



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

Proceedings of the Third International Conference

**Computational
Linguistics in
Bulgaria**

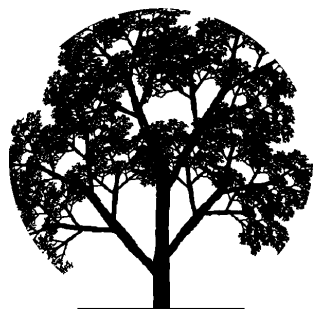


**27-29 May 2018
Sofia, Bulgaria**

The Third International Conference *Computational Linguistics in Bulgaria (CLIB 2018)* is organised with the support of the National Science Fund of the Republic of Bulgaria under the project *Towards a Semantic Network Enriched with a Variety of Relations*, Grant Agreement DN10/3/2016.




CLIB 2018 is organised by:



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

PUBLICATION AND CATALOGUING INFORMATION

Title:	Proceedings of the Third International Conference <i>Computational Linguistics in Bulgaria (CLIB 2018)</i>
ISSN:	2367 5675 (online)
Published and distributed by:	The Institute for Bulgarian Language Bulgarian Academy of Sciences
Editorial address:	Institute for Bulgarian Language Bulgarian Academy of Sciences 52 Shipchenski Prohod Blvd., Bldg. 17 Sofia 1113, Bulgaria +3592/ 872 23 02
Copyright:	Copyright of each paper stays with the respective authors. The works in the Proceedings are licensed under a Creative Commons Attribution 4.0 International Licence (CC BY 4.0).  License details: http://creativecommons.org/licenses/by/4.0 Copyright © 2018

Proceedings of the
Third International Conference
*Computational Linguistics in
Bulgaria*



Sofia, Bulgaria
28-29 May 2018

PROGRAMME COMMITTEE

Chair:

Svetla Koeva – Institute for Bulgarian Language (BAS)

Co-chairs:

Mila Dimitrova-Vulchanova – Norwegian University of Science and Technology

Cvetana Krstev – University of Belgrade

Tania Avgustinova – Saarland University

Iana Atanassova – Centre Tesnière, Université de Franche-Comté, UFR SLHS

Verginica Barbu Mititelu – Research Institute for Artificial Intelligence, Romanian Academy

Mariana Damova – Mozaika, Bulgaria

Ivan Derzhanski – Institute of Mathematics and Informatics (BAS)

Radovan Garabík – L’udovít Štúr Institute of Linguistics, Slovak Academy of Sciences

Stefan Gerdjikov – Sofia University

Filip Ginter – University of Turku

Kjetil Rå Hauge – University of Oslo

Zornitsa Kozareva – Google

Ivan Koychev – Sofia University

Hristo Krushkov – Plovdiv University

Eric Laporte – Université Paris-Est Marne-la-Vallée

Denis Maurel – University of Tours

Stoyan Mihov – Institute of Information and Communication Technologies

Ruslan Mitkov – University of Wolverhampton

Preslav Nakov – Qatar Computing Research Institute

Karel Oliva – Institute of the Czech Language, Academy of Sciences of the Czech Republic

Maciej Ogrodniczuk – Institute of Computer Science, Polish Academy of Sciences

Maciej Piasecki – Wrocław University of Technology

Dragomir Radev – Yale University

Marko Tadić – University of Zagreb

Hristo Tanev – Joint Research Centre of the European Commission

Tinko Tinchev – Sofia University

Dan Tufis – Research Institute for Artificial Intelligence, Romanian Academy

Duško Vitas – University of Belgrade

Radka Vlahova – Sofia University

Victoria Yaneva – University of Wolverhampton

ORGANISING COMMITTEE

Svetlozara Leseva – Institute for Bulgarian Language (BAS)

Iana Atanassova – Centre Tesnière, Université de Franche-Comté, UFR SLHS

Rositsa Dekova – Plovdiv University, Faculty of Philology, Department of English Studies

Tsvetana Dimitrova – Institute for Bulgarian Language (BAS)

Dimitar Hristov – Institute for Bulgarian Language (BAS)

Alexander Popov – Institute of Information and Communication Technologies

Borislav Rizov – Institute for Bulgarian Language (BAS)

Katya Saint-Amand – Senior Linguist, Chenope

Valentina Stefanova – Institute for Bulgarian Language (BAS)

Ivelina Stoyanova – Institute for Bulgarian Language (BAS)

Ekaterina Tarpomanova – Sofia University, Faculty of Slavic Studies

Maria Todorova – Institute for Bulgarian Language (BAS)

Martin Yalamov – Institute for Bulgarian Language (BAS)

Victoria Yaneva – University of Wolverhampton

PLENARY TALKS

LINGUISTIC INTELLIGENCE: COMPUTERS VS. HUMANS

Prof. Ruslan Mitkov (University of Wolverhampton)

Computers are ubiquitous – they are and are used everywhere. But how good are computers at understanding and producing natural languages (e.g. English or Bulgarian)? In other words, what is the level of their linguistic intelligence? This presentation will examine the linguistic intelligence of the computers and will look at the challenges ahead...

