes:
   - POS tagging and lemmatisation;
   - identification of NPs and syntactic filtering;
   - identification of MWEs;
   - categorisation of recognised MWEs and identification of certain types of NEs.

The MWE categorisation method is applied independently of MWE recognition although they generally complement each other. The MWE recognition method used in the experiments is outlined below, but it falls outside of the scope of the present work. The experiments are limited to several NP constructions: (A) N; N N; N P N; and N P (A) N.

The method for MWE identification combines collocation extraction with syntactic filtering to eliminate invalid or rare constructions. The method is described by Justeson and Katz (1995). It gives relatively good results taking into account its simplicity and the limited resources it requires (only POS annotation is needed). However, this method is best suited for extracting MWEs with adjacent components and additional processing is required to adapt it for the task of identifying non-adjacent MWEs.

In our application of the method, mutual information (MI) is adopted as the associaton measure used for deciding whether the cooccurring words form a collocation (Manning and Schutze, 1999). Other measures have also been experimented with, such as the Chi-square, Log-likelihood, Dice coefficient, but they have not proven to be empirically superior to MI for our data. It is recognised that MI, as well as most of the other statistical measures, does not work well for low frequency events so we only consider N-grams with frequency of over 10 occurrences.

In order to evaluate the performance of the MWE categorisation method (in the second set of tasks) independently of the quality of MWE recognition, we perform the method on automatically annotated and manually verified MWEs from `Wiki1000+`. However, it should be noted that in real-life applications MWE categorisation is interweaved with MWE recognition and thus the performance of the categorisation is influenced by the results of the recognition.

The two sets of tasks are evaluated independently in order to establish whether MWE categorisation can be used for MWE identification as well. The nature of the rules suggested that the method can be applied with relative independence for the identification of some categories of NEs, although it is not suitable for non-NE MWE identification in general.

4.3. Results

Table 4 presents the results for different MWE categories in terms of precision and recall. The simple rule-based approach on already recognised MWEs reaches precision of 91.51% with variation of $\pm$ 4% (except category 6, see Table 4), while on unlabelled NPs the precision varies considerably between categories and ranges between 25.11% and 81.43%. Even for the categories with best results (category 1, with the vast majority of entities being personal names) the precision without prior MWE recognition is considerably lower (81.43%) than the precision after MWE recognition (94.10%) although the recall is slightly better.

The results confirm the hypothesis that the method is unsuited for MWE recognition on its own and does not obtain satisfactory results when applied independently on general NPs.

Category	Label	Precision	Recall
Non-decomposable and Idiosyncratically decomposable	A and B	77.5	82.9
Category	1	94.1	96.5
Category	2	89.0	91.7
Category	3	0	0
Category	4	0	0
Category	5	89.4	71.0
Category	6	79.6	90.8
Category	7	90.1	87.2
Category	8	87.4	87.3
Category	9	0	0

Table 4: Results (precision and recall) for different categories after MWE recognition

Categories (3) and (9) are grouped with category (7), and category (4) is grouped with (5), they are not recognised separately, therefore they appear with zero precision and recall in Table 4. In the application of the method after MWE recognition, errors are mainly due to combination of categories, errors in tagging, or specific cases of capital letter use. For more precise results it is required to pose additional constraints on the rules or involve more detailed structural and semantic information. Moreover, improving MWE recognition methods will invariably lead to improvement in MWE categorisation.

5. Related Work

Research in the field of automatic MWE recognition and analysis in the last few decades has been clearly divided into two main trends – on the one hand, unsupervised resource-light highly efficient but less effective statistical approaches, and on the other hand, linguistically based resource-dependent but often inefficient methods. Recent research suggests that successful MWE recognition and tagging lies in the balanced hybrid approaches.

The detailed linguistically motivated characteristic of MWEs both as morpho-syntactic and semantic units, is a necessary prerequisite for successful automatic rendition. In this respect our research relies on the theoretical and applied studies focused on MWE classification by Baldwin et al. (2003), Baldwin (2004), Nunberg et al. (1994), Sag et al. (2002), among others.

Hybrid methods for MWE identification have been applied and described by Justeson and Katz (1995), Smadja (1993), Baldwin et al. (2003), Widdows (2008), Nakov (2008), Giesbrecht (2009) and many others. The specific problems of the description and automatic recognition of MWEs and NEs in Bulgarian have been discussed by Koeva (2006), Koeva (2007), Todorova (2006), Todorova and Obreshkov (2008), Leseva and Stoyanova (2008).

6. Conclusion

In conclusion, the methods described in the paper are relatively simple and do not require elaborate linguistic resources. Thus, they are suitable for morphologically rich languages, such as Bulgarian.

We need to emphasize that the results presented here are only valid for noun phrases of a limited variety of syntactic structures, and the possible generalisation of the observations over the whole group of MWEs is still to be evaluated.

However, we can conclude that the approach described here can potentially be developed into a successful methodology by considering the parameters of the particular research purpose – whether we need to simply identify MWEs, or discriminate between categories, as well as the granularity of the categorisation. It is also important to consider the characteristics of the resources as they influence highly the results, and take into account the specific features of the analysed corpora and the employed dictionaries in the anlysis and evaluation. The extensive application and testing of methods for MWE identification remains one of the major tasks in natural language processing of Bulgarian.

References

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An Empirical Model of Multiword Expression Decomposability. In *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96.

Baldwin, T. (2004). *Multiword Expressions, Advanced course at the Australasian Language Technology Summer School (ALTSS)*.

Baldwin, T. (2006). *Compositionality and Multiword Expressions: Six of One, Half a Dozen of the Other?, COLING/ACL 2006 Workshop on MWEs*.

Giesbrecht, E. (2009). In Search of Semantic Compositionality in Vector Spaces. In Rudolph, S., Dau, F., and Kuznetsov, F. O. (eds.), *ICCS 2009, LNAI*, volume 5662, page 173–184. Springer-Verlag Berlin Heidelberg.

Justeson, J. and Katz, S. (1995). Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering*, pages 9–27.

Koeva, S. and Genov, A. (2011). Bulgarian Language Processing Chain. In *Proceedings of Integration of Multilingual Resources and Tools in Web Applications. Workshop in conjunction with GSCL 2011, 26 September 2011, University of Hamburg*.

Koeva, S. (2006). Inflection Morphology of Bulgarian Multiword Expressions. In *Computer Applications in Slavic Studies*, pages 201–216. Boyan Penev Publishing Centre, Sofia.

Koeva, S. (2007). Multi-word Term Extraction for Bulgarian. In *Balto-Slavonic Natural Language Processing 2007*, pages 59–66, Prague.

Landauer, T. K., McNamara, D., Dennis, S., and Kintsch, W. (eds.). (2007). *Handbook of Latent Semantic Analysis*. Mahwah NJ: Lawrence Erlbaum Associates.

Leseva, S. and Stoyanova, I. (2008). Treatment of Named Entities in Machine Translation. In Blanko, X. and Silberztein, M. (eds.), *Proceedings of the 2007 International Nooj Conference*, pages 254–272, Barcelona, Spain. Cambridge Scholars Publishing.

Leseva, S. and Stoyanova, I. (2014). Wikipedia as a Source for Linguistic Resources – Corpora, Dictionaries and Language Models. In Koeva et al. (ed.), *Language Resources and Technologies for Bulgarian (in Bulgarian)*. In press.

Manning, C. and Schutze, H. (1999). *Foundations of Statistical NLP*. MIT Press.

Nakov, P. (2008). Noun Compound Interpretation Using Paraphrasing Verbs : Feasibility Study. In *Proceedings of the 13th International Conference on Artificial Intelligence Methodology Systems Applications AIMSA08 (2008)*, pages 103–117.

Nunberg, G., Sag, I. A., and Wasow, T. (1994). Idioms. *Language*, 70:491–538.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, Mexico.

Smadja, F. (1993). Retrieving Collocations from Text: Xtract. *Computational Linguistics*, 19:143–177.

Stoyanova, I. (2012). *Automatic Recognition and Tagging of Compound Lexical Units in Bulgarian (PhD thesis, in Bulgarian)*.

Todorova, M. and Obreshkov, N. (2008). Compilation of Inflectional Dictionaries Using WordEditor. In Tadić, M., Koeva, S., and Dimitrova-Vulchanova, M. (eds.), *Proceedings of the 6th International Conference on Formal Approaches to South Slavic and Balkan Languages*, pages 131–138.

Todorova, M. and Stoyanova, I. (2014). MWE Dictionaries in Bulgarian. In Koeva et al. (ed.), *Language Resources and Technologies for Bulgarian (in Bulgarian)*. In press.

Todorova, M. (2006). On The classification of Bulgarian Non-Free Phrases. In Vulchanova, M. D., Koeva, S., Krapova, I., and Vulchanov, V. (eds.), *Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages*, pages 251–256, Sofia, Bulgaria.

Widdows, D. (2008). Semantic Vector Products: Some Initial Investigations. In *Proceedings of the Second AAAI Symposium on Quantum Interaction*.

Temporal Adverbs and Adverbial Expressions in a Corpus of Bulgarian and Ukrainian Parallel Texts

Ivan Derzhanski
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
iad58g@gmail.com

Olena Siruk
Institute of Philology
Taras Shevchenko National University of Kyiv
olebosi@gmail.com

Abstract

This paper presents a comparative bilingual corpus-based study of the use of several frequent temporal adverbs and adverbial expressions ('always', 'sometimes', 'never' and their synonyms) in Bulgarian and Ukrainian. The Ukrainian items were selected with the aid of synonym dictionaries of words and of set expressions, the corpus was used to identify their most common Bulgarian counterparts, and the frequencies of the correspondences were compared and scrutinised for possibly informative regularities.

1. Introduction

Although corpus-based contrastive research is quickly gaining momentum today, it is still at a very early stage, because parallel corpora are as yet available for few language pairs and the methods of their processing have only started being developed. In this paper we make a contribution to this field by addressing a pair of languages, Bulgarian and Ukrainian, which have received little attention in this regard, and evolving an algorithm for comparative analysis of lexical and phraseological units from a chosen lexical and semantic field on the basis of corpus and dictionary data.

The paper presents a comparative bilingual corpus-based study of the use of several frequent time adverbs and adverbial expressions with the meanings 'always', 'sometimes' and 'never' in Bulgarian and Ukrainian.

The working Bulgarian–Ukrainian parallel corpus (Siruk, Deržans'kyj, 2013; Siruk, Derzhanski, 2013) is composed entirely of fiction (mainly novels, although some shorter works have been included as well), including both original Bulgarian and Ukrainian texts and translations from other languages. The overall word count is currently 6.35 million in Bulgarian and 5.58 million in Ukrainian.[1] All texts have been aligned at sentence level with Hunalign (Varga et al., 2005), with subsequent manual correction of alignment errors.

A number of Ukrainian adverbs and adverbial expressions with the meanings 'always', 'sometimes' and 'never' were selected at the first stage of the research by means of a Ukrainian synonym dictionary (Burjačok et al., 1999) and a dictionary of phraseological synonyms (Kolomijec', Rehuševs'kyj, 1998). They formed a lexical and semantic group with three subgroups, composed of 14, 21 and 16 units respectively, not counting variants.

The use of both a word dictionary of synonyms and a phraseological one is expedient in order to achieve a more complete coverage of the semantic field because it is not always easy to estimate the power of the semantic cohesion of adverbial expressions. The dynamics of the forming of the common meaning of a cliché's constituents is also seen in the variations in the orthography of many of the

[1] The difference is due both to the contrast between the syntactic characters of the Bulgarian and the Ukrainian language (analytic *versus* synthetic) and the prevailing tendencies of the translators of the two schools towards comprehensiveness and conciseness, respectively.

expressions in both languages, whose components are written separately at first, but later can become a hyphenated or solid word.

At the second stage of the work, the corpus was used to find (through regular expression search) instances of the Ukrainian temporal adverbs and adverbial expressions drawn from the dictionaries and to identify their Bulgarian translation correspondences. These correspondences were checked against a Bulgarian dictionary of synonyms (Nanov, Nanova, 2000) and of phraseological synonyms (Nanova, 2005). Then they served in turn to locate further Ukrainian translation correspondences with the goal of expanding the lexical and semantic group. Finally, the frequencies (more precisely, the numbers of occurrences) of the matches were compared and studied.

2. 'always'

The lexical and semantic group ZAVŽDY[2] is defined in the Ukrainian synonym dictionary as 'all the time or over a certain temporal segment—invariably' and contains 14 items, not counting variants: *postijno, povsjakčas, vsjakčas [usjakčas], doviku, poviky, povik, povik-viky, povik-vikiv, vvik [uvik]* (colloq.), *zavše (*colloq.), *zavsidy [zavsihdy]* (colloq.), *vse [use]* (colloq.), *zajedno* (dial.), *skriz'* (rare) (Burjačok et al., 1999, vol. 1: 511, s.v. ZAVŽDY). All items in the group are single-word adverbs. The dictionary of Ukrainian phraseological synonyms does not feature such a group at all.

In the texts that make up the parallel corpus *zavždy* dominates absolutely (3,697 occurrences), followed by *nazavždy* 'for ever' 366, *vično* 'eternally' 278, *na viky* or *navik(y)* 271, *zavše* 201 and *postijno* 'constantly' 131.[3]

The makeup of the corresponding lexical and semantic group in Bulgarian is very similar. It is led by *vinagi* (3,462 occurrences), followed by *zavinagi* 'for ever' 446, *večno* 392, *postojanno* 'constantly' 390, *vsjakoga* (a close synonym of *vinagi*) 263, *naveki* 143 and *neizmenno* 'invariably' 94. The entry in the dictionary of synonyms (Nanov, Nanova, 2000: 61, s.v. VINAGI) also lists the stylistically marked items *vsjakogaž* (folk), *vsegda* (bookish, obs.) and *sjavga* (dial.), which do not occur in our corpus.

The numbers of occurrences of the translation correspondences involving the two languages' most common 'always' items are given in Table 1. As in all tables in this paper, the Bulgarian words label the rows and the Ukrainian ones the columns.

	zavždy	*zavše*	*postijno*	*vično*	*nazavždy*	*naviky*
vinagi	**2,191**	**129**	20	13	20	3
vsjakoga	**156**	4			2	1
postojanno	**53**	1	**41**	3	3	
večno	**48**	1	3	**202**	1	1
zavinagi	20	2		4	**250**	**67**
naveki	1	2		2	8	**86**

Table 1: Correspondences between items expressing the meaning 'always'.

Of some interest here is the high frequency with which Bulgarian *zavinagi* corresponds to Ukrainian *naviky*, and Bulgarian *postojanno* and *večno* to Ukrainian *zavždy*.

(1) Uk: *Odyn raz vidmovyšsja vid svobody, a todi naviky zabudeš, ščo to take.* 'You'll relinquish freedom once, and then you'll forget what it is for good.'

Bg: *Otkažeš li se vednăž ot svobodata, zabravjaš zavinagi kakvo e tja.*

(Pavlo Zahrebelnyi, *Roksolana*)

(2) Uk: *Vona zavždy tak žorstoko mene obražala!* 'She always wounded me so cruelly.'

[2] We use the 1898 scientific transliteration system that is predominant in international linguistic publications on Cyrillic-written Slavic languages for both Bulgarian and Ukrainian.
http://en.wikipedia.org/wiki/Scientific_transliteration_of_Cyrillic

[3] Strictly speaking, 'for ever' is a different semantic field, but it has a significant overlap with 'always', and Bulgarian words from one field often correspond to Ukrainian words from the other.

Bg: *Tja postojanno me zasjagaše žestoko.*

(Charlotte Brontë, *Jane Eyre*)

(3) Uk: *— I Flores, vidljud'ko, zavždy poxmuryj Flores zasmijavsja.* 'And Flores, the anchorite, Flores the always frowning, smiled.'

Bg: *— i Flores, samoživecăt, večno namrăštenijat, mračnijat Flores se zasmja.*

(Alexander Belyaev, *The Shipwreck Island*)

Generally, adverbs derived from the noun *vik* are more numerous and more frequent in Ukrainian than derivations of its etymological counterpart *vek* are in Bulgarian, in line with the fact that in Ukrainian this noun has a wider range of meanings ('life, lifetime' as well as 'century' and 'age, epoch', which are shared by the Bulgarian word as well).

3. 'sometimes'

For the semantic field 'sometimes' the Ukrainian dictionary of synonyms lists 19 adverbs and set expressions, not counting variants (Burjačok et al., 1999, vol. 1: 644, s.v. *INODI*). The group only includes items which allow bounding from above ('not always', 'not often'), as demonstrated by their ability to co-occur with the restrictive modifiers *lyše* and *til'ky* 'only'. The dictionary of phraseological synonyms (Kolomijec', Rehuševskyj 1998: 62) gives four set expressions (clichés), two of which (*čas vid času* and *vid času do času*, actually variants of the same) are also given in the synonym dictionary. The group contains both adverbs and adverbial expressions with various structures.

In the corpus only four of these appear with significant frequencies, namely *inodi*, *časom*, *čas vid času* (with variants *čas od času*, *čas do času* and *vid/od času do času*) and *inkoly*, with 860, 828, 459 and 385 occurrences, respectively. They are followed by *raz u raz* (*raz po raz*, *raz za razom*) 'time and again' 317 and, far behind, by *zridka* 109, *vrjady-hody* (*urjady-hody*) 82 and *podekoly* 49.