I shall begin by a brief historical flashback. I shall plot the timeline of the linguistic intelligence of computers against that of humans. Natural Language Processing (NLP) advances in the last 20 years have made it possible for the linguistic intelligence of computers to increase significantly but they are still behind humans...

The presentation will explain why it is so difficult for computers to understand, generate and in general, to process natural language texts – it is a steep road/learning curve, it is long and winding road for both computers and researchers who seek to develop intelligent programs. The talk will also briefly present well-established NLP techniques computers follow when 'learning' to 'speak' our languages, including rule-based and knowledge-based methods initially and machine learning and deep learning methods more recently, the latter being regarded as highly promising. A selection of Natural Language Processing applications will be outlined next. Finally, a preview will be offered of selected slides from my plenary talk at CLIB'2018 (which will be given on the following day).

I am not a clairvoyant, but at some of my plenary talks I have been asked to predict how far will computers go... At the end of my presentation in Sofia I shall share with you what I predict for the future and in general, what my vision is.

WITH A LITTLE HELP FROM NLP: MY LANGUAGE TECHNOLOGY APPLICATIONS WITH IMPACT ON SOCIETY

Prof. Ruslan Mitkov (University of Wolverhampton)

The talk will present three original methodologies developed by the speaker, underpinning implemented Language Technology tools which are already having an impact on the following areas of society: e-learning, translation and interpreting and care for people with language disabilities.

The first part of the presentation will introduce an original methodology and tool for generating multiple-choice tests from electronic textbooks. The application draws on a variety of Natural Language Processing (NLP) techniques which include term extraction, semantic computing and sentence transformation. The presentation will include an evaluation of the tool which demonstrates that generation of multiple-choice tests items with the help of this tool is almost four times faster than manual construction and the quality of the test items is not compromised. This application benefits e-learning users (both teachers and students) and is an example of how NLP can have a positive societal impact, in which the speaker passionately believes.

The talk will go on to outline two other original recent projects which are also related to the application of NLP beyond academia. First, a project, whose objective is to develop next-generation translation memory tools for translators and, in the near future, for interpreters, will be briefly presented. Finally, an original methodology and system will be outlined which helps users with autism to read and better understand texts.

BUILDING CONVERSATIONAL ASSISTANTS USING DEEP LEARNING

Dr Zornitsa Kozareva (Google)

Over the years there has been a paradigm shift in how humans interact with machines. Today's users are no longer satisfied with seeing a list of relevant web pages, instead they want to complete tasks and take actions. This raises the questions: "How do we teach machines to become useful in a human-centered environment?" and "How do we build machines that help us organize our daily schedules, arrange our travel and be aware of our preferences and habits?". In this talk, I will describe these challenges in the context of conversational assistants. Then, I will delve into deep learning algorithms for entity extraction, user intent prediction and question answering. Finally, I will highlight findings on user intent prediction from shopping, movies, restaurant and sport domains.

NEURAL GRAPH LEARNING

Dr Sujith Ravi (Google)

Recent machine learning advances have enabled us to build intelligent systems that understand semantics from speech, natural language text and images. While great progress has been made in many AI fields, building scalable intelligent systems from "scratch" still remains a daunting challenge for many applications. To overcome this, we exploit the power of graph algorithms since they offer a simple elegant way to express different types of relationships observed in data and can concisely encode structure underlying a problem. In this talk I will focus on "How can we combine the flexibility of graphs with the power of machine learning?"

I will describe how we address these challenges and design efficient algorithms by employing graph-based machine learning as a computing mechanism to solve real-world prediction tasks. Our graph-based machine learning framework can operate at large scale and easily handle massive graphs (containing billions of vertices and trillions of edges) and make predictions over billions of output labels while achieving $O(1)$ space complexity per vertex. In particular, we combine graph learning with deep neural networks to power a number of machine intelligence applications, including Smart Reply, image recognition and video summarization to tackle complex language understanding and computer vision problems. I will also introduce some of our latest research and share results on "neural graph learning", a new joint optimization framework for combining graph learning with deep neural network models.