In the matching Bulgarian sentences two items dominate: these are *ponjakoga* (with its rare variant *ponjavga*) and *ot vreme na vreme* (also written *otvreme-navreme*) 'from time to time', with 1,836 and 761 occurrences. Next come the pointedly colloquial *segiz-togiz* 59, *čas po čas* 51 and the archaic *navremeni* 47. The idiom *ot dăžd na vjatăr* (lit. 'from rain to wind') only occurs seven times in the corpus. No occurrences were found of *ponjakogaž*, *sporadično*, *izrjadko*, *čat-pat* and *napăti*, which are also listed in the dictionary of synonyms (Nanov, Nanova, 2000: 436, s.v. *NJAKOGA*).

The distribution of translation correspondences is given in Table 2.

	inodi	*inkoly*	*časom*	*čas vid času*	*podekoly*	*zridka*	*vrjady-hody*	*raz u raz*
ponjakoga	**603**	**255**	**452**	43	**35**	13	10	8
ot vreme na vreme	55	28	102	**295**	12	**34**	**35**	**40**
navremeni	3	4	**14**	6		2	1	2
segiz-togiz	4	4	8	**18**		6	5	1
čas po čas	1		6		1			**18**

Table 2: Correspondences between items expressing the meaning 'sometimes'.

It is obvious that the three frequent Ukrainian adverbs *inodi*, *inkoly* and *časom* are very similar in behaviour, and indeed the choice between them seems to be largely a matter of individual preference: there are texts in the corpus which use almost exclusively *inodi*, or nearly nothing but *časom*, or all three to an approximately equal extent. We may note, however, that *časom* corresponds to Bulgarian *navremeni* more often than the others, which may be accidental (given the shallow amount of data), though the correlation with the fact that both adverbs are derived from the nouns meaning 'time' (*čas* and *vreme*, respectively) is certainly interesting.

Another difference, concerning the co-occurrence of the adverbs and adverbial expressions with Bulgarian samo and Ukrainian lyše and til'ky 'only', is shown in Table 3.

	total	with 'only'	percentage
ponjakoga	1,835	28	1.53%
ot vreme na vreme	761	38	4.99%
navremeni	46	3	6.52%
segiz-togiz	59	9	**15.25%**
inodi	860	24	2.79%
inkoly	385	12	3.12%
časom	828	5	0.60%
čas vid času	459	10	2.18%
zridka	109	35	**32.11%**
vrjady-hody	82	14	17.07%
podekoly	49	0	0.00%

Table 3: Co-occurrence of some 'sometimes' items with 'only'.

If *inodi*, *inkoly* and *časom* are counted together, we see a strong correlation between them and the adverb *ponjakoga* on one hand, and between the set expression *čas vid času* (with its variants) and its near-literal counterpart *ot vreme na vreme* 'from time to time', on the other. The translators' tendency to stay close to the originals should explain this to some extent, but not entirely. A further contrast is shown in Table 4: the single-word adverbs are the only 'sometimes' items that often correspond to 'often' in the other language. This may indicate imprecise translation on some occasions, but the frequency with which it happens is too great to overlook, and suggests a semantic reason as well (greater proximity to the upper end of the frequency scale).

	inodi + *inkoly* + *časom*	*čas vid času*	*raz u raz*	*zridka*	*často* 'often'	*ridko* 'seldom'
ponjakoga	**1,310**	43	8	13	**19**	1
ot vreme na vreme	185	**295**	**40**	**34**	1	
često 'often'	**70**	9	18			
rjadko 'seldom'	4	1		**21**		

Table 4: Correspondences between 'sometimes' and 'often' or 'seldom' items.

One can note that Bulgarian *ot vreme na vreme* often corresponds to the structurally similar Ukrainian *raz u raz*.

The existence in Ukrainian of the adverb *zridka* (related to *ridko* 'seldom'), preferred host of *lyše* and *til'ky* 'only' and frequent translation correspondence of Bulgarian *rjadko* but with no precise counterpart in Bulgarian, constitutes yet another major difference between the two systems of expressions that lexicalise the meaning 'sometimes'.

(4) Uk: […] *moja Kateryna tak varyt' galušky, ščo j het'manovi zridka dovodyt'sja jisty taki.* 'My Kateryna cooks such dumplings that even the hetman seldom gets to eat the like.'
Bg: *Mojata Katerina pravi takiva galuški, kakvito i hetmanăt rjadko može da jade.*
(Nikolai Gogol, *A Terrible Vengeance*)

Concerning the adjacent semantic field of 'sometimes' with no upper bound (items absent from the entry s.v. *INODI* in the Ukrainian synonym dictionary), the most conspicuous observations from the corpus are the frequent use of Ukrainian *raz u raz* and its variants and the high frequency of the iterative verb *buvaty* 'be occasionally, be regularly, happen' as a main verb or a parenthetic word, which corresponds to the Bulgarian adverb *ponjakoga* on 72 occasions:

(5) Uk: *"Istoryčna misija", — kazav, buvalo, Brjans'kyj…* '"An historical mission," Bryansky used to say.'

Bg: *"Istoričeska misija" — kazvaše ponjakoga Brjanski…*

(Oles Honchar, *Guide-on Bearers*)

The Bulgarian verb *slučvam se* 'happen' has a similar function, but a much lower frequency; it corresponds to a Ukrainian adverb (*inodi*, *časom*, *inkoly*, *vrjady-hody*) only 37 times.

(6) Uk: *Lyše inodi vin zryvajet'sja na kil'ka hodyn i raptom padaje, mov jastrub, pronyzanyj striloju.* 'Only sometimes it starts up for several hours and suddenly falls down like a hawk pierced by an arrow.'

Bg: *Slučva se da duha i samo njakolko časa i izvednăž sekva, kato orel, pronizan ot strela.*

(Bolesław Prus, *Pharaoh*)

4. 'never'

The lexical and semantic group *NIKOLY*, defined in both source dictionaries as 'at no time, under no circumstances'), consists of nine adverbs, mostly stylistically marked ones, in the synonym dictionary (*zrodu* emph., colloq., *zrodu-viku [zrodu-zviku]* emph., colloq.; *doviku* emph., *povik* emph., *poviky* rare, *vik* emph., colloq., *povik-viky [povik-vikiv]* emph., poet., *vvik [uvik]* emph., colloq., *vovik* [*voviky*] arch., emph., colloq. (Burjačok et al. 1999, vol. 1: 1021, s.v. *NIKOLY*) and seven set expressions, not counting variants, in the dictionary of phraseological synonyms (Kolomijec', Rehuševskyj, 1998: 82). The data from the two dictionaries don't intersect. The group is large in size, and its elements vary in structure. Among the three groups, this is the only one to contain set expressions with a high level of semantic cohesion, and two of these expressions were found in the parallel corpus, both times with different but likewise idiomatic Bulgarian translation counterparts:

(7) Uk: *Nu, to pobačyš joho, jak svoje vuxo.* 'Well, you'll see him as [you'll see] your ear.'

Bg: *Šte go vidiš, kogato si vidiš vrata.* '… when you see your neck.'

(Henryk Sienkiewicz, *The Teutonic Knights*)

(8) Uk: *Jak rak svysne?* 'When pigs fly?', lit. 'When the crayfish whistles?'

Bg: *Na kukovo ljato?* ditto, lit. 'At cuckoo's summer?'

(Bogomil Raynov, *Typhoons with Tender Names*)

In Bulgarian between 30 and 60 set expressions with the meaning 'never', not counting variants, are registered (Ničeva et al., 1974; Nanova, 2005):

(9) Uk: *Ajakže, čorta puxloho dočekaješsja!…* 'Oh sure, the hell you'll live to see it!…'

Bg: *Kak ne, na kukovden!…* 'Sure thing, on the first of Never!…', lit. 'on Cuckoo's day'.

(Mykhailo Kotsiubynsky, *Fata Morgana*)

That said, the corpus in fact does little justice to the wealth of set expressions for 'never' that exists in either language. It does, however, feature some of the Ukrainian adverbs, especially *zrodu* (*zrodu-viku*, *zrodu-zviku*) with 187 occurrences, *doviku* with 94 and *povik* (*povik-viku*) with 33 (recall that the latter two have also the meaning 'always' or 'for ever' when used in affirmative contexts). Contrary to the synonym dictionary's explicit statement, *zrodu* proves not to be limited to the past; it is applicable to the future as well:

(10) Uk: *Koly rozpovidaješ jim pro svoho novoho pryjatelja, vony zrodu ne pocikavljat'sja najistotnišym.* 'When you're telling them about your new friend, they will never be interested in the most substantial.'

Bg: *Kogato im razpravjate za njakoj nov prijatel, te nikoga ne vi pitat za naj-săštestvenoto.*

(Antoine de Saint-Exupéry, *The Little Prince*)

While unusual in the use of *zrodu* for the future, this example is typical in that the Bulgarian uses the regular adverb *nikoga* 'never'. The evidence of the corpus shows that Bulgarian has no other 'never' item comparable to Ukrainian *zrodu* in frequency, and the derivatives of *vek* are used with a negative meaning less often (and in the corpus not at all) than their Ukrainian etymological counterparts.

5. Conclusions

The comparative analysis of the lexical and semantic field of temporal adverbs and adverbial expressions on the basis of parallel texts makes it evident that this field is richer in synonyms in Ukrainian, whereas in Bulgarian it is generally more monolithic (this conclusion seems to be in variance with dictionary data,

but this can be explained with the fact that dictionaries cover specific ranges of genres, not restricted to fiction). The comparison of the frequencies with which words and expressions of one language correspond to words and expressions of the other in parallel sentences reveals subtle semantic oppositions and demonstrates the structure of the semantic fields and the relations between them.

As a side result of the search in the bilingual corpus, some items not marked in dictionaries as rare are shown to be so, which raises the question of checking the actual frequency of their use by the help of larger (and balanced) monolingual corpora.

The method employed in this investigation, which is readily applicable to other temporal adverbs and adverbial expressions and to other semantic fields, contributes to the comparative study of different languages' pictures of the world and, on a more practical level, holds potential for the improvement of synonym and bilingual dictionaries.

References

Burjačok et al., 1999: Бурячок, А.А., Гнатюк, Г. М. (ред.) (1999). *Словник синонімів української мови.* Наукова думка, Київ.

Kolomijec', Rehuševskyj, 1998: Коломиец Н. Ф., Регушевский Е. С. (1998). *Словник фразеологічних синонімів.* Ред. В.О. Винник. Издательство Радянська школа, Київ.

Nanov, Nanova, 2000: Нанов, Л., Нанова, А. (2000). *Български синонимен речник.* София: Хейзъл.

Nanova, 2005: Nanova, A. (2005). *Фразеологичен синонимен речник на българския език.* София; Хейзъл.

Ničeva et al., 1974: Ничева, К., Спасова-Михайлова, С. Чолакова, Кр. (1974). *Фразеологичен речник на българския език.* Изд. на БАН: София.

Siruk, Deržans'kyj, 2013: Сірук, О., Держанський, І. (2013), Лексичні перекладні еквіваленти в болгарських і українських паралельних текстах. *Українське мовознавство*, 43(2013):75–86.

Siruk, O., Derzhanski, I. (2013). Linguistic Corpora as International Cultural Heritage: The Corpus of Bulgarian and Ukrainian Parallel Texts. *Digital Presentation and Preservation of Cultural and Scientific Heritage* (III/2013):91–98.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V. (2005). Parallel Corpora for Medium Density Languages. In *Proceedings of the RANLP'2005*, pages 590–596. The program Hunalign: `http://mokk.bme.hu/resources/hunalign/`

Historical Corpora of Bulgarian Language and Second Position Markers

Tsvetana Dimitrova
Institute for Bulgarian Language
Bulgarian Academy of Sciences
cvetana@dcl.bas.bg

Andrej Bojadžiev
Faculty of Slavic Studies
Sofia University
aboy@uni-sofia.bg

Abstract

This paper demonstrates how historical corpora can be used in researching language phenomena. We exemplify the advantages and disadvantages through exploring three of the available corpora that contain textual sources of Old and Middle Bulgarian language to shed light on some aspects of the development of two words of ambiguous class. We discuss their behaviour to outline certain conditions for diachronic change they have undergone. The three corpora are accessible online (and offline – for downloading search results, xml files, etc.).

1. Introduction

This paper presents part of an ongoing work on the historical evolvement of clausal second position clitics and the clitic cluster in Bulgarian which attempts at explaining the conditions for the placement and movement of clitics and clitic-like elements towards the second position in the phrase and/or clause (it is the position immediately after the first emphatic (strong or stressed) syntactic constituent – the so-called Second Wackernagel position where reflexive, discourse, interrogative, and pronominal clitics can be found in different periods in the history of Bulgarian language). In this paper, we discuss the behaviour of two words – *бо* (*bo* "for, then") and *оубо* (*oubo* "then, indeed, therefore") that are often found in second position, in the context of methodological issues in development of historical corpora.

In the next section, we present the three corpora we have used for our study with a brief overview of their characteristics. In section 3., we discuss a couple of practical issues in dealing with historical corpora. Section 4. contains an empirical study of the two words that are often classified as conjunctions or particles in the traditional literature with an outline of the conditions when the research has to employ the data from historical corpora available.

2. The Corpora

We started our study by excerpting data from three corpora with Old Church Slavonic/Old Bulgarian texts. They are representative of the textual collections available nowadays for linguists to work with. The first – PROIEL corpus[1] – contains annotated texts without considering the variation in data, redactions, and transparent access to parallel data (the corpus contains parallel texts but they have been used for automatic and semi-automatic annotation and texts are not readily available in parallel). The second – Old Church Slavonic subcorpus in the TITUS database[2] – gives parallelized texts but they have been lemmatised only; parallel data involves the gospel text. The third corpus – the Historical Corpus of Bulgarian Language[3] – has being developed for a couple of years to give access to an impressive electronic collection of texts – broad and diverse, although lacking transparent annotation so far.

[1] http://foni.uio.no:3000/users/sign_in
[2] http://titus.uni-frankfurt.de/indexe.htm
[3] http://histdict.uni-sofia.bg/textcorpus/list

The PROIEL corpus has been developed at the University of Oslo since 2008. The corpus contains the gospel text from *Codex Marianus* (following the edition of Vatroslav Jagić, cf. Jagić, 1883), parts of the gospel text according to *Codex Zographensis* (again following the Jagić's edition, cf. Jagić, 1879) that is missing in *Codex Marianus* (Matthew 1:1 – 1:27) and texts from *Codex Suprasliensis* (this part of the corpus is still under preparation, and not all texts from *Codex Suprasliensis* are included and annotated; here, we use only the available texts[4]). Although the texts are annotated (normalized wordform, lemma, part-of-speech, and applicable morphological information, plus syntactic annotation and attempt at information structure annotation), there is no readily usable marking of corresponding passages across languages and texts. We have isolated the patterns (syntactic, with respect to word ordering and right and left adjoined constituents) that we are interested in for the discussion in Section 3. However, texts are translations, so the access to sources pertaining to different redactions and/or translations, is needed to support the comparative research across texts and language phenomena (as shown by example (4) in 4.1., there are well known differences between the texts according to different manuscripts).

The TITUS corpus gives a valuable access to aligned and parallelized texts albeit not annotated with morphological and/or syntactic information. However, they are lemmatized and it is easy to search for different inflectional and orthographic forms. Access to parallel texts with corresponding passages across texts and in comparison to Greek New Testament (NT) is easy although it does not resolve the issue of handy access to different sources within the Byzantine tradition. There is no marking of the common passages across texts either (quotations, idiomatic constructions, etc.).

The third corpus – the Historical Corpus of Bulgarian Language (HCBL) – gives access to a great variety of texts (104 as of June 2014), some of which are of very late dating. The still missing annotation makes comparative research a bit complicated but the collection is extremely valuable because it covers texts according to manuscripts (and not editions), some rare and very interesting non-canonical texts, and late developments. This corpus is open-ended in the sense that non-canonical and non-literary materials can be added such as inscriptions, dialect data, databases of toponyms, personal names, etc.

Conditions	**PROEIL**	**TITUS**	**HCBL**
Metadata	Bibliographic reference to the edition only	Mirroring reference to the editions of the manuscripts	Reference to the manuscripts
Access to source	No	No	No
Annotation	Morphological, syntactic, lexical	Lemmatised only	No
Parallel data	No	Yes (no marking of parallel passages, citations, etc.)	No
Search engine	Yes	Yes	No
Text diversity	No	No	Extensive time period and genres

Table 1: Summary of the most important characteristics of the three corpora.

3. Practical Issues

Historical language study relies almost exclusively on written data as there are no sources that are more reliable for this research purpose. Corpus data is the empirical basis for diachronic linguistics, and by analysing it, we build hypotheses about linguistic processes within or outside a particular linguistic theory.

[4] *Codex Suprasliensis* is included as part of the work in the UNESCO-funded project *The Tenth Century Cyrillic Manuscript Codex Suprasliensis* that aimed at digitizing this largest Old Church Slavonic manuscript.

As historical linguists do not have ready and non-compromised access to balanced corpora with well described sources covering entire periods, diverse content and genres, they often search for open-ended databases to collect materials they need. In this context, the notion of corpus may need broadening to cover different resources such as electronic text collections, editions, linguistic atlases, and dictionaries (Kytö, 2011). The Historical Corpus of Bulgarian Language is the only one among the three corpora used for our research that contains texts of diverse time periods and genres. However, it is still neither a corpus because it lacks annotation and metadata, nor a database because it is not really searchable. Therefore, here we define it as an open-ended e-text collection.