Table of Contents

Ruslan Mitkov <i>With a little help from NLP: My Language Technology applications with impact on society</i>	1
Vito Pirrelli <i>NLP-based Assessment of Reading Efficiency in Early Grade Children</i>	5
Mila Vulchanova and Valentin Vulchanov <i>Figurative language processing: A developmental and NLP Perspective</i>	7
Nikola Taushanov, Ivan Koychev and Preslav Nakov <i>Abstractive Text Summarization with Application to Bulgarian News Articles</i>	15
Maria Gritz <i>Towards Lexical Meaning Formal Representation by virtue of the NL-DL Definition Transformation Method</i>	23
Junya Morita <i>Narrow Productivity, Competition, and Blocking in Word Formation</i>	34
Cvetana Krstev, Ranka Stanković and Duško Vitas <i>Knowledge and Rule-Based Diacritic Restoration in Serbian</i>	41
Anton Zinoviev <i>Perfect Bulgarian Hyphenation, or How not to Stutter at End-of-line</i>	52
Anna Roitberg and Denis Khachko <i>Russian Bridging Anaphora Corpus</i>	62
Ekaterina Tarpomanova <i>Aspectual and Temporal Characteristics of the Past Active Participles in Bulgarian – a Corpus-based Study</i>	69
Olena Siruk and Ivan Derzhanski <i>Unmatched Femininitives in a Corpus of Bulgarian and Ukrainian Parallel Texts</i>	77
Viktoriya Petrova <i>The Bulgarian Summaries Corpus</i>	85
Natalia Loukachevitch and Boris Dobrov <i>Ontologies for Natural Language Processing: Case of Russian</i>	93
Ranka Stanković, Miljana Mladenović, Ivan Obradović, Marko Vitas and Cvetana Krstev <i>Resource-based WordNet Augmentation and Enrichment</i>	104
Svetlozara Leseva, Ivelina Stoyanova and Maria Todorova <i>Classifying Verbs in WordNet by Harnessing Semantic Resources</i>	115

Maria Mitrofan, Verginica Barbu Mititelu and Grigorina Mitrofan <i>A Pilot Study for Enriching the Romanian WordNet with Medical Terms</i>	126
Ivelina Stoyanova <i>Factors and Features Determining the Inheritance of Semantic Primes between Verbs and Nouns within WordNet</i>	135
Borislav Rizov and Tsvetana Dimitrova <i>Online Editor for WordNets</i>	146
Amir Bakarov <i>The Effect of Unobserved Word-Context Co-occurrences on a Vector-Mixture Approach for Compositional Distributional Semantics</i>	153
Rositsa Dekova and Adelina Radeva <i>Introducing Computational Linguistics and NLP to High School Students</i>	162
Ivan Derzhanski and Milena Veneva <i>Linguistic Problems on Number Names</i>	169
Marina Dzhonova, Kjetil Rå Hauge and Yovka Tisheva <i>Parallel Web Display of Transcribed Spoken Bulgarian with its Normalised Version and an Indexed List of Lemmas</i>	177
Georgi Dzhumayov <i>Integrating Crowdsourcing in Language Learning</i>	185
Todor Lazarov <i>Bulgarian–English Parallel Corpus for the Purposes of Creating Statistical Translation Model of the Verb Forms. General Conception, Structure, Resources and Annotation</i>	193
Branislava Šandrih <i>Fingerprints in SMS messages: Automatic Recognition of a Short Message Sender Using Gradient Boosting</i>	203

With a little help from NLP: My Language Technology applications with impact on society

Ruslan Mitkov

University of Wolverhampton

Abstract

The keynote speech presents the speaker's vision that research should lead to the development of applications which benefit society. To support this, the speaker will present three original methodologies proposed by him which underpin applications jointly implemented with colleagues from across his research group. These Language Technology tools already have a substantial societal impact in the following areas: learning and assessment, translation and care for people with language disabilities.

1. Impact on learning and assessment

The first part of the presentation will introduce an original methodology and tool for generating multiple-choice tests from electronic documents. Multiple-choice tests are sets of test items, the latter consisting of a question or *stem* (e.g. Who was FIFA player of the year for 2017?), the correct *answer* (e.g. Ronaldo) and *distractors* (e.g. Messi, Neymar, Buffon). This type of test has proved to be an efficient tool for measuring students' achievement and is used on a daily basis both for assessment and diagnostics worldwide. According to Question Mark Computing Ltd (p.c.), who have licensed their Perception software to approximately three million users so far, 95% of their users employ this software to administer multiple-choice tests. Despite their popularity, the manual construction of such tests remains a time-consuming and labour-intensive task. One of the main challenges in constructing a multiple-choice test item is the selection of plausible alternatives to the correct answer which will better distinguish confident students from unconfident ones.

As an illustration, consider the sentence "Syntax is the branch of linguistics which studies the way words are put together into sentences". This sentence can be transformed into the questions "Which branch of linguistics studies the way words are put together into sentences?", "Which discipline studies the way words are put together into sentences?" or "What studies the way words are put together into sentences?". All these phrases can act as stems in multiple-choice test items. If we assume that the stem of a test item is one of the questions above, the distractors to the correct answer *syntax* should preferably be concepts semantically close to it. This is vital because in this case the distractors will be more plausible and therefore better at distinguishing good, confident students from poor and uncertain ones. For this particular test item, *semantics* or *pragmatics* would be a much better distractors than *chemistry* or *football*, for instance.

Mitkov and Ha (2003) and Mitkov et al. (2006) offered an alternative to the lengthy and demanding activity of developing multiple-choice test items by proposing an NLP-based methodology for construction of test items from instructive texts such as textbook chapters and encyclopaedical entries. This methodology makes use of NLP techniques including shallow parsing, term extraction, sentence transformation and semantic distance computing and employs resources such as corpora and ontologies like WordNet. More specifically, the system identifies important terms in a textbook text, transforms declarative sentences into questions and mines for terms which are semantically close to the correct answer, to serve as distractors.

The system for generation of multiple-choice tests described in Mitkov and Ha (2003) and in Mitkov et al. (2006) was evaluated in practical environment where the user was offered the option to post-edit and in general to accept or reject the test items generated by the system. The formal

evaluation showed that even though a significant part of the generated test items had to be discarded, and that the majority of the items classed as ‘usable’ had to be revised and improved by humans, the quality of the items generated and proposed by the system was not inferior to the tests authored by humans, were more diverse in terms of topics and very importantly – their production needed 4 times less time than the manually written items. The evaluation was conducted both in terms of measuring the time needed to develop test items and in terms of classical test analysis to assess the quality of test items.

A later study (Mitkov et al. 2009) sought to establish which similarity measures generate better quality distractors of multiple-choice tests. Similarity measures employed in the procedure of selection of distractors were collocation patterns, four different methods of WordNet-based semantic similarity (extended gloss overlap measure, Leacock and Chodorow’s, Jiang and Contrath’s as well as Lin’s measures), distributional similarity, phonetic similarity as well as a mixed strategy combining the aforementioned measures. The evaluation results showed that the methods based on Lin’s measure and on the mixed strategy outperform the rest, albeit not in a statistically significant fashion.

The system for generation of multiple-choice tests has been taken up by the National Board of Medical Examiners (NBME) based in Philadelphia, USA. NBME are the only organisation in USA who are licenced to administer and asses exams for the to-be-doctors. NBME have been using our system for delivery of low-stake tests for more than 10 years already.

2. Impact on translation

The quest for reliable tools assisting professional translators goes back to 1971 when Krollman (1971) put forward the reuse of existing human translations. A few years later, Arthern (1979) went further and proposed the retrieval and reuse not only of identical text fragments (exact matches) but also of similar source sentences and their translations (fuzzy matches). It took another decade before the ideas sketched by Krollman and Arthern were commercialised as a result of the development of various computer-aided translation (CAT) tools such as Translation Memory (TM) systems in the early 1990s. These translation tools revolutionised the work of translators and the last two decades saw dramatic changes in the translation workflow.

The TM memory systems indeed revolutionised the work of translators and nowadays the translators not benefiting from these tools are a tiny minority. However, while these tools have proven to be very efficient for repetitive and voluminous texts, are they intelligent enough? Unfortunately, they operate on fuzzy (surface) matching (Levenstein distance) mostly, cannot reuse already translated sentences which are part of another complex sentence nor texts which are synonymous to (or paraphrased versions of) the text to be translated and can be ‘fooled’ on numerous occasions. A recent study (Mitkov et al., forthcoming) shows that TM systems (Trados, MemoQ, Wordfast, Omega T) spectacularly fail to offer matches for sentences already translated but which have undergone (slight) transformations which include among others: change active to passive voice and vice versa, change word order and replace one word with synonym.

A way forward would be to equip the TM tools with Natural Language Processing (NLP) capabilities. This idea was suggested first by Mitkov (2005) at panel discussion held during 27th annual conference *Translating and the Computer* in London and the first experiments were reported by Mitkov and Pekar (2007). In the second part of his presentation, the speaker will explain how two NLP methods/tasks, namely clause splitting and paraphrasing, make it possible for TM systems to identify semantically equivalent sentences which are not necessarily identical or close syntactically and enhance performance. The results reported in Timonera and Mitkov (2015) show that TM systems which are enhanced with a clause splitting component perform with a dramatic increase of recall which is statistically significant. Adding a paraphrasing module increases further the performance and experiments with the S to XL package sizes of the Paraphrase Database PPDB show that the larger the database, the better the results.