The trend, though, makes even harder to collect and align the materials to extract and observe the data because if we aim at studying the language system and its change in time (Mair, 2008), we need to take into account the linguistic phenomena as attested over time. Thus, although we may not be interested in the history of individual texts as instances of the output of the language system, we still have to take into account textual history (and the history of sources) to interpret the data we collect and analyze.

Moreover, if researchers do not have access to thoroughly described and annotated textual data, they may make use of design and arrangement of the data in a way that will rely on already available knowledge (reflected in traditional grammars and dictionaries, already annotated corpora, dialect atlases, and other handy data collections). One such approach involves heuristic alignment of historical texts with contemporary editions and/or translations of the same texts or editions of other texts that are readily available. For example, the TITUS database offers a parallel view of Old Church Slavonic NT text according to different manuscripts (*Codex Marianus, Codex Zographensis, Codex Assemanius, and Codex Sabbae*), Greek NT and Modern Russian NT translation. This parallel view is a fantastic tool for studying parallel constructions and specific phenomena.

In the next section, we will employ the three corpora for a field study on behavior of two words attested as early as the period of the earliest sources and preserved in some contemporary dialects. While summarizing our findings, we will sketch out the specifics of the three corpora.

4. Empirical Study

Our empirical study covers the words *бо*[5] *(bo* "for, then"*)* and *оубо (oubo* "then, indeed, therefore"*)*, with additional notes on *ибо* (*ibo* "because") – the origin of all of them can be traced to *бо*. *Бо* and *оубо* are predominantly found in the second clausal and/or phrasal position after (prosodically and syntactically) strong constituent (in the Second Wackernagel position or 2P). The first strong constituent can be a wh-word in complementizer function such as *кѣто* (*kăto* "who"), *чѣто* (*chăto* "what"), etc., including a prepositional phrase with a wh-word such as *по чѣто* (*po chăto* "why"). The strong constituent (verb, noun, adjective, adverb) in the first position can be preceded by a conjunction or a subjunction, negation particle *не* (*ne* "not"), and/or followed by the reflexive particle *сѧ* (*sen* "self"), discourse particle *же* (*zhe*), pronominal clitics such as *тѧ* (*ten* "you-ACC,Sg"[6]), *ти* (*ti* "you-DAT,Sg"), etc. These are mostly prosodically weak constituents – proclitics or enclitics (depending on whether the strong constituent is after or before them). In section 4.1., we discuss our observations on an annotated corpus (PROIEL), with additional data from the parallel texts included in TITUS. For further analysis, we need the Greek correspondences but parallel and comparable corpora of these sources are not readily available (and annotated). Therefore, we need to look further into traditional critical editions to extract the information about the Greek equivalents (Nestle-Aland, 2013).

4.1. Earlier Texts

In this section, we will present our observations on the earlier texts that are part of the PROIEL corpus with some raw and inconclusive numbers (instances of both *бо* and *оубо* in the two large annotated textual segments of *Codex Marianus* and *Codex Suprasliensis* – respectively, Cod. Mar. and Cod. Supr.).

[5] As the words will be repeated in the next pages, the transliteration will not be repeated and translation is to be given only to differentiate specific meanings in appropriate discussion passages.

[6] The following abbreviations and conventional labels are used in the paper: ACC – accusative; DAT – Dative; GEN – Genitive; Sg – Singular; Pl – Plural; FUT – Future tense form; CL – clitic; QuCL – interrogative clitic; Pron – pronoun; PP – prepositional phrase.

Overall, *Codex Marianus* attests for 172 instances of *оубо* and 343 of *бо*, and the texts of the *Codex Suprasliensis* included in PROIEL contain 272 instances of *оубо* and 442 of *бо*.

Conditions	**Cod.Mar.**	**Cod.Supr.**	**Cod.Mar.**	**Cod.Supr.**
	бо	*бо*	*оубо*	*оубо*
After wh-pronoun (incl. wh in PP)	9	12	47	27
After a verb (incl. *бꙑти* (*byti* "be")	136	128	32	62
After a noun (incl. pronoun, etc.)	109	170	27	50
After any constituent followed by *же*	0	0	10	11
After any constituent followed by *сѧ*	0	0	8	11
After any constituent followed by weak pronoun	0	0	4	7
After any constituent followed by *ли (li* – interrogative particle*)*	0	0	2	2
After *иже (izhe* "who/what"*)*	18	12	3	6
After *аще (ashte* "if"*)*	3	21	16	9
Before *же*	0	0	0	0
Before *сѧ*	13	24	0	0
Before a weak pronoun	21	21	0	0
Before *ли*	0	0	0	0
Before *аще*	10	2	0	0

Table 2: Positions of *бо* and *оубо* after and before other constituents as attested in *Codex Marianus*, and the texts from *Codex Suprasliensis* (in the annotated texts in PROIEL)

Originally, *бо* was a particle for emphasis and verification (Sławski, 1974: 285–286) of the preceding constituent – the emphasized word (often syntactically focused constituent). In the data, *бо* is almost exclusively preceded by only one constituent, except for *дроугъ къ дроугоу* (*drug kŭ drugou "*one another"), and the preceding constituent can be preceded only by a preposition or a negation (*не* "not", *ни* "neither"). Other syntactically weak constituents such as *сѧ* and pronominal clitics are placed after it.

The origin and clausal position of *бо* are parallel to the Greek *γάϱ* (*gar* "for, indeed") that was colloquially used to highlight the faculty or the property of something or someone. In the history of Bulgarian language, *бо* was gradually adopted for various functions, which, on the one hand, overlapped (partially or fully with the meaning of *оубо*), and, on the other, were very close to those of *же* in its function of emphatic particle (there is no co-occurrence of *же* and *бо* alone – not as *никътоже бо*, *иже бо*, etc. - in the texts here). It was also adopted to function as a conjunction – in our data *бо* is found after the negation particle alone (without a preceding constituent). The conditions for the overlap depend on its position and function to emphasize the meaning of the preceding word (just like *же*), as: 1) a marker of cause or reason - "for" (introducing the reasoning); 2) a marker of clarification - "for, you see"*;* 3) a marker of inference - "certainly, by all means, so, then".

The derivation variants of *бо* are many – *ибо* (*ibo,* "for, because"), and *оубо,* among others. They were often used in earlier Old Bulgarian texts to translate specific Greek constructions and are mostly calques (unlike *бо*). The following examples show co-occurrence of *бо* and *и* in the form of *ιбо* (phonetic variant of *ибо* used to translate parallel constructions in Greek (with *καὶ (kai* "and"*)* and γὰρ (*gar* "for, indeed"; see also the occurrence of *и* in the meaning of "even, also" after *ибо*), as in:

(1) a. ı***бо*** *и* бесѣда твоѣ авѣ тѧ творитъ Cod. Mar. Mt. 26:73
indeed even speech your out you give[7]
καὶ γὰρ ἡ λαλιά σου δῆλόν σε ποιεῖ

b. ı***бо*** *и* пси подъ трапезоиѫ ѣдѧтъ Cod. Mar. Mk. 7:28
indeed and dogs under table eat
καὶ γὰρ τὰ κυνάρια ὑποκάτω τῆς τραπέζης ἐσθίουσιν[8]

c. ı***бо*** сънъ чловѣчьскъı не приде Cod. Mar. Mk. 10:45
indeed son human not come
καὶ γὰρ ὁ υἱὸς τοῦ ἀνθρώπου οὐκ ἦλθεν

d. ı***бо*** азъ чловкъ есмъ подъ властелъı оучиненъ Cod. Mar. Lk. 7:8
indeed I man am under authority appointed
καὶ γὰρ ἐγὼ ἄνθρωπός εἰμι ὑπὸ ἐξουσίαν τασσόμενος

The use of *оубо* as particle for explanation and emphasis, if synonymous with *бо*, is considered the earliest (Tseytlin, 1994: 721–722). The further use of *оубо* was dependent on its use after pronouns and pronominal adverbs, mainly in interrogative clauses (after a wh-word) – it is probably among its first functions as it is closest to the particle function (Tseytlin, 1994: 721-722).

(2) a. отъ коудѫ ***оубо*** иматъ плѣвелъ Cod. Mar. Mt. 13:27
from where then have weed
πόθεν ***οὖν*** ἔχει ζιζάνια

b. Кто ***оубо*** естъ вѣрънъı рабъ и мѫдръı. Cod. Mar. Mt. 24:45
who then be faithful servant and wise
Τίς ***ἄρα*** ` ἐστὶν ὁ πιστὸς δοῦλος καὶ φρόνιμος,

c. почто ***ѻубо*** ѻсѫждаıѥши· ıѥгоже богъ не ѻсѫждаıѥтъ·
why therefore judge whom God not judge
Τί ***τοίνυν*** κρίνεις ὅν ὁ Θεὸς κατακρίνει
Cod. Supr. 359:1 (PROIEL Supr. 31:147-148)

d. бракъ ***оубо*** готовъ естъ Cod.Mar. Mt. 22:8
marriage truly ready be
Ὁ ***μὲν*** γάμος ἕτοιμός ἐστιν,

The corresponding Greek constituents vary a lot – *ἄρα* (*ara* "then"), *μὲν* (*men* "indeed"), *οὖν* (*un* "therefore"), *τοίνυν* (*toinun* "indeed, therefore"). The conjunction *οὖν* "then, therefore" is overwhelmingly placed in second position and is also found as *εἰ οὖν* (*ei un*) – *аще оубо* (*ashte oubo* "if then"). The adverb *μὲν* "indeed, truly" in (2d) occurs after the article in the NT Greek text while *оубо* is in 2P.

оубо can be found (albeit sporadically) in the first clausal position – typical for subjunctions and conjunctions (5 instances in Cod. Mar., and 2 in Cod. Supr.), and in the last position (as some adverbs, 1 in Cod. Mar., 2 in Cod. Supr.). *оубо* is also found immediately after a weak constituent such as the conjunction *и* (*i* "and") and *да* (*da* "to"). If there is another clitic, *оубо* is usually found after it or after clitics in the clitic cluster (unlike *бо*). This means that it is placed (almost) exclusively after weak constituents such as *сѧ*, *ли, же* – (3a) and (3b), and pronominal clitics such as *ти* (*ti* "you-DAT") and *ми (mi* "me-DAT") – (3c).

[7] Glosses are given only if there is no appropriate translation, i.e., *dogs* instead of dog-PL, but Israel-DAT (for the Dative form).

[8] Nestle, Aland, 1979: 113, readings from various witnesses. The version of PROIEL follows Tischendorf, 1869: καὶ τὰ κυνάρια.

(3) a. слъшасте ли ***оубо*** Cod. Supr., 1, 3, 14a, 12 (27)
heard QuCL indeed

b. състарѣвъ же сѧ ***оубо*** Cod. Supr., 1, 16, 104b, 2 (208)
he became old Disc CL ReflCL indeed

c. подобааше ти ***оубо*** Cod. Mar. Mt. 25:27
ἔδει σε ***οὖν***
suited you-DAT indeed

There are isolated examples of immediate closeness to *оубо* and *бо* that can be interpreted as a result of an overlap in their functions. In TITUS, there is even a disagreement in translations in the parallel corpus (*бо оубо* in *Codex Marianus*, only *оубо* in *Codex Assemanius*, and *оудобь* in *Codex Zographensis*).

(4) a. ѣко ***бо оубо*** събираѭтъ плѣвелы. Cod. Mar. Mt. 13:40
as therefore is granted the weeds
ὥσπερ ***οὖν*** συλλέγεται τὰ ζιζάνια

b. Ѣкоже ***оубо*** плѣвелꙑ събираѭтъ сѧ · Cod. Assemanius Mt. 13:40
as therefore weeds granted

c. ѣко ***оудобь*** събираѭтъ плѣвелы Cod. Zogr. Mt. 13:40
as conveniently(?) granted weeds

The example with the variant readings in (4) shows that the correct interpretation of the language phenomena with respect to the language change requires access to parallel data.

4.2. Open-ended Text Collection

In this section, we discuss the additional data available through an open-ended text collection where we follow the changes in the phenomena. Sources are part of the Historical Corpus of Bulgarian Language which comprises diverse texts, with some very late ones such as *Damascenus Troianensis* (17th c.; NBKM № II, 11 or Kodov 88).

The raw statistics (without taking into account different meanings) shows interesting results with many later non-canonical sources exhibiting higher number for *оубо* and not for *бо* (in contrast to the earlier sources). The observations give a complex picture of the interplay between *бо* and *оубо*.

Source	бо	оубо
Zlatoust of Jagić (13th c.; RNB, St. Petersburg, Q.п.I.56)	*525*	*17*
Manasii Chronicle (14th c.; GIM, Moscow, Syn 38)	249	434
Borili Regis Synodicum (14th c.; NBKM 289)	1[9]	37
Codex of German (14th c.; Library of Romanian Patriarchy, №1)	486[10]	115
Laudatio sanctae magnae martyris Dominicae (1479; Rila Mon. 4/8, 603v-611v)	45	42
Laudation sanctorum magnorum aeqalium apostolic regum Constantini et Helenae (1483; Rila Mon. 4/5, 424r-439r)	44	82
Vita et acta sancti patris nostril Hilarionis episcope ex Moglen (1483; Rila Mon. 4/5, 161r-175r)	41	57

[9] Co-occurring with *оубо*.
[10] With one co-occurrence: *ꙗко бо ѳбо и колико иж*.

Vita et acta sancti patris nostril Ioannis in monte (14th c.; Zogr. Mon. 172 (olim 103 II g.6), 93r-104r)	30	81
Vita et acta sanctae matris nostrae Parascevae (14th c.; Zogr. Mon. 172 (olim 103 II g.6, 93r-104r), 74r-82v)	36	38

Table 3: Occurrences of *бо* and *оубо* in later texts from the Historical Corpus of Bulgarian Language

In the latest source – *Damascenus Troianensis* – there are no instances of *оубо* and *бо*. Historical-apocalyptic literature consistently prefers *оубо* instead of *бо* in later texts. *In Homilia Hypatii Ephesiensis* there is only *бо* (disregarding the meaning), as in the following examples:

(5) a. бь **бо** вь шесты дн҃ь се в'се бзѣ сьдѣлавь · послѣднеѥ дѣло[11] ·
God indeed on sixthday (in the name of the God made last thing)

b. надь тѣми **бо** вьтораꙗ смр'ть не имать власти
over them-INST indeed second death not has power

The same is observed in *Visio Danielis propheti. De regibus. De novissimis diebus. De fine saeculi*:

(6) и сьразеть **бо** се бранию крѣпькою
and (stroke down) indeed (with the fierce battle)

A possible explanation extends to postulated stylistic differences between *бо* и *оубо*. In *S. Methodii episcopi revelatione de regibus et novissimis diebus*, all 17 instances of *бо* are associated with different meanings; *оубо* is found only once but in the same discourse contexts as *бо* – in (7d) below, where we give the translation of the segment with the difference in the meaning between the two words.

(7) a. рече **бо** бь Изл҃ю ·
said then God Israel-DAT

b. вь ти **бо** дн҃и · боу(д)ть чл҃вци ·
in these then days be-FUT men

c. творити **бо** нач'н]еть тьг(д)а · знам[ениꙗ и] чюд[еса многа
create then start then signs and wonders many

d. тог(д)а всѣке **бо** хетрости/!/ то диꙗволоу съкр[а]тѣють
then every then skills Conj Demon-DAT go short of

и не оуспѣють ничесоже сиі **оубо** нечисти скврьньни гноусни ѥзыци
and not succeed nothing-GEN this truly sinful unclean disgusting people
"**then** every Devil's skills will disappear, and these all **truly** sinful unclean disgusting people will not succeed"

The observations are additionally hampered by the orthographic variants such as *бѡ* and *бо*; *оубо*, *8бо*, *8бѡ*, *оубѡ*, etc. Variation in graphics and the changes in lexical and morphological forms of the words are among the greatest obstacles to the annotation and structuring of these data.

Nowadays, *бо* can be found in most Slavic languages (Trubachev, 1975: 141–142). It has preserved its particle function, and keeps the second position. In Russian dialects, *бо* is synonymous with *же* as in:

[11] The examples are excerpted from the corpus so there is no reference to edition (`http://histdict.uni-sofia.bg/textcorpus/list`).

Садись бо, *принеси бо* "Take a sit *then*, bring along *then*". If it is kept as a conjuncition, it moves towards the first position in the clause as in the Russian Smolensk dialect *Ня пойде, бо боится яго* "(He) didn't go *because* he is afraid of him" (Filin, 1968: 34–35). The last example shows that *бо* has kept its unique syntactic function of connecting two clauses while it is placed in the second clause but not in the first position of the clause it introduces (unlike most conjunctions).

Some authors (Mladenov, 1941: 36) have stipulated that Bulgarian dialects keep traces of *бо* in *боедно (boedno)*, *боедна (boedna), боедно (boedno)* (with variants of *буд-* (*bud-*), *бад-* (*bad-*) in the Rhodope and Southern Bulgarian dialects) to be traced back to *бо един*, *бо едно*, *бо една* with the meaning of the indefinite pronoun *някой* (*nyakoy "somebody-M"), някоя* (*nyakoya "somebody-F"), някое* (*nyakoe "somebody-N"*), and sporadically can be interpreted as negative pronouns *никой* (*nikoy "nobody-M"), никоя* (*nikoya "nobody-F"), никое* (*nikoe "nobody-N"*) (Mirchev, 1932). However, the *Bulgarian Etymological Dictionary* suggests etymology from **любо едьнъ* (BER, 1971). *Бо* can be found very later, although sporadically, as a conjunction in the meaning of "because" (Ilchev, 1974: 37).

5. Closing Remarks

The discussion above shows that the benefits of a corpus study for an observation on the evolvement of language phenomena in context. However, neither available collection of historical texts of Bulgarian language offers working access to structured comprehensive data. The lack of context means that valuable linguistic information on syntax, for example, remains hidden which hampers the access to syntax-semantics information for the status of the markers we have studied in this paper.

The historical linguists interested in the history of Bulgarian still need structured resources with user-friendly marking (annotation) of the linguistic information, metadata (sources, dating, editions, etc.) and visualization and search interface to allow them to make use of valuable data.

Acknowledgements

The present paper was partially prepared within the project *Integrating New Practices and Knowledge in Undergraduate and Graduate Courses in Computational Linguistics* (BG051PO001-3.3.06-0022) implemented with the financial support of the Human Resources Development Operational Programme 2007 – 2013 co-financed by the European Social Fund of the European Union. The authors take full responsibility for the content of the present paper.

References

Jagić, V. (1879). *Quattuor evangeliorum codex glagoliticus olim Zographensis nunc Petropolitanus.* Berlin.

Jagić, V. (1883). *Quattuor Evangeliorum versionis palaeoslovenicae Codex Marianus Glagoliticus.* Saint Petersburg.

Kytö, M. (2011). Corpora and Historical Linguistics. *Revista Brasileira de Linguística Aplicada.* 11(2): 417-457.

Mair, C. (2008). Corpora and the Study of Recent Change in Language. In *Corpus Linguistics: an International Handbook.* Berlin/New York: Walter de Gruyter.

Nestle-Aland (1979). *Greek-English New Testament. 26th revised edition*. Stuttgart: Deutsche Bibelgesellschaft.

Nestle-Aland (2013). *Novum Testamentum Graece. 28th revised edition*. Stuttgart: Deutsche Bibelgesellschaft.

Sławski, F. (1974). *Słownik prasłowiański*. Tom I. A–B. Wrocław-Warszawa-Kraków-Gdańsk: Wydawnictwo Polskiej Akademii nauk.

Tischendorf, C. v. (1869). *Novum Testamentum Grecae*. Editio octava critica maior. Leipzig: Gieseche & Devrient.

BER, 1971: *Български етимологичен речник.* (1971). Том 1. Под ред. на Владимир Георгиев. София: Изд. на БАН.

Ilchev, 1974: Илчев, Ст. (1974). *Речник на редки, остарели и диалектни думи в литературата ни от XIX и XX век.* София: БАН.

Mirchev, 1932: Mirchev, K. (1932). Източномакедонското и родопското "боедин", словенското (n)obeden и западнославянското "žádný – żaden". *Македонски преглед* 8 (2). 9–22.

Mladenov, 1941: Младенов, Ст. (1941). *Етимологически и правописен речник на българския книжовен език.* София: Хр. Г. Данов.

Trubachev, 1975: Трубачев, О. Н. (1975). *Этимологический словарь славянских языков (Праславянский лексический фонд)*. Вып. 2. Москва: Наука.

Filin, 1968: Филин, Ф. (1968). *Словарь русских народных говоров*. Вып. 3. Москва: Наука.

Tzeytlin, 1994: Цейтлин, Р., Вечерка, Р., Благова, Э. *Старославянский словарь (по рукописям X- XI веков).* Москва: Русский язык.

Machine Translation Based on WordNet and Dependency Relations

Luchezar Jackov
Institute for Bulgarian Language
Bulgarian Academy of Sciences
lucho@skycode.com

Abstract

The proposed machine translation (MT) approach uses WordNet (Fellbaum, 1998) as a base for concepts. It identifies the concepts and dependency relations using context-free grammars (CFGs) enriched with features, role markers and dependency markers. Multiple interpretation hypotheses are generated and then are scored using a knowledge base for the dependency relations. The hypothesis with the best score is used for generating the translation. The approach has already been implemented in an MT system for seven languages, namely Bulgarian, English, French, Spanish, Italian, German, and Turkish, and also for Chinese on experimental level.

1. Introduction

Any translation must properly convey the concepts and the relations between them from the source to the target language. This includes correct identification of the concepts (i.e., word sense disambiguation) and correct identification of the relations between them (their dependency relations). These concepts and relations must be properly projected into the target language so that they can be correctly identified (understood) by the recipient of the translation.

The article proposes an approach for generation and semantically driven evaluation of interpretation hypotheses as part of an MT system. The derived hypotheses embed and evaluate the morphological, syntactic and semantic information simultaneously instead of in a pipeline. Recent developments (Bohnet et al., 2013) show the advantages of performing morphological and syntactic analysis jointly, obviating the use of a part-of-speech tagger. Our approach goes further by performing morphological, syntactic and semantic analysis jointly. The best hypotheses are chosen by using a semantic scoring mechanism that works on the relations that each hypothesis identifies. A method for performing parse selections based on semantic knowledge has been proposed in (Fujita et al., 2010).

The article presents work in progress, and no extensive comparison of the translation results has been done yet. However, the proposed MT approach is used in the SkyCode machine translation system. It has been implemented in C++ and has a very compact binary data representation, approx. 60MB for 7 languages and 42 language translation directions. It has been used in offline translation applications for mobile devices, outperforming Google Offline Translator in both quality and size (the latter needs about 1.05GB of data for 7 languages). The system has also participated successfully in the *iTranslate4* project, and can be tested online at http://itranslate4.eu (the SkyCode vendor). The system consists of a lemmatizer, a concept binder, a hypothesis generator, a dependency relations scorer and a synthesis unit.

2. Lemmatizer

The lemmatizer analyzes the smallest bits that the system works on: the tokens. For every token the lemmatizer yields a list of all lemmas that have a word form equal to the token. Each entry in the list consists of the lemma identifier in the database and the morphological features of the word form.

The lemmatizer database consists of entries where each entry holds an identifier, a lemma (or a base form of the word), an inflection group identifier, and a paradigm identifier. Each inflection group is a set of inflection entries consisting of a suffix and its respective features. In this way, all word forms of the lemma are defined. The input word form can be lemmatized with the inflection features extracted, and any word form can be generated by specifying the lemma and the respective features.

The result of applying the lemmatizer over each token is a list of lemma entries. Each entry consists of a lemma identifier and a set of features. The lemma entries list is used by the concept binder to yield initial interpretation hypotheses for the token. For instance, "water" will yield two lemma entries, one for the noun and one for the verb. The Bulgarian surface form of ми (*mi*, "me") will yield an entry for the dative/genitive/possessive form of the personal pronoun аз (*az*, "I") and another two for the second and third person past forms of the verb мия (*miya* "to wash").

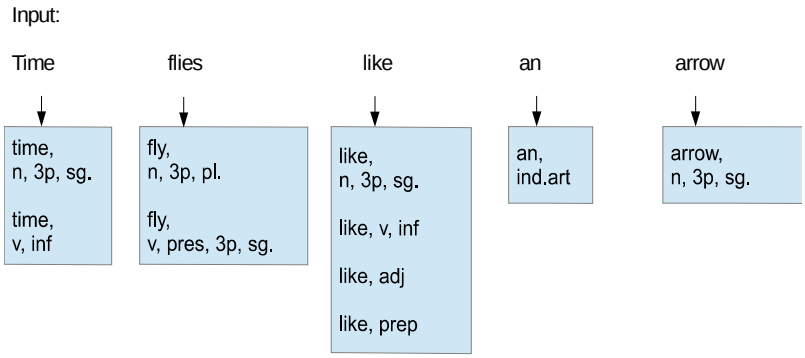


Figure 1: Lists of lemma entries resulting from the application of the lemmatizer over each token. The lemma and language identifiers for each entry are omitted for brevity.

The proposed approach considers every possible lemmatization of the token producing one or more interpretation hypotheses. The lemmatization disambiguation occurs naturally when scoring the different hypotheses and disregarding the low-scored ones. This obviates the use of part-of-speech taggers, which are known to introduce errors that cannot be handled further in the process.

We have developed dictionaries containing 115,735 lemmas for English, 102,393 for Bulgarian, 38,445 for Turkish, 135,171 for German, 68,026 for Spanish, 65,866 for French, and 59,883 for Italian as part of the SkyCode MT system.

3. Concept Binder

The concept binder database links each WordNet concept (its synset identifier) to a list of one or several lemmas that observe agreement restrictions. The database is used by the concept binder to identify concepts in the input language and to generate translations in the output language.

3.1. Database Structure

The concept binder database consists of entries having the following fields:

- a language identifier;
- a base form (a descriptive string, usually matching the base form of the constituting lemmas);
- a hypothesis type identifier (HTI);
- a list of lemma identifiers;
- a WordNet synset identifier;
- restrictions on features of each lemma;
- unification of features of each lemma;
- a list of additional features.

The base form is used only for easy lookup and management of the database.

The hypothesis type identifier (HTI), as used in this article, corresponds to some extent to the non-terminal symbols of a classical CFG. Here are some of the HTIs used in the system: *Verb, Adjective, Noun, Personal_pronoun, Demonstrative_pronoun, Direct_object, Indirect_object, Verb_phrase, Noun_phrase, Prepositional_phrase, Subject_phrase, Sentence*, etc.

The list of lemma identifiers is used for both concept identification and translation generation. The restrictions on features of each lemma allow identifying concepts that are defined by a specific word form and not by all of the word forms, which is usually encountered in multiword expressions (MWEs). The unification of features of each lemma is used for MWEs. The list of additional features is used to define sub-categorization frames, mass/plural count nouns, etc.

The SkyCode MT system currently has 285,171 concept binder entries for English, 166,948 for Bulgarian, 118,832 for Turkish, 213,421 for German, 162,545 for Spanish, 183,479 for French, and 140,836 for Italian. The concept binder data for English has been automatically imported from the Princeton WordNet 3.0, while the rest has been developed independently. Similar resources exist for some of the languages (e.g., Bulgarian – cf. (Koeva, 2010), but they were either not available or not freely accessible when the development of the system started.

3.2. Identifying Concepts

The concept binder works on the lists of lemmatized tokens created by the Lemmatizer. It generates all the possible interpretations for one or more consecutive tokens and the result comprises the initial interpretation hypotheses on which the hypothesis generator works. For instance, running the concept binder over the token “water” (lemmatized to [water, n, English] and [water, v, English]) will yield the following interpretation hypotheses: 6 instances with HTI of “noun” bearing the respective WordNet synset identifiers and 5 instances with HTI of “verb” bearing the respective WordNet synset identifiers.

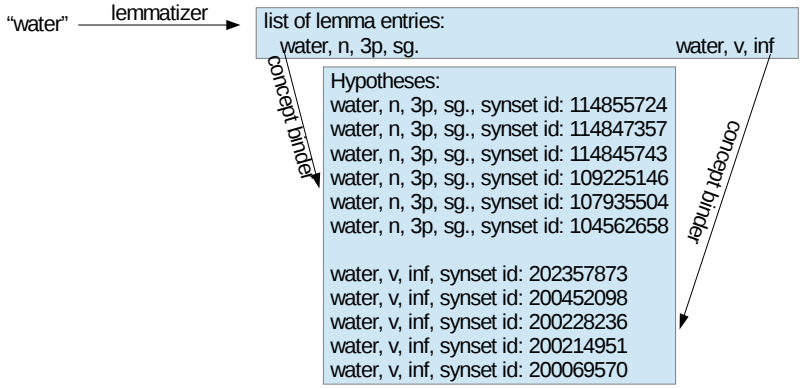


Figure 2: Concept binder being run over the output of the lemmatizer. The lemma, language and concept binder identifiers are omitted for brevity.

The concept binder is run for spans up to 9 tokens to find multiword expressions (such as “guinea pig”) and yield interpretation hypotheses for them. Each hypothesis comprises a particular WordNet concept and a particular projection (translation) of the WordNet concept in the target language if there is more than one translation of the concept.

The hypotheses derived from several language units by the concept binder are considered along with the hypotheses created by applying the rules over the single-lemma hypotheses. For instance, “to kick the bucket” will be considered as a hypothesis for a single concept (“to die”) having HTI of “Verb”. It will also be considered as a hypothesis with HTI of “Verb_phrase” and roles and dependencies identified in concert with the literal meaning of *to kick a bucket*.

3.3. Generating Translations for Concepts

The concept binder database is also used to generate translations in the target language. For each source language concept one or several translations are retrieved from the database by filtering the entries that

match the target *language id* field and the *WordNet synset id* field of the source concept. Each translation is generated by looking for the lemmas in the lemmatizer database and inflecting each of them into the appropriate word form.

4. Hypotheses Generator and Parsing Rules

The hypotheses generator groups hypotheses of adjacent spans of the input text by trying to apply each of the parsing rules (based on enriched CFGs) over them. A parsing rule can be applied if the hypotheses to be grouped meet the parsing rule criteria, thus yielding new interpretation hypotheses for the span that includes the adjacent spans whose hypotheses are grouped. The hypothesis generator (parse generator) uses the Cocke–Younger–Kasami (CYK) algorithm (Cocke et al., 1970; Younger, 1967; Kasami, 1965), modified with scoring and pruning to prevent search space explosion.

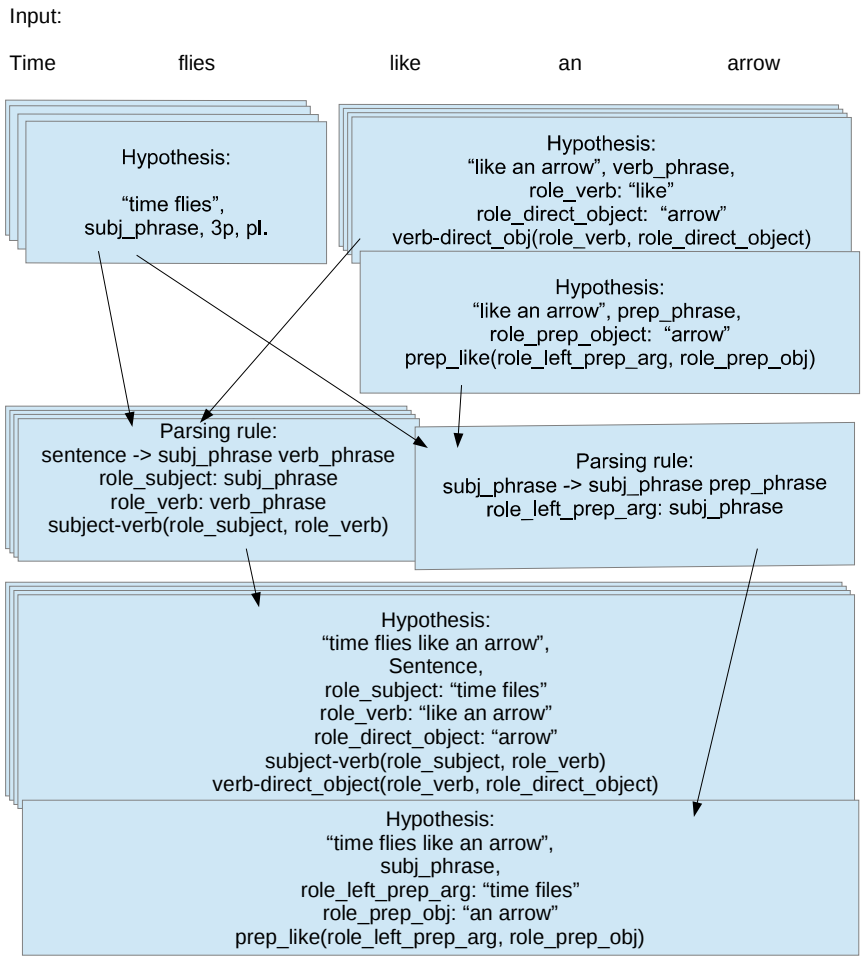


Figure 3: Parsing rules being applied to hypotheses yield hypotheses for broader spans. Even though the illustrated hypotheses seem unlikely for the sample input text, this may not be so for other input text (e.g., “time travels seem an illusion”). The likeliness is evaluated as a hypothesis score by looking up for the identified dependency relations in the knowledge base. Note that Figure 3 shows just one of the possible splits but other splits are also considered, such as the correct one, [S → SP VP (“time”, *subj_phrase*) (“flies like an arrow”, *verb_phrase*)]. When having good knowledge base, the latter hypothesis will receive the best score.

Each parsing rule used by the hypothesis generator assumes roles and dependency relations. The result of the successful application of a parsing rule is a new interpretation hypothesis that includes the assumed roles and dependency relations as part of it. The hypothetical dependency relations between the assumed roles are scored by the dependency relations scorer using the dependency relations knowledge

base. Thus, each interpretation hypothesis is scored and the worst hypotheses are pruned to prevent search space explosion. Currently, the system identifies the following relations and roles (inexhaustive): *subject-verb, verb-direct_obj, verb-indirect_obj, modal_verb-verb, adjective-noun*, etc. It also identifies a number of language-dependent prepositional relations, such as *prep_in(left_prep_argument, prep_object)*.

4.1. Parsing Rules

The parsing rules are the equivalent to the rewriting rules of classical CFG. A classical CFG rewriting rule, when used for analysis, selects or restricts the non-terminals that would build the resulting non-terminal. Unlike CFG, the parsing rules hold data for additional restrictions over the features of the constituent hypotheses. Such restrictions are used to define rules for specific sub-categorizations, agreement rules, etc. The parsing rules are manually developed. Each parsing rule can be either unary or binary. It consists of:

1. A list of one (unary rule) or two (binary rule) entries. Each entry defines the restrictions on the hypothesis that would take the entry position. The following data restricts the candidate hypothesis:
    - A hypothesis type identifier (HTI);
    - A list of restrictions over the features of the hypothesis;

Example: *VP → V NP should be restricted only for transitive verbs. Such verbs have the "transitive"feature defined in the concept binder. The restriction for transitiveness is in this list.*

    - A list of features being inherited (i.e., feature unification data);

Example: *In composite past tenses in Bulgarian, the auxiliary verb does not have a gender feature, but has person and number features. The past participle has gender and number features. Gender is inherited from the past participle, while person and number features are inherited from the auxiliary verb. The resulting hypothesis has unified gender, number, and person features that will be used later to account for the subject-verb agreement on these features.*

    - Role markers: *(e.g., role_subject, role_verb, role_direct_object, role_indirect_object, role_prep_object, role_left_prep_argument)*;
    - A list of role markers (to be inherited).

Complex interpretation hypotheses may identify more than one role. When grouping such hypotheses, the parsing rule inherits the pointers to the role markers from the hypotheses that are being grouped.

Example: *A unary rule for the preposition "in" introduces the relation prep_in(role_left_prep_argument, role_prep_object). Another parsing rule groups the preposition hypothesis with a noun phrase hypothesis and sets its role to prep_object to yield a prepositional phrase hypothesis. This hypothesis carries the preposition_role and the prep_object role pointing to the particular concepts within the hypothesis. Another rule binds a noun phrase to the prepositional phrase. This rule inherits the role pointers to the preposition and to the prepositional object.*

    - A list of features that the hypothesis should agree with any of the roles that the parsing rule identifies.

Example: *In S → NP VP the verb phrase should agree with the subject noun phrase. The rule marks the first entry (the NP) with "role_subject" and defines that the feature list [gender, person, number] of the second entry should agree with "role_subject". If the agreement is not met, the rule is not applied.*

2. A list of dependency relations where each entry holds:
    - A relation identifier;
    - Role markers for the first argument and for the second argument.

Example: *A parsing rule for a subject phrase with a verb phrase subcategorized for possession. The rule introduces a possession dependency relation between the subject and the direct object. A general rule for non-possession verbs would introduce only the subject-verb and verb-direct_obj relations.*

3. Resulting HTI and features

Example: *VP → V NP will have HTI of "V" for the first entry, HTI of "NP" for the second entry, and a resulting HTI of "S". The resulting features are used to add information on what the parse tree lying under the hypothesis contains. For instance, a verb phrase with that-clause is unlikely to be bound to a*

prepositional phrase. This can be described by having a resulting feature "+that-clause" on the VP → V CP rule, and having a "not(+that-clause)" restriction on the VP → VP PP rule.

4. A list of languages that the parsing rule can be applied on.
5. A list of languages that the rule can be used to translate into.
6. Rule score that is added to the total hypothesis score.

Example 1: *Rules that handle commonly encountered but grammatically incorrect constructions.*
Example 2: *Rules that handle inverse word order in free word order languages. Such rules are defined with a lower score, giving precedence to the rules that would handle the canonical word order.*

A parsing rule is applied to adjacent interpretation hypotheses if they obey the feature and agreement restrictions. When the feature and agreement restriction lists are empty, the rule will not apply any feature restrictions.

The data structure holding each newly yielded interpretation hypothesis preserves pointers to its constituents, the rule that has been applied, the roles that have been identified, and the dependency relations that have been introduced, so that the hypothesis can be scored.

Example: *(DO → NP) a unary rule for a noun in accusative case (for case languages) that generates a new hypothesis with HTI of "Direct_object".*
Example: *(DO → Ppr) a unary rule for a personal pronoun in accusative case (e.g., in Bulgarian) that generates a new hypothesis with HTI of "Direct_Object".*
Example: *(VP → Vtr DO) a binary rule that binds a transitive verb with the direct object. The first entry has the following data:*

- HTI is "Verb".
- It must have a sub-categorization feature "transitive_verb".
- Its role marker is set to "role_verb".

The second entry has the following data:

- HTI is "Direct_object".
- Its role marker is set to "role_Direct_object".

The parsing rule introduces the dependency relation verb-direct_obj (role_verb, role_direct_object). It will group hypotheses with HTI of "Verb" with hypotheses having HTI of "Direct_Object" to yield a new hypothesis with HTI of "Verb_Phrase" (Verb_Phrase → Verb Direct_Object). This parsing rule is common for English, German, Spanish, French, Italian, and Bulgarian. This specific rule will not cover all cases for all languages, as the direct object can stand before the verb in German, and in Bulgarian for cases where a pronoun is the direct object.

There are 5,598 parsing rules, of which 2,085 rules are shared by more than one language.

4.2. Hypothesis Generator

The hypothesis generator is a modified version of the CYK algorithm. Given a list of language units from 1 to n, it sequentially derives hypotheses for spans starting from 1 and having length of 1, then, length of 2, then length of 3, and so on to length of n-1 by applying the parsing rules on every possible split of the span being considered. Each interpretation hypothesis for each span is stored in a three-dimensional array where the first index denotes the span start, the second index denotes the span length, and the third index denotes the hypothesis position in the hypotheses list.

4.2.1. General Algorithm Description

Let's assume that the input text is "Time flies like an arrow". The hypothesis generator will first derive interpretation hypotheses for span of length 1 starting at position 1 ([1,1]), i.e., for the token "time" by running the concept binder over the lemmatizer output of "time". Then it will derive hypotheses for "Time flies" by first deriving hypotheses for "flies", i.e., span of length 1 starting at position 2 ([2,1]). Then it will try to apply parsing rules over the two spans [1,1] and [2,1], yielding hypotheses for span [1,2] ("time flies"). It will continue by deriving hypotheses for span [3,1] ("like"), [2,2] ("flies like"), [1, 3] ("time flies like"). Eventually it will generate hypotheses for the span [1,4] ("Time flies like an arrow").

Multiple hypotheses are derived for each span (see Figure 3). For instance, “flies” is the third person singular present form of the verb “fly”, but it is also the plural of the noun “fly”. The verb “fly” has 14 WordNet senses and for each sense the concept binder yields an interpretation hypothesis. Each hypothesis holds particular bindings to the WordNet concepts and the presumed relations between them, which makes it possible for the dependency scorer to look up the dependency relation instances in the knowledge base.

4.2.2. Application of the Parsing Rules

The parsing rules are applied on adjacent spans by trying to apply each parsing rule over the Cartesian product of the hypotheses for the two spans. Let's assume that “flies” yields two hypotheses, one as a noun and one as a verb. Let's have two parsing rules, S → NP VP, NP → N N. Applying the parsing rules over the two fragments [(“time”, N)] and [(“flies”, N), (“flies”,V)] will yield [(“time flies”, S), (“time flies”, NP)]. Even though the second hypothesis is unacceptable from a semantic point of view, it is a legitimate syntactic parse and a legitimate hypothesis. However, this hypothesis will receive a low score and will eventually be pruned, since the hypothesized dependencies between the hypothesized concepts do not have a match in the dependency relations knowledge base.

4.2.3. Telling the Good Hypotheses from the Bad Ones

The data structure behind each interpretation hypothesis stores the roles and the dependency relations identified as part of the hypothesis. Each dependency relation that has its arguments (role markers) bound to particular concepts, is scored by the dependency relations scorer.

5. Dependency Relations Knowledge Base and Scoring

The dependency relations knowledge base consists of quadruples containing a relation identifier, two concept identifiers for the relation arguments, and scoring weight. The weight can be positive or negative. The scorer evaluates each hypothesis by looking in the knowledge database for all of the dependency relations between the particular concepts that the hypothesis has identified and summing the weights, thus forming the hypothesis score.

By applying the parsing rules, the hypothesis generator defines particular dependency relations between the concepts of each generated interpretation hypothesis. For instance, it hypothesizes the relation *subject-verb(time, fly)* for the hypothesis (“time flies”, S), and *attrib_english(time, fly)* for the hypothesis (“time flies”, NP). The knowledge base consists of entries giving scores for such instances (e.g., *relation*(subject-verb, time, fly) = 1, *relation*(attrib_english, time, fly) = 0). The scorer looks up for the particular dependency relations entries in the knowledge base and adds the entry score to the hypothesis score whenever it finds a matching entry. Thus, each hypothesis receives a score, and low-scored hypotheses are pruned to prevent search space explosion.

5.1. Knowledge Base over WordNet Synsets

Having a knowledge base over WordNet synsets allows reusing it for analyzing different languages that have WordNets bound to the Priceton WordNet synsets. Each knowledge base entry consists of a relation identifier, two synset identifiers, and relation score (usually 0, 1 or -1). Each hypothesis has a number of hypothesized relations, namely a relation identifier and two concepts (i.e., two synset identifiers). For each such relation instance, the scorer looks up for matches of the triple (*rel_id, synset_id1, synset_id2*) in the knowledge base and adds the resulting score to the hypothesis score.

5.2. Knowledge Base over Lemmas

Having a knowledge base over lemmas allows making fine distinctions between members of the same WordNet synset in the translation synthesis. Each knowledge base entry consists of a relation identifier, two concept binder base forms and relation score. For each relation identified by a given hypothesis, the scorer retrieves the concept binder base forms of the translated concepts and forms a triple having (*rel_id, arg1_base_form, arg2_base_form)*. The scorer looks up for matches of this triple and adds the resulting score to the hypothesis score.

5.3. Data Sparseness

The main challenge to the proposed system is the data sparseness of the dependency relations knowledge base. One way of overcoming this challenge is to use the WordNet hypernym relations and manually populate relation instances over hypernyms. For instance, the relation *verb-direct_obj(play, musical instrument)* can yield the same relation for the "musical instrument" hyponyms. Unfortunately, this approach is not productive enough.

Another way is to use the MT system for automatic collection of lemma-based dependency relations knowledge from monolingual corpora. This can be achieved by translating sentences of the corpora and recording the dependency relations over the particular source language lemmas identified by the best interpretation hypothesis. This data can be used in the hypothesis scoring by using the translated concepts as relation arguments when looking up the lemma-based knowledge base. The data can be used to infer relation instances between WordNet synsets by running the system over a set of several languages (e.g., English, French, Spanish, German, Italian, and Bulgarian) and picking the most complete synset clusters.

6. Translation Synthesis

Each hypothesis is a parse tree consisting either of sub-trees or of concept binder entries. Creating a translation of the hypothesis includes constituent reordering, various agreements, etc. for each parsing rule. There is a set of synthesis rules for each parsing rule that takes care of word reordering, insertion, deletion, etc. when creating the translation output. The rules are manually written. For instance, when translating "I gave him the book", the hypothesis generator identifies the structure [I (*subj*) [[gave him (*verb-ind_obj*)] the book (*v_ind_obj-dir_obj*)]. When translating it into Bulgarian, there is a synthesis rule for the *verb-ind_obj* rule that checks whether the indirect object is a pronoun, whether the rest of the translation has a missing subject, or whether it is negative, to achieve the correct word order:

I gave him the book. = Дадох му книгата. (Dadoh mu knigata)
I gave John the book. = Дадох книгата на Джон (Dadoh knigata na Dzhon)
John gave him the book. = Джон му даде книгата. (Dzhon mu dade knigata)
I haven't given him the book. = Не му дадох книгата. (Ne mu dadoh knigata)

The leaves of the hypothesis parse tree are concept binder entries and are translated by looking up the concept binder database for entries that match the source concept synset in the target language and inflecting them (see 3.3).

More than one synthesis rule can be defined for each parsing rule. The competing synthesis rules add language-specific relation dependencies that are also scored by the Dependency relations scorer.

Example: *A noun phrase with a simple prepositional phrase can be expressed in English as an attributive, e.g.,"months of spring" and "spring months". There are two competing synthesis rules, one introducing prep_of relation and the other introducing attrib_english relation. The rule that gets the higher score is chosen over the other rule.*

7. Conclusion

The article provides an overview of a machine translation system based on WordNet and dependency relations. There is a working prototype of this system implemented in C++ for seven languages (42 language directions): English, French, German, Spanish, Italian, Turkish, and Bulgarian that can be tested online at `http://itranslate4.eu` (SkyCode translation vendor).

One of the main challenges to the proposed system is the populating of the knowledge base and mitigating the data sparseness. There are several approaches to overcome this.

One approach involves manual population of the knowledge base; it has proven to yield very good results in terms of parsing accuracy for any given sentence. This is further improved by populating relations over the WorNet concepts (hypernyms) and the Dependency scorer is modified to look for relations between the hypernyms of the arguments when no direct match is found.

Another approach includes automatic collection of relation instances over lemmas. The system produces scored hypotheses with dependency relations over the lemmas. The best hypothesis can be used

to populate language-specific lemma-based knowledge base. This knowledge base can be reused when translating into the language that the knowledge base is for. Running the system over the Europarl corpus yielded some 33 million knowledge entries for six languages (English, French, German, Italian, Spanish, and Bulgarian).

A third approach employs automatic derivation of WordNet-based dependency relations by picking a lemma-based relation, generating all possible WordNet-based hypotheses, and choosing the one that is most consistent with the lemma knowledge base in different languages and WordNet synonyms. A test version of the relation inference module over the 33 million lemma-based knowledge entries yielded some 1.32 million synset-based knowledge entries.

Acknowledgements

The present paper was partially prepared within the project *Integrating New Practices and Knowledge in Undergraduate and Graduate Courses in Computational Linguistics* (BG051PO001-3.3.06-0022) implemented with the financial support of the Human Resources Development Operational Programme 2007 – 2013 co-financed by the European Social Fund of the European Union. The author takes full responsibility for the content of the present paper.

References

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Bohnet, B., Nivre, J., Boguslavsky, I., Farkas, R., Ginter, F., and Hajic, J. (2013). Joint Morphological and Syntactic Analysis for Richly Inflected Languages. *Transactions of the Association for Computational Linguistics*, 1:415–428.

Fujita, S., Bond, F., Oepen, S., and Tanaka, T. (2010). Exploiting Semantic Information for HPSG Parse Selection. *Research on Language and Computation.* Online publication date: 20-Oct-2010.

Koeva, S. (2010). Bulgarian Wordnet – Current State, Applications and Prospects. In *Bulgarian-American Dialogues,* pages 120-132. Sofia: Prof. M. Drinov Academic Publishing House.

Cocke, J. and Schwartz, J. T. (1970). *Programming Languages and Their Compilers: Preliminary Notes. Technical report.* Courant Institute of Mathematical Sciences, New York University.

Kasami, T. (1965). *An Efficient Recognition and Syntax-analysis Algorithm for Context-free Languages*. Scientific report AFCRL-65-758. Bedford, MA: Air Force Cambridge Research Lab.

Younger, D. H. (1967). Recognition and Parsing of Context-free Languages in Time n3. *Information and Control,* 10 (2):189–208.

Recognize the Generality Relation between Sentences using Asymmetric Association Measures

Sebastião Pais
MINES ParisTech
Centre de Recherche en Informatique
77305 Fontainebleau, France
pais@cri.ensmp.fr

Gaël Dias and **Rumen Moraliyski**
Normandie University
UNICAEN, GREYC CNRS
F-14032 Caen, France
firstname.lastname@unicaen.fr

Abstract

In this paper we focus on a particular case of entailment, namely entailment by generality. We argue that there exist various types of implication, a range of different levels of entailment reasoning, based on lexical, syntactic, logical and common sense clues, at different levels of difficulty. We introduce the paradigm of Textual Entailment (TE) by Generality, which can be defined as the entailment from a specific statement towards a relatively more general statement. In this context, the Text T entails the Hypothesis H, and at the same time H is more general than T. We propose an unsupervised and language-independent method to recognize TE by Generality given a case of $Text - Hypothesis$ or $T - H$ where entailment relation holds.

1. Introduction

We introduce the paradigm of TE by Generality, which can be defined as the entailment from a specific sentence towards a more general sentence. For example, from sentences (1) and (2) extracted from RTE-1, we would easily state that $(1) \rightarrow (2)$ as their meaning is roughly the same and sentence (2) is more general than sentence (1).

(1) Mexico City has a very bad pollution problem because the mountains around the city act as walls and block in dust and smog.

(2) Poor air circulation out of the mountain-walled Mexico City aggravates pollution.

To understand how TE by Generality can be modeled for two sentences, we propose a new paradigm based on the Asymmetric InfoSimba Similarity (AIS) measure. Instead of relying on the exact matches of words between texts, we propose that one sentence entails the other one in terms of generality if two constraints hold: (a) if and only if many of the words in T are semantically similar to the words that make H, and (b) if most of the words of H are more general than the words of T. As far as we know, we are the first to propose an unsupervised, language-independent, threshold free methodology in the context of TE by Generality, although the approach of Glickman and Dagan (2005) is based on similar assumptions. This new proposal is exhaustively evaluated against the first five RTE datasets. In particular, the RTE-1 is the only dataset for which there exist comparable results with linguistic-free methodologies (Glickman and Dagan, 2005; Perez et al., 2005; Bayer et al., 2005).

In this paper we hypothesize the existence of a special mode of TE, namely TE by Generality. Thus, the main contribution of our study is to highlight the importance of this inference mechanism.

2. Variants of the Entailment

Pazienza et al. (2005) define three types of entailment:

1. *Semantic Subsumption* - T and H express the same fact, but the situation described in T is more specific than the situation in H. The specificity of T is expressed through one or more semantic operations. For example, in the sentential pair:

   - H: The cat eats the mouse. | T: The cat devours the mouse.

   T is more specific than H, as eat is a semantic generalization of devour.

2. *Syntactic Subsumption* - T and H express the same fact, but the situation described in T is more specific than the situation in H. The specificity of T is expressed through one or more syntactic operations. For example, in the pair:

   - H: The cat eats the mouse. | T: The cat eats the mouse in the garden.

   T contains a modifying prepositional phrase.

3. *Direct Implication* - H expresses a fact that is implied by a fact in T. For example:

   - H:The cat killed the mouse. | T: The cat devours the mouse.

   H is implied by T, as it is supposed that killed is a precondition for devour. In Dagan and Glickman (2004) syntactic subsumption roughly corresponds to the restrictive extension rule, while direct implication and semantic subsumption correspond to the axiom rule.

We want to regard entailment by generality as a relation between utterances (that is, sentences in context), where the context is relevant to understand the meaning. In relation to the classification proposed by Pazienza et al. (2005), entailment by generality is comparable to *Semantic Subsumption* kind of TE. Thus, Entailment by Generality can be defined as the entailment from specific sentence towards a more general sentence.

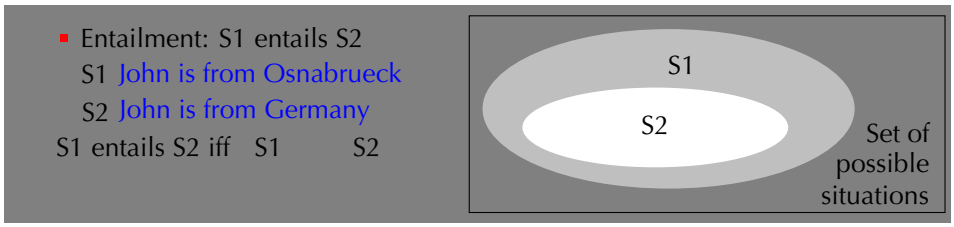


Figure 1: Venn diagram: entailment by generality.

2.1. Context Textual Entailment

Within TE framework, a text T is said to entail a textual hypothesis H if the truth of H can be inferred from T. This means that most people would agree that the meaning of T implies that of H. Somewhat more formally, we say that T entails H when some representation of H can be "matched" with some (or part of a) representation of T, at some level of granularity and abstraction.

Dagan and Glickman (2004) define TE as a relationship between a coherent textual fragment T and a language expression, which is considered as a hypothesis H. Entailment holds (i. e. $T \rightarrow H$) if the meaning of H can be inferred from the meaning of T, as interpreted by a typical language user. This relationship is directional and asymmetric since the meaning of one expression may usually entail the other while entailment in the other direction is less certain.

For instance, a Question Answering (QA) system has to identify texts that entail the expected answer. Given the question "Who painted the Mona Lisa?", the text "Among the works created by Leonardo da Vinci in the 16th century is the small portrait known as the Mona Lisa or la 'Gioconda'", entails the expected answer "Leonardo da Vinci painted the Mona Lisa". Similarly, in Information Retrieval (IR) relevant documents should entail the combination of semantic concepts and relations denoted by the query. In Information Extraction (IE), entailment holds between different text variants expressing

the same target relation (Romano et al., 2006). In text summarization, an important processing stage is sentence extraction, which identifies the most important sentences of the texts to be summarized; especially when generating a single summary from several documents (Barzilay and McKeown, 2005), it is important to avoid selecting sentences that convey the same information as other sentences that have already been selected, i.e. ones that entail such sentences.

3. Recognizing Textual Entailment

Basically, Recognizing Textual Entailment (RTE) is the task of deciding, given two text fragments, whether the meaning of one of the texts is entailed (can be inferred) from the other text. Also, this task captures generically a broad range of inferences that are relevant for multiple applications. A necessary step in transforming textual entailment from a theoretical idea into an active empirical research field was the introduction of benchmarks and an evaluation forum for entailment systems.

3.1. Unsupervised and Language-Independent Methodologies

Different approaches have been proposed to recognize Textual Entailment: from unsupervised language-independent methodologies (Glickman and Dagan, 2005; Perez et al., 2005; Bayer et al., 2005) to deep linguistic analysis. We will particularly detail the unsupervised language-independent approaches, to which our work can be directly compared, at least to a certain extent.

One of the most simple proposals (Perez et al., 2005) explores the *BLEU algorithm* (Papineni et al., 2002). First, for several values of n (typically from 1 to 4), they calculate the percentage of n-grams from the text T, which appear in the hypothesis H. The frequency of each n-gram is limited to the maximum frequency with which it appears in any text T. Then, they combine the marks obtained for each value of n as a weighted linear average and finally apply a brevity factor to penalize short texts T. The output of BLEU is then taken as the confidence score. Finally, they perform an optimization procedure to choose the best threshold according to the percentage of success of correctly recognized entailment. This procedure achieves 0.495 accuracy in recognizing TE.

In Bayer et al. (2005) the entailment data is treated as an aligned translation corpus. In particular, they use the *GIZA++* toolkit (Och and Ney, 2003) to induce alignment models. However, the alignment scores alone were next to useless for the RTE-1 development data, predicting entailment correctly only slightly above chance. As a consequence, they introduced a combination of metrics intended to measure translation quality. Finally, they combined all the alignment information and string metrics with the classical K Nearest Neighbors (K-NN) classifier to choose for each test pair the dominant truth value among the five nearest neighbors in the development set. This method achieves 0.586 accuracy.

The most interesting work is certainly the one described in Glickman and Dagan (2005), who propose a general probabilistic setting that formalizes the notion of TE. Here, they focus on identifying when the lexical elements of a textual hypothesis H are inferred from a given text T. The probability of lexical entailment is derived from Equation 1 where $hits(.,.)$ is a function that returns the number of documents containing its arguments.

$$P(H|T) = \prod_{u \in H} max_{v \in T} \frac{hits(u,v)}{hits(v)} \tag{1}$$

The text and hypothesis of all pairs in the development and test sets were tokenized and stop words were removed to empirically tune a decision threshold, λ. Thus, for a pair $T-H$, they tagged an example as true (i.e. entailment holds) if $P(H|T) > \lambda$, and as false otherwise. The threshold was empirically set to 0.005. With this method accuracy of 0.586 is achieved. The best results from these three approaches are obtained by Glickman and Dagan (2005), who introduce the notion of asymmetry within their model. The underlying idea is based on the fact that for each word in H the best asymmetrically co-occurring word in T is chosen to evaluate $P(H|T)$. Although all three approaches show interesting properties, they all depend on tuned thresholds, which can not reliably be reproduced and need to be changed for each new application. Moreover, they need training data, which may not be available. Our idea aims at generalizing the hypothesis made by Glickman and Dagan (2005).

4. Asymmetric Word Similarities

Two different types of knowledge can be acquired depending on the basic textual unit under study. On the one hand, analyzing word similarities evidences intrinsic knowledge about the language (i.e. information about the language which is not explicitly encoded in texts). Traditional examples are collocations and word semantic relations such as hypernymy/hyponymy, meronymy/holonymy, synonymy or antonymy, which must be mined from texts. On the other hand, explicit knowledge about the language (i.e. information about the message conveyed by the texts) can be extracted from the evaluation of sentence, passage and text similarities[1]. There are obviously some exceptions.

4.1. Asymmetric Association Measures (AAMs)

In order to stay within the domain of language-independent and unsupervised methodologies, a number of asymmetric association measures have been proposed (Pecina and Schlesinger, 2006; Tan et al., 2004) and applied to the problems of taxonomy construction (Sanderson and Croft, 1999; Cleuziou et al., 2010), cognitive psycholinguistics (Michelbacher et al., 2007) and general-specific word order induction (Dias et al., 2008). Sanderson and Croft (1999) is certainly one of the first studies to propose the use of the conditional probability for taxonomy construction.

They assume that a term t_2 subsumes a term t_1 if the documents in which t_1 occurs are a subset of the documents in which t_2 occurs constrained by $P(t_2|t_1) \geq 0.8$ and $P(t_1|t_2) < 1$. By gathering all subsumption relations, they build the semantic structure of any domain, which corresponds to a directed acyclic graph. In Sanderson and Lawrie (2000), the subsumption relation is indicated by the following expressions $P(t_2|t_1) \geq P(t_1|t_2)$ and $P(t_2|t_1) > t$ where t is a given threshold and all term pairs found to have a subsumption relationship are passed through a transitivity module, which removes extraneous subsumption relationships in the way that transitivity is preferred over direct pathways, thus leading to a non-triangular directed acyclic graph.

Eight of the AAMs used in that work will be evaluated in the context of asymmetric similarity between sentences: the Added Value (Equation 2), the Braun-Blanket (Equation 3), the Certainty Factor (Equation 4), the Conviction (Equation 5), the Gini Index (Equation 6), the J-measure (Equation 7), the Laplace (Equation 8) and the Conditional Probability (Equation 9).

$$AV(x\|y) = P(x|y) - P(x). \quad (2) \qquad BB(x\|y) = \frac{f(x,y)}{f(x,y) + f(\bar{x},y)}. \quad (3)$$

$$CF(x\|y) = \frac{P(x|y) - P(x)}{1 - P(x)}. \quad (4) \qquad CO(x\|y) = \frac{P(x) \times P(\bar{y})}{P(x,\bar{y})}. \quad (5)$$

$$GI(x\|y) = P(y) \times (P(x|y)^2 + P(\bar{x}|y)^2) - P(x)^2 \times P(\bar{y}) \times (P(x|\bar{y})^2 + P(\bar{x}|\bar{y})^2) - P(\bar{x})^2. \quad (6)$$

$$JM(x\|y) = P(x,y) \times \log \frac{P(x|y)}{P(x)} + P(\bar{x},y) \times \log \frac{P(\bar{x}|y)}{P(\bar{x})}. \quad (7)$$

$$LP(x\|y) = \frac{N \times P(x,y) + 1}{N \times P(y) + 2} \quad (8) \qquad P(x|y) = \frac{P(x,y)}{P(y)} \quad (9)$$

4.2. Asymmetric Attributional Word Similarities

The InfoSimba (IS) aims to measure the correlations between all the pairs of words in two word context vectors instead of just relying on their exact match as with the cosine similarity measure. Further, IS guarantees to catch similarity between pairs of words even when they do not share contexts, for example due to data sparseness. IS takes under account the fraction of similar contexts instead. It is defined in Equation 10 where $S(.,.)$ is any symmetric similarity measure and each W_{ik} corresponds to the attribute word at the k^{th} position in the vector X_i, p and q are the lengths of the vectors X_i and X_j respectively.

[1] From now on, we will refer to sentences, passages and texts simply as texts.

$$IS(X_i, X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{q} X_{ik} \times X_{jl} \times S(W_{ik}, W_{jl})}{\left(\begin{array}{l} \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik} \times X_{il} \times S(W_{ik}, W_{il}) + \\ \sum_{k=1}^{q}\sum_{l=1}^{q} X_{jk} \times X_{jl} \times S(W_{jk}, W_{jl}) - \\ \sum_{k=1}^{p}\sum_{l=1}^{q} X_{ik} \times X_{jl} \times S(W_{ik}, W_{jl}) \end{array}\right)}. \quad (10)$$

Although there are many asymmetric similarity measures, they evidence problems that may reduce their utility. On the one hand, asymmetric association measures can only evaluate the generality/specificity relation between words that are known to be in a semantic relation (Sanderson and Croft, 1999; Dias et al., 2008). Indeed, they generally capture the direction of association between two words based on document contexts and only take into account a loose semantic proximity between words. For example, it is highly probable to find that *Apple* is more general than *iPad*, which can not be considered to be an hypernymy/hyponymy or meronymy/holonymy relation. On the other hand, asymmetric attributional word similarities only take into account common contexts to assess the degree of asymmetric relatedness between two words. To leverage these issues, we propose the Asymmetric InfoSimba (AIS) measure whose underlying idea is to say that one word x is semantically related to word y and x is more general than y if x and y share as many similar contexts as possible and each context word of x is likely to be more general than most of the context words of y. The AIS is defined in Equation 11, where $AS(.\|.)$ is any asymmetric similarity measure, likewise for the IS in Equation 10 where $S(.,.)$ stands for any symmetric similarity measure. We also define its simplified version $AISs(.\|.)$ in Equation 12.

$$AIS(X_i\|X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{q} X_{ik} \times X_{jl} \times AS(W_{ik}\|W_{jl})}{\left(\begin{array}{l} \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik} \times X_{il} \times AS(W_{ik}\|W_{il}) + \\ \sum_{k=1}^{q}\sum_{l=1}^{q} X_{jk} \times X_{jl} \times AS(W_{jk}\|W_{jl}) - \\ \sum_{k=1}^{p}\sum_{l=1}^{q} X_{ik} \times X_{jl} \times AS(W_{ik}\|W_{jl}) \end{array}\right)}. \quad (11)$$

$$AISs(X_i\|X_j) = \sum_{k=1}^{p}\sum_{l=1}^{q} X_{ik} \times X_{jl} \times AS(W_{ik}\|W_{jl}). \quad (12)$$

5. Asymmetry between Sentences

A number of ways to compute the similarity between two sentences were proposed in the literature. Most similarity measures determine the distance between two vectors associated to two sentences (i.e. the vector space model). However, when applying the classical similarity measures between two sentences, only the identical indexes of the row vector X_i and X_j are taken into account, which may result in misleading values. To deal with this problem, different methodologies have been proposed, but the most promising one is certainly the one proposed by Dias et al. (2007), the InfoSimba informative similarity measure, expressed in Equation 10.

Although there exsist many asymmetric similarity measures between words, there does not exist any attributional similarity measure capable to assess whether a sentence is more specific/general than another one. To overcome this issue, we introduce the asymmetric InfoSimba similarity measure (AIS), which underlying idea is to say that a sentence T is semantically related to sentence H and H is more general than T, if H and T have many related words in common and each word of H is likely to be more general than most of the words of T. The AIS is defined in Equation 11.

As AIS is computationally expensive, we also define its simplified version $AISs(.\|.)$ in Equation 12, which we will specifically use in our experiments.

As a consequence, entailment by generality ($T \overset{G}{\rightarrow} H$) will hold if and only if

$$AISs(T\|H) < AISs(H\|T).$$

Due to its asymmetric definition, in contrast to existing methodologies, we do not need to define or tune thresholds.

6. Three Levels of Pre-Processing

We consider three approaches for selecting the words for the calculation of the asymmetry between sentences. Thus, we can assess which approach performs best to identify entailment by generality. In the first approach, we chose to do the calculations without preprocessing, i.e., do the calculations with all the words. The next approach was to use a list of Stop Words[2].

Finally, in the last approach, we used the Software for the Extraction of N-ary Textual Associations (SENTA) (Dias et al., 1999), in order to extract important Multiword Units (MWU). This system is parameter free and language independent, thus allowing to extract MWU from raw text.

In summary, our experiments are based on three approaches to the calculations to which we refer bellow as *With All Words*, *Without Stop Words* and *With MWU*.

7. Evaluation

In order to evaluate our methodology against well known test data used to compare a number of methodologies our evaluation is based on analysis of Confusion Matrix and values calculated from it. An important performance measure is classification Accuracy (AC) and Precision (P). More specifically, in our work we used the following performance measures – *Average Accuracy*, *Average Precision* and *Weighted Average Accuracy*, *Weighted Average Precision*. Although the obtained results are not excellent, they are promising and encouraging.

Averages ACCURACY by RTE Challenges — Measures versus Approach			
AAM	**Arithmetic Average by Approach**		
	With All Words	**Without Stop Words**	**With MWU**
ADDED VALUE	0.54	0.52	0.53
BRAUN-BLANKET	**0.55**	**0.53**	**0.54**
CERTAINTY FACTOR	0.53	**0.53**	0.53
CONDITIONAL PROBABILITY	0.53	**0.53**	0.53
CONVICTION	0.52	0.51	0.51
GINI INDEX	0.54	0.51	0.53
J-MEASURE	0.53	0.51	0.52
LAPLACE	0.53	**0.53**	0.53
AAM	**Weighted Average by Approach**		
	With All Words	**Without Stop Words**	**With MWU**
ADDED VALUE	0.53	0.52	0.53
BRAUN-BLANKET	**0.54**	0.52	**0.56**
CERTAINTY FACTOR	0.53	**0.53**	0.52
CONDITIONAL PROBABILITY	0.53	**0.53**	0.53
CONVICTION	0.52	0.50	0.51
GINI INDEX	0.53	0.52	0.53
J-MEASURE	**0.54**	0.51	0.52
LAPLACE	0.52	0.52	**0.56**

Table 1: Accuracy Averages | Measures versus Approach

Regarding the **Arithmetic Average** (Table 1), the combination that has the best performance is the *Braun-Blanket* measure on *All Words*. Best **Weighted Average** is achieved on *With WMU* approach by *Braun-Blanket* and *Laplace* measures. Overall, the worst result was obtained with the measure *Conviction* in the approach *Without Stop Words*.

Accuracy values of our experiments on RTE Challenges span a relatively short range between 0.50 and 0.56.

[2]Obtained using `http://www.microsoft.com/en-us/download/confirmation.aspx?id=10024` [Last access: 14th December, 2013]

Average PRECISION - ENTAILMENT by RTE Challenges — Measures versus Approach			
AAM	**Arithmetic Average by Approach**		
	With All Words	**Without Stop Words**	**With MWU**
ADDED VALUE	0.66	0.65	**0.78**
BRAUN-BLANKET	0.53	0.60	0.63
CERTAINTY FACTOR	0.63	0.62	0.46
CONDITIONAL PROBABILITY	0.64	0.64	0.48
CONVICTION	0.60	0.54	0.54
GINI INDEX	0.67	0.55	0.53
J-MEASURE	**0.81**	**0.75**	0.63
LAPLACE	0.65	0.64	0.47
AAM	**Weighted Average by Approach**		
	With All Words	**Without Stop Words**	**With MWU**
ADDED VALUE	0.60	0.59	**0.74**
BRAUN-BLANKET	0.49	0.54	0.58
CERTAINTY FACTOR	0.58	0.57	0.46
CONDITIONAL PROBABILITY	0.58	0.58	0.48
CONVICTION	0.59	0.53	0.53
GINI INDEX	0.62	0.52	0.53
J-MEASURE	**0.73**	**0.66**	0.63
LAPLACE	0.59	0.57	0.48

Table 2: PRECISION – ENTAILMENT Averages | Measures versus Approach

Table 1 points out the approach *Without Stop Words* as the one with worst performance in terms of accuracy, while *All Words* achieves slightly better accuracy compared to *With MWU*.

In Table 2, the combination with the best performance on the **Arithmetic Average Precision** is the *J-measure* with approach *All Words*. For the **Weighted Average Precision**, the *Added Value* shows the best result *With MWU*. The worst result is obtained with the measure *Certainty Factor With MWU* – 0.46.

With respect to the **Precision – Entailment** criterion, the approach that achieves the best results is *With All Words*.

In contrast to the results for **Precision – Entailment**, our method shows unsatisfactory behavior when considered from the perspective of **Precision – No Entailment** (see Table 3). For **Arithmetic Average** the best combination is *Certainty Factor*, *Conditional Probability* and *Laplace With MWU*. For **Weighted Average**, *Laplace* has the best performance *With MWU* approach. Note the low results obtained by the *J-measure* and *Added Value*. In Table 3 the approach with the best performance is *With MWU*, and the worst performing approach is *Without Stop Words*.

After an exhaustive analysis of the results obtained, we can compare our results with the results of the methodologies presented in Section 3.1. Precisely, Bayer et al. (2005), Glickman and Dagan (2005) and Perez et al. (2005) obtained accuracy of 0.586, 0.586 and 0.495, respectively. We prove that our methodology has better performance compared to what was possible in previous works. On RTE-1 Challenge *With MWU* approach, our methodology achieved its best results. The measures *Braun-Blanket* and *Laplace* achieve good results in **Weighted Average Accuracy**, namely 0.61.

8. Conclusion

We study the behavior of our methodology for recognizing TE by Generality. Also, we provide a thorough comparison to related works. This is done taking into account the limitations of typical language-independent and unsupervised learning techniques. In order to obtain fair comparison, we used a well known dataset studied in the RTE Challenge as our test-bed. Further, as we are interested in a special kind of TE, we built a suitable corpus.

Average PRECISION - NO ENTAILMENT by RTE Challenges — Measures versus Approach			
AAM	**Arithmetic Average by Approach**		
	With All Words	**Without Stop Words**	**With MWU**
ADDED VALUE	0.40	0.39	0.28
BRAUN-BLANKET	**0.56**	0.46	0.45
CERTAINTY FACTOR	0.43	0.44	**0.59**
CONDITIONAL PROBABILITY	0.42	0.41	**0.59**
CONVICTION	0.44	**0.49**	0.48
GINI INDEX	0.39	0.48	0.53
J-MEASURE	0.26	0.27	0.42
LAPLACE	0.40	0.41	**0.59**
AAM	**Weighted Average by Approach**		
	With All Words	**Without Stop Words**	**With MWU**
ADDED VALUE	0.49	0.48	0.32
BRAUN-BLANKET	**0.62**	**0.53**	0.51
CERTAINTY FACTOR	0.50	0.51	0.60
CONDITIONAL PROBABILITY	0.50	0.49	0.61
CONVICTION	0.46	0.49	0.52
GINI INDEX	0.46	**0.53**	0.55
J-MEASURE	0.37	0.38	0.43
LAPLACE	0.47	0.48	**0.62**

Table 3: PRECISION – NO ENTAILMENT Averages | Measures versus Approach

In this process we learned that detecting entailment between sentences is not an exact science. We saw that each new RTE Challenge required different approach to the problem. Thus, we do not provide a measure or an approach that pretends to solve the problem. We can only conclude, based on evidences from Table 2 that for some combinations of measure and preprocessing approach our method shows good precision in recognizing TE.

Comparing our results, with the results of other relevant methodologies, presented in Section 3.1., we prove that our methodology achieves higher performance figures. The measures *Braun-Blanket* and *Laplace* achieve better results for **Weighted Average Accuracy**, namely 0.61.

With this paper, we contribute an original proposal to RTE. Our methodology is unsupervised and language-independent, and accounts for the asymmetry of the studied phenomena by means of asymmetric similarity measures.

References

Barzilay, R. and McKeown, K. R. (2005). Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, 31(3):297–328.

Bayer, S., Burger, J., Ferro, L., Henderson, J., and Yeh, E. (2005). Mitre's Submission to the EU Pascal RTE Challenge. In *PASCAL. Proc. of the First Challenge Workshop. Recognizing Textual Entailment*, pages 41–44.

Cleuziou, G., Dias, G., and Levorato, V. (2010). Modélisation Prtopologique pour la Structuration Sémantico-Lexicale. In *Proceedings of the 17èmes Rencontres de la Société Francophone de Classification (SFC 2010)*.

Dagan, I. and Glickman, O. (2004). Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability. In *Learning Methods for Text Understanding and Mining*.

Dias, G., Guilloré, S., and Lopes, J. (1999). Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text Corpora. In *Proceedings of the 6ème Confrence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN 1999)*, pages 333–339.

Dias, G., Alves, E., and Lopes, J. (2007). Topic Segmentation Algorithms for Text Summarization and Passage Retrieval: An Exhaustive Evaluation. In *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI 2007)*, pages 1334–1340.

Dias, G., Mukelov, R., and Cleuziou, G. (2008). Unsupervised Graph-Based Discovery of General-Specific Noun Relationships from Web Corpora Frequency Counts. In *Proceedings of the 12th International Conference on Natural Language Learning (CoNLL 2008)*.

Glickman, O. and Dagan, I. (2005). Web Based Probabilistic Textual Entailment. In *Proceedings of the 1st Pascal Challenge Workshop*, pages 33–36.

Michelbacher, L., Evert, S., and Schtze, H. (2007). Asymmetric Association Measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, pages 1–6.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL 2002, pages 311–318.

Pazienza, M. T., Pennacchiotti, M., and Zanzotto, F. M. (2005). A Linguistic Inspection of Textual Entailment. In *Proceedings of the 9th Conference on Advances in Artificial Intelligence*, AI*IA 2005, pages 315–326.

Pecina, P. and Schlesinger, P. (2006). Combining Association Measures for Collocation Extraction. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL 2006)*, pages 651–658.

Perez, D., Alfonsecaia, E., and Rodríguez, P. (2005). Application of the Bleu Algorithm for Recognising Textual Entailments. In *Proceedings of the Recognising Textual Entailment Pascal Challenge*.

Romano, L., Kouylekov, M., and Szpektor, I. (2006). Investigating a Generic Paraphrase-Based Approach for Relation Extraction. In *EACL 2006*.

Sanderson, M. and Croft, B. (1999). Deriving Concept Hierarchies from Text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 1999, pages 206–213.

Sanderson, M. and Lawrie, D. (2000). Building, Testing, and Applying Concept Hierarchies. *Advances in Information Retrieval*, 7:235–266.

Tan, P.-N., Kumar, V., and Srivastava, J. (2004). Selecting the Right Objective Measure for Association Analysis. *Information Systems*, 29(4):293–313.

Unsupervised and Language-Independent Method to Recognize Textual Entailment by Generality

Sebastião Pais
MINES ParisTech
Centre de Recherche en Informatique
77305 Fontainebleau, France
pais@cri.ensmp.fr

Gaël Dias and **Rumen Moraliyski**
Normandie University
UNICAEN, GREYC CNRS
F-14032 Caen, France
firstname.lastname@unicaen.fr

João Cordeiro
University of Beira Interior
HULTIG
6200 Covilhã, Portugal
jpaulo@di.ubi.pt

Abstract

In this work we introduce a particular case of textual entailment (TE), namely *Textual Entailment by Generality* (TEG). In text, there are different kinds of entailment yielded from different types of implicative reasoning (lexical, syntactic, common sense based), but here we focus just on TEG, which can be defined as an entailment from a specific statement towards a relatively more general one. Therefore, we have $T \xrightarrow{G} H$ whenever the premise T entails the hypothesis H, the hypothesis being more general than the premise. We propose an unsupervised and language-independent method to recognize TEGs, given a pair $\langle T, H \rangle$ in an entailment relation. We have evaluated our proposal through two experiments: (a) Test on $T \xrightarrow{G} H$ English pairs, where we know that TEG holds; (b) Test on $T \rightarrow H$ Portuguese pairs, randomly selected with 60% of TEGs and 40% of TE without generality dependency (TEnG).

1. Introduction

TE aims to capture major semantic inference needs across applications in Natural Language Processing (NLP). Automatic identification of TEs has become a relevant issue promoted by the series of challenges on Recognizing Textual Entailment (RTE), where it is defined as a directional relationship between pairs of text expressions denoted by T (the entailing *"Text"*) and H (the entailed *"Hypothesis"*). We say that T entails H if humans reading T would typically infer that H is most likely true (Dagan et al., 2005). Basically, RTE is the task of deciding, given two text fragments, whether the meaning of one of the texts is entailed (can be inferred) from the other one. As noted by Dagan et al. (2005), this definition is based on common human understanding of language, much like the definition of any other language understanding task. Accordingly, it enables the creation of gold-standard evaluation data sets for the task, where humans can judge whether the entailment relation holds for a given $\langle T, H \rangle$ pair. This setting is analogous to the creation of gold standards for other text understanding applications like Question Answering (QA) and Information Extraction (IE), where human annotators are asked to judge whether the target answer or relation can indeed be inferred from a candidate text.

We introduce the TEG paradigm, which can be defined as the entailment from a specific sentence towards a more general one. For example, the pair $\langle S_1, S_2 \rangle$, taken from the RTE-1 corpus, naturally evidences that S_1 entails/implies S_2, and the latter is more general. Therefore, we have TEG from S_1 to S_2, denoted as: $S_1 \xrightarrow{G} S2$.

S_1: *Mexico City has a very bad pollution problem because the mountains around the city act as walls and block in dust and smog.*

S_2: *Poor air circulation out of the mountain-walled Mexico City aggravates pollution.*

To understand how TE by Generality can be modeled, we propose a new paradigm based on a new *Informative Asymmetric Measure* (IAM), called the *Asymmetric InfoSimba Similarity* (AIS) measure. Instead of relying on the exact matches of words between texts, we propose that one sentence entails the other by generality if two constraints hold: (a) if and only if both sentences share many related words and (b) if most of the words of a given sentence are more general than the words of the other one. As far as we know, we are the first to propose an unsupervised, language-independent, threshold-free methodology in the context of TEG.

In order to evaluate our methodology, it was necessary to create a corpus of pairs $T \rightarrow H$ and a set of TEG pairs ($T \overset{G}{\rightarrow} H$). This was achieved through the *CrowdFlower*[1] system, a convenient and fast way to collect annotations from a broad base of paid non-expert contributors over the Web. The corpus is composed of $T \rightarrow H$ pairs collected from the RTE challenge (RTE-1 through RTE-5). Only positive pairs of TE were submitted to CrowdFlower for annotation, together with a small set of carefully selected cases of known categorization that are used to train the participating annotators and to exercise quality control.

2. Corpus Construction

Large scale annotation projects such as TreeBank (Marcus et al., 1993), PropBank (Palmer et al., 2005), TimeBank (Pustejovsky et al., 2003), FrameNet (Baker et al., 1998), SemCor (Miller et al., 1993), and others play an important role in NLP research, encouraging the development of new ideas, tasks, and algorithms. The construction of these datasets, however, is extremely expensive in both annotator-hours and financial cost. Since the performance of many NLP tasks is limited by the amount and quality of data available to them (Banko and Brill, 2001), one promising alternative for some tasks is the collection of non-expert annotations. The availability and the increasing popularity of crowdsourcing services have been considered as an interesting opportunity to meet the aforementioned needs and design criteria.

Crowdsourcing services have been recently used with success for a variety of NLP applications (Callison-Burch and Dredze, 2010). Although MTurk is directly accessible only to US citizens, the CrowdFlower service provides a crowdsourcing interface to MTurk for non-US citizens.

The main idea in using crowdsourcing to create NLP resources is that the acquisition and annotation of large datasets needed to train and evaluate NLP tools and applications can be carried out in a cost-effective manner by defining simple Human Intelligence Tasks (HITs) routed to a crowd of non-expert workers, called *Turkers*, who are hired through online marketplaces.

2.1. Building Methodology - Quantitative Analysis

Our approach builds on a pipeline of HITs routed to MTurk workforce through the CrowdFlower interface. The objective is to collect $\langle T, H \rangle$ pairs where entailment by generality holds.

Our building methodology has several stages. First we select the positive pairs of TE from the first five RTE challenges. These pairs are then submitted to CrowdFlower through a job that we have built online, to be evaluated by Turkers. In CrowdFlower each $\langle T, H \rangle$ pair is a unit. The Turkers are asked to choose one of the following *Entailment by Generality* (TEG), *Entailment, but not Generality* (TEnG) or *Other*, whichever is most appropriate for the $\langle T, H \rangle$ pair under consideration.

Table 1 summarizes the work involved in the annotation of the entailment cases of the RTE-1 through RTE-5 datasets with the TEG, TEnG and *Other* labels. A total of 2,000 $\langle T, H \rangle$ pairs known to be in an entailment relation were uploaded, from which 1,740 were submitted for evaluation, and the remaining 260 constitute our *Gold* units.

[1] `http://crowdflower.com/` [Last access: 14th December, 2013]

	RTE-1	RTE-2	RTE-3	RTE-4	RTE-5
# Input Pairs[2]	400	400	400	500	300
# Pairs to Launch[3]			1,740		
# Gold Pairs[4]			260		
# Output Pairs[5]			1,203		
# Discarded Pairs[6]			797		
# *Trusted Turkers*			2,308		
# Trusted Judgments			5,220 (1,740*3)		
# Untrusted Judgments			60,482		
Evaluation Time			≈43 days		
Cost ($)			108.08		

Table 1: Summary of RTE by Generality corpus annotation task

In Table 1 we can see that 1,203 $\langle T, H \rangle$ pairs were annotated as TEG. Each pair was evaluated by three Turkers, and the final average inter-annotator agreement of 0.8 was verified.

This task proved to be hard for the Turkers, as it is difficult for human annotators to identify the entailment relation and entailment by generality in particular. This is proved by the time spent to complete the task (*Evaluation Time*) and the total number of *Judgments* (*Trusted* + *Untrusted*) needed to achieve the final objective.

The resulting manually annotated corpus is the first large-scale dataset containing a reasonable number of TEG pairs and constitutes one of the contributions of our work. It is an important resource available to the research community.

3. Asymmetric Association Measures

Most of the existing measures that evaluate the degree of similarity between words are symmetric (Pecina and Schlesinger, 2006; Tan et al., 2004). In order to avoid as much as possible the necessity of training data, different works propose the use of asymmetric association measures. Some have been introduced in the domain of taxonomy construction (Sanderson and Croft, 1999), others in cognitive psycholinguistics (Michelbacher et al., 2007) and in word order discovery (Dias et al., 2008).

Sanderson and Croft (1999) is one of the first studies to propose the use of *conditional probability* for taxonomy construction. They assume that a term t_2 subsumes a term t_1 if the documents in which t_1 occurs are a subset of the documents in which t_2 occurs constrained by $P(t_2|t_1) \geq 0.8$ and $P(t_1|t_2) < 1$. By gathering all subsumption relations, they build the semantic structure of any domain, which corresponds to a directed acyclic graph. In Sanderson and Lawrie (2000), the subsumption relation is relieved to the following expression $P(t_2|t_1) \geq P(t_1|t_2)$ and $P(t_2|t_1) > t$ where t is a given threshold and all term pairs found to have a subsumption relationship are passed through a transitivity module which removes extraneous subsumption relationships in such a way that transitivity is preferred over direct pathways, thus leading to a non-triangular directed acyclic graph.

In Michelbacher et al. (2007) the plain *conditional probability* and the *ranking measure* based on the Pearson's χ^2 test were used as a model for directed psychological association in the human mind. In particular, $R(t_2\|t_1)$ returns the rank of t_2 in the association list of t_1 given by the order obtained with the Pearson's χ^2 test for all the words co-occurring with t_1. So, when comparing $R(t_2\|t_1)$ and $R(t_1\|t_2)$, the smaller rank indicates the strongest association.

In the specific domain of word order discovery, Dias et al. (2008) proposed a methodology combining directed graphs with the TextRank algorithm (Mihalcea and Tarau, 2004) to automatically induce a general-specific word order for a given vocabulary based on Web corpora frequency counts.

[2]Number of pairs $T \rightarrow H$ uploaded
[3]Number of pairs $T \rightarrow H$ submitted for evaluation
[4]Number of *Gold* pairs $T \rightarrow H$
[5]Number of pairs $T \rightarrow H$ classified as *Entailment by Generality*
[6]Number of pairs classified as *Entailment, but not Generality* or *Other*

In order to compute the general-specific relations between sentence pairs we have employed eight Asymmetric Association Measures (AAM) defined in the following equations: *Added Value* (Equation 1), *Braun-Blanket* (Equation 2), *Certainty Factor* (Equation 3), *Conviction* (Equation 4), *Gini Index* (Equation 5), *J-measure* (Equation 6), *Laplace* (Equation 7), and *Conditional Probability* (Equation 8).

$$AV(x\|y) = P(x|y) - P(x) \quad (1) \qquad BB(x\|y) = \frac{f(x,y)}{f(x,y) + f(\overline{x},y)} \quad (2)$$

$$CF(x\|y) = \frac{P(x|y) - P(x)}{1 - P(x)} \quad (3) \qquad CO(x\|y) = \frac{P(x) \times P(\overline{y})}{P(x,\overline{y})} \quad (4)$$

$$GI(x\|y) = P(y) \times (P(x|y)^2 + P(\overline{x}|y)^2) - P(x)^2 \times P(\overline{y}) \times (P(x|\overline{y})^2 + P(\overline{x}|\overline{y})^2) - P(\overline{x})^2. \quad (5)$$

$$JM(x\|y) = P(x,y) \times \log \frac{P(x|y)}{P(x)} + P(\overline{x},y) \times \log \frac{P(\overline{x}|y)}{P(\overline{x})} \quad (6)$$

$$LP(x\|y) = \frac{N \times P(x,y) + 1}{N \times P(y) + 2} \quad (7) \qquad P(x|y) = \frac{P(x,y)}{P(y)} \quad (8)$$

3.1. Asymmetry between Sentences

There are a number of ways to compute the similarity between two sentences. Most similarity measures determine the distance between two vectors associated with two sentences (i.e. the vector space model). However, when applying the classical similarity measures between two sentences, only the identical indexes of the row vector X_i and X_j are taken into account, which may lead to miscalculated similarities. To deal with this problem, different methodologies have been proposed. A promising one is the InfoSimba informative similarity measure (Dias et al., 2007), expressed in Equation 9.

$$IS(X_i, X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{q} X_{ik} \times X_{jl} \times S(W_{ik}, W_{jl})}{\left(\begin{array}{l} \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik} \times X_{il} \times S(W_{ik}, W_{il}) + \\ \\ \sum_{k=1}^{q}\sum_{l=1}^{q} X_{jk} \times X_{jl} \times S(W_{jk}, W_{jl}) - \\ \\ \sum_{k=1}^{p}\sum_{l=1}^{q} X_{ik} \times X_{jl} \times S(W_{ik}, W_{jl}) \end{array} \right)}. \quad (9)$$

Here $S(.,.)$ is any symmetric similarity measure and each W_{ik} corresponds to the attribute word at the k^{th} position in the vector X_i, and p and q are the lengths of the vectors X_i and X_j respectively. This measure aims to compute the correlations between all pairs of words in two word context vectors instead of just relying on their exact match as with the cosine similarity measure. Furthermore, InfoSimba guarantees to capture similarity between pairs of sentences even when they do not share words. For example, this can happen when one sentence is a paraphrased version of the other and all the content words are substituted for similar words.

3.2. Asymmetric Similarities

Although there are many asymmetric similarity measures, they pose problems that may reduce their utility. On the one hand, asymmetric association measures can only evaluate the generality/specificity relation between words that are known to be in a semantic relation (Sanderson and Croft, 1999; Dias et al., 2008). Indeed, they generally capture the direction of association between two words based on document contexts and only take into account a loose semantic proximity between words. For example, it is highly probable to find that *Apple* is more general than *iPad*, which cannot be considered as a hypernymy/hyponymy or a meronymy/holonymy relation. On the other hand, asymmetric attributional word similarities only take into account common contexts to assess the degree of asymmetric relatedness between two words. To overcome this limitation, we introduce the *Asymmetric InfoSimba Similarity*

measure (AIS), whose underlying idea is to say that one word x is semantically related to a word y and x is more general than y, if x and y share as many contexts as possible and each context word of x is likely to be more general than most of the context words of y. The AIS is defined in Equation 10, where $AS(.\|.)$ is any asymmetric word similarity measure, likewise for IS in Equation 9 where $S(.,.)$ stands for any symmetric similarity measure.

$$AIS(X_i\|X_j) = \frac{\sum_{k=1}^{p}\sum_{l=1}^{q} X_{ik} \times X_{jl} \times AS(W_{ik}\|W_{jl})}{\left(\begin{array}{l} \sum_{k=1}^{p}\sum_{l=1}^{p} X_{ik} \times X_{il} \times AS(W_{ik}\|W_{il}) + \\ \\ \sum_{k=1}^{q}\sum_{l=1}^{q} X_{jk} \times X_{jl} \times AS(W_{jk}\|W_{jl}) - \\ \\ \sum_{k=1}^{p}\sum_{l=1}^{q} X_{ik} \times X_{jl} \times AS(W_{ik}\|W_{jl}) \end{array}\right)}. \quad (10)$$

We now apply this idea to the RTEG problem, where each sentence is characterized by its content words and a sentence T is semantically related to sentence H and H is more general than T (i.e. $T \overset{G}{\rightarrow} H$), if H and T share as many related words as possible and each context word of H is likely to be more general than most of the words of T.

As a result, we propose a new simple and effective method for entailment identification through the AIS measure. We state that an entailment ($T \overset{G}{\rightarrow} H$) will hold if and only if $AIS(T\|H) < AIS(H\|T)$. Note that, contrarily to the existing methodologies, we do not need to define or tune any threshold at all. Indeed, due to its asymmetric definition, the *Asymmetric InfoSimba* similarity measure allows us to compare both sides of a candidate entailment.

Since we only want to compare $AIS(T\|H)$ and $AIS(H\|T)$, the denominator of AIS in both cases does not change. Thus we have defined an equivalent (with respect to the task) but simplified version of AIS – the $AISs(.\|.)$ in Equation 11, which ended up to be the one used in our experimentation.

$$AISs(X_i\|X_j) = \sum_{k=1}^{p}\sum_{l=1}^{q} X_{ik} \times X_{jl} \times AS(W_{ik}\|W_{jl}). \quad (11)$$

3.3. Three Levels of Word Granularity

It is evident that even through the simplified version of our proposed measure ($AISs$) we end up with a considerable amount of computation complexity – $O(n^2)$ – for comparing two sentences. Therefore, we have also considered two additional possibilities to reduce the number of words in each sentence without losing effectiveness. These are: (1) stop-word[7] removal and (2) multiword units (MWU) replacement, by identifying MWUs in the sentences. The MWUs were automatically computed using SENTA[8] (Dias et al., 1999) from the first five RTE datasets.

In summary, our experiments are based on three approaches to the calculations – using all words, using a list of stop words and finally using MWUs.

4. Experimentation and Results

In order to assess the effectiveness and general quality of our proposed measures for TEG identification, we have performed a comparative test on the corpus described in Section 2. We have tested our proposed $AISs$ measure with all word-similarity functions, mentioned in Section 3. Sentence similarity is computed in three different manners, as described previously in Section 3.3.

The evaluation functions used are based on the confusion matrix, in particular the accuracy and the precision. More specifically, we dealt with *Average Accuracy*, *Average Precision*, *Weighted Average Accuracy*, and *Weighted Average Precision*.

[7] A list of English stop-words, obtained using http://www.microsoft.com/en-us/download/confirmation.aspx?id=10024 [Last access: 14th December, 2013]

[8] The *Software for the Extraction of N-ary Textual Associations*.

4.1. The TEG Corpus

Here we report the obtained results of our methodology on the TEG corpus. These are the results we are most interested in as they concern the problem on which we are focusing our attention, namely identification of entailment by generality.

With respect to **Accuracy**, as seen in Table 2, the best performance, 0.85, is achieved by the measure *Braun-Blanket* in conjunction with the *MWU* method. The second best measure was *Added Value* with an accuracy of 0.69. It is important to highlight the significant difference between these two AAMs.

The measure *Braun-Blanket* remains the best one in the stop-word removal approach with an accuracy of 0.73, and *Gini Index* and *J-measure* achieved the second best results with an accuracy of 0.64. In *All Words* we have two measures with the best performance – *Conviction* and *J-measure* achieving respectively 0.70 and 0.69 of accuracy.

From Table 2 we may conclude that although *Conviction* is the best measure with *All Words* with respect to **Accuracy**, its performance is virtually equivalent to that of a random guesser for the *Without Stop Words* and *With MWU* approaches.

AAM	**Accuracy**		
	All Words	*Without Stop Words*	*With MWU*
AV	0.67	0.63	0.69
BB	0.62	0.73	0.85
CF	0.65	0.63	0.64
P	0.61	0.60	0.64
CO	0.7	0.59	0.54
GI	0.65	0.64	0.68
JM	0.69	0.64	0.6
LP	0.64	0.62	0.6

Table 2: Accuracy by AAM

AAM	**Precision for A**		
	All Words	*Without Stop Words*	*With MWU*
AV	0.81	0.74	0.82
BB	0.69	0.80	0.93
CF	0.74	0.67	0.63
P	0.72	0.70	0.64
CO	0.74	0.63	0.56
GI	0.74	0.72	0.65
JM	0.83	0.78	0.64
LP	0.71	0.69	0.58
AAM	**Precision for B**		
	All Words	*Without Stop Words*	*With MWU*
AV	0.45	0.47	0.49
BB	0.51	0.62	0.73
CF	0.51	0.56	0.68
P	0.45	0.44	0.63
CO	0.64	0.52	0.5
GI	0.51	0.51	0.71
JM	0.5	0.43	0.55
LP	0.52	0.51	0.63

Table 3: Precision by AAM

In terms of **Precision**, the *Braun-Blanket* measure, in conjunction with the *MWU* approach, achieved

the best results for both entailment types: *Entailment by Generality* (**A**) and *Entailment, but no Generality* (**B**), with 0.93 and 0.73 points respectively. On **Precision A** the worst result is achieved by *Conviction* – 0.56 with *MWU*, and for **Precision A**, the worst result is achieved by *J-measure* with stop words removed: 0.43.

4.2. A Portuguese TEG Corpus

In this section we present the results of an experiment parallel to the one discussed in Section 4.1. The main idea was to measure the degree to which our methodology was capable to recognize TEGs in different languages. To this end, we have randomly selected a subset of 100 $\langle T, H \rangle$ pairs from the TEG Corpus, preserving the proportion of 60 $\langle T, H \rangle$ TEG pairs (Entailment by Generality) and 40 TEnG $\langle T, H \rangle$ pairs (Entailment, but no Generality). This subset of 100 TE pairs was translated into Portuguese using the *Google Translate* service.

Machine translation is a viable alternative to manual translation due to a combination of two factors. First, since our intention was to be as much language independent as possible, our methodology does not use morpho-syntactic analysis and language specific word order knowledge. On the other hand, *Google Translate* is reasonably successful in correct content word substitution. Thus, from the perspective of our bag-of-words approach *Google Translate* preserves well the important information. This supposition is in line with the fact that our results in Portuguese are comparable to the corresponding results in English.

With respect to **Accuracy** the best performance is achieved with the *Braun-Blanket* measure in conjunction with the *With MWU* approach, with a result of 0.76, as shown in Table 4. In this approach the second best measure is *Added Value*, with a result of 0.69. Similarly, *Braun-Blanket* achieves the best performance in the *Without Stop Words* approach, with a result of 0.71, followed by *Gini Index*, with 0.66. In *All Words*, the measure with the best **Accuracy** is *J-measure* (0.72).

In Table 4, the three measures with the lowest **Accuracy** are *Conditional Probability* in the approaches *All Words* and *Without Stop Words*, and *Conviction* – in *With MWU*.

AAM	**Accuracy**		
	All Words	*Without Stop Words*	*With MWU*
AV	0.63	0.62	0.69
BB	0.62	0.71	0.76
CF	0.64	0.62	0.63
P	0.59	0.57	0.6
CO	0.68	0.6	0.5
GI	0.66	0.66	0.68
JM	0.72	0.58	0.6
LP	0.61	0.62	0.63

Table 4: Accuracy by AAM

Considering the **Accuracy** figures for English and for Portuguese, presented in Table 2 and Table 4, which show similar scale and variations, we conclude that the performance of our methodology is not significantly influenced by the language.

With respect to **Precision – Entailment by Generality** the measure *Braun-Blanket* in conjunction with the approach *With MWU* presents the best results (0.88), followed by *J-measure* in conjunction with the approach *All Words* (0.85). The worst results are achieved by *Certainty Factor* and *Laplace* in *With MWU* (0.6).

With respect to **Precision – Entailment, but no Generality** the results are markedly lower. The best results are achieved in *With MWU* by *Certainty Factor*, *Gini Index* and *Laplace* (0.68). The worst results are achieved by *Added Value* in *All Words* (0.38).

Both the **Accuracy** and the **Precision** figures show that whether applied to a corpus in English or in Portuguese, our methodology provides a classification capability that is significantly better than a random guessing baseline and virtually indistinguishable with respect to the language.

AAM	Precision for A		
	All Words	*Without Stop Words*	*With MWU*
AV	0.78	0.78	0.85
BB	0.65	0.78	0.88
CF	0.68	0.65	0.6
P	0.68	0.65	0.62
CO	0.73	0.65	0.55
GI	0.72	0.75	0.68
JM	0.85	0.72	0.62
LP	0.7	0.72	0.6
AAM	**Precision for B**		
	All Words	*Without Stop Words*	*With MWU*
AV	0.40	0.38	0.45
BB	0.58	0.6	0.58
CF	0.58	0.58	0.68
P	0.45	0.45	0.58
CO	0.60	0.52	0.43
GI	0.58	0.53	0.68
JM	0.53	0.38	0.58
LP	0.48	0.48	0.68

Table 5: Precision by AAM

5. Conclusion

This work presents a new methodology for recognizing TEG and studies its behavior in a detailed experimental configuration, achieving significant results. As seen in Table 2 and Table 3, there is always a measure and an approach that stand out, namely the *Braun-Blanket* measure in *With MWU*. However, *J-measure* and *Conviction* also have good results – (a) in **Precision – Entailment by Generality** *J-measure* with *All Words* has the second best performance (0.83) – in other words, *J-measure* with *All Words* has a good performance to identify entailment by generality between sentences; (b) *Conviction* ranks second for **Accuracy** (0.7) and achieves good results in **Precison – Entailment, but no generality or Other**, both in the *All Words* approach.

We may conclude that our methodology is language independent since results for Portuguese are comparable to those for English although with less significant discrimination between the first and the second measure. However, in terms of **Accuracy** (Table 4) and **Precision – Entailment by Generality** (Table 5) *Braun-Blanket* achieves the best performance in the approach *With MWU*.

With this paper we also contribute to the consideration of a new kind of textual entailment, providing also new experimental resources (TEG Corpus). Our methodology is unsupervised and language independent, and accounts for the asymmetry of the studied phenomena by means of asymmetric similarity measures. Using our methodology we have demonstrated excellent results in identifying textual entailment by generality.

References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL 1998)*, volume 1, pages 86–90.

Banko, M. and Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL 2001)*, pages 26–33.

Callison-Burch, C. and Dredze, M. (2010). Creating Speech and Language Data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT 2010)*, pages 1–12.

Dagan, I., Glickman, O., and Magnini, B. (2005). The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005*, pages 177–190.

Dias, G., Guilloré, S., and Lopes, J. (1999). Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text Corpora. In *Proceedings of the 6ème Confrence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN 1999)*, pages 333–339.

Dias, G., Alves, E., and Lopes, J. (2007). Topic Segmentation Algorithms for Text Summarization and Passage Retrieval: An Exhaustive Evaluation. In *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI 2007)*, pages 1334–1340.

Dias, G., Mukelov, R., and Cleuziou, G. (2008). Unsupervised Graph-Based Discovery of General-Specific Noun Relationships from Web Corpora Frequency Counts. In *Proceedings of the 12th International Conference on Natural Language Learning (CoNLL 2008)*.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a Large Annotated Corpus of English: the penn treebank. *Computational Linguistics*, 19(2):313–330.

Michelbacher, L., Evert, S., and Schtze, H. (2007). Asymmetric Association Measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2007)*, pages 1–6.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing Order into Texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A Semantic Concordance. In *Proceedings of the Workshop on Human Language Technology*, HLT 1993, pages 303–308.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Pecina, P. and Schlesinger, P. (2006). Combining Association Measures for Collocation Extraction. In *Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL 2006)*, pages 651–658.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., and Lazo, M. (2003). The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics 2003*, pages 647–656.

Sanderson, M. and Croft, B. (1999). Deriving Concept Hierarchies from Text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999)*, pages 206–213.

Sanderson, M. and Lawrie, D. (2000). Building, Testing, and Applying Concept Hierarchies. *Advances in Information Retrieval*, 7:235–266.

Tan, P.-N., Kumar, V., and Srivastava, J. (2004). Selecting the Right Objective Measure for Association Analysis. *Information Systems*, 29(4):293–313.

Cover 3

http://dcl.bas.bg/clib

ISSN 2367-5578