In (Gupta et al. 2016) we presented a novel and efficient approach which incorporates semantic information in the form of paraphrasing in the edit-distance metric. The approach computes edit-distance while efficiently considering paraphrases using dynamic programming and greedy approximation. In addition to using automatic evaluation metrics such as BLEU and METEOR, we have carried out an extensive human evaluation in which we measured post-editing time, keystrokes, HTER, HMETEOR, and carried out three rounds of subjective evaluations. Our results show that

paraphrasing substantially improves TM matching and retrieval, resulting in translation performance increases when translators use paraphrase-enhanced TMs. Finally, the speaker will present a new metric developed by members of his group (Gupta et al. 2014) which is capable of comparing semantic similarity of sentences and thus becomes highly eligible for inclusion in a new generation TM matching algorithm.

The speaker will promise to go beyond the translation world: he is already thinking not only about the next-generation translation memory tools for translators but also about the future interpreting memory tools for interpreters. The presentation will sketch how this is envisaged to be developed and how it will work. In addition to the interpreting memory, the speaker will outline other tools which will be developed as support to interpreters.

3. Impact on people with language disabilities

The last part of the keynote speech will focus on the work within the recent EC-funded project FIRST whose objective was to develop a tool customised for the needs of people with autism (ASD) by allowing easy comprehension of texts which otherwise would have been challenge for them (Mitkov 2011; Orasan et al. 2012; Orasan et al. 2017). Autistic Spectrum Disorder (ASD) is a neurodevelopmental disorder which has a life-long impact on the lives of people diagnosed with the condition. In many cases, people with ASD are unable to derive the gist or meaning of written documents due to their inability to process complex sentences, understand non-literal text, and understand uncommon and technical terms. The idea put forward by the speaker was to develop a tool which would enable readers or carers to convert documents into easier-to-understand ones by (i) reducing complexity at morphological and syntactical level, (ii) by removing ambiguity in terms of lexical polysemy, anaphoric interpretation and figurative language and (iii) by improving readability through adding pictures, document navigation tools, providing concise summaries of long documents and replacing technical words with more common ones. The project FIRST produced a powerful editor called OpenBook which is operational for English, Spanish and Bulgarian and which enables carers of people with ASD to prepare texts suitable for this population. Assessment of the texts generated using the editor showed that they are not less readable than those generated more slowly as a result of onerous unaided conversion and were significantly more readable than the originals.

The speaker intends to go beyond the topic of autism and plans to develop tools customised for people with dementia. The first goal of this project will consist of the data collection of speech samples to build a corpus of transcribed speech of Alzheimer's disease and control subjects. Language technology and machine learning techniques will be employed to measure a set of speech and language markers and to assess their change. The project is expected to contribute to the understanding of changes in language use of people with dementia. It will also enhance understanding of communication in this population and will suggest improved therapeutic strategies involving the use of language and information technology to automatically correct some of the communication deficiencies identified in people with dementia.

References

- Arthern, P. J. 1979. "Machine Translation and computerized terminology systems: A translator's viewpoint". In *Translating and the computer, proceedings of a seminar. London, 14th November, 1978*, ed. Snell, B. M., 77-108. Amsterdam: North-Holland.
- Gupta, R., Bechara, H., El Maarouf, I. and Orasan, C. 2014. "UoW: NLP techniques developed at the University of Wolverhampton for Semantic Similarity and Textual Entailment". In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 785-789. Dublin, Ireland.
- Gupta, R., Orasan, C., Zampieri, M., Vela, M., Mihaela Vela, van Genabith, J. and Mitkov, R. 2016. "Improving Translation Memory matching and retrieval using paraphrases", *Machine Translation*, 30(1), 19-4
- Gupta, R., Orasan, C., Liu, Q. and Mitkov, R. 2016. A Dynamic Programming Approach to Improving Translation Memory Matching and Retrieval using Paraphrases. In *Proceedings of the 19th International Conference on Text, Speech and Dialogue (TSD)*, Brno, Czech Republic.

- Krollmann, F. 1971. "Linguistic data banks and the technical translator". *Meta* 16(1-2), 117-124.
- Mitkov, R. 2011. A Flexible Interactive Reading Support Tool (FIRST). *Presentation at the FIRST Kick-Off Meeting*. Brussels, 17th October 2011.
- Mitkov, R. 2005. "New Generation Translation Memory systems". *Panel discussion at the 27th annual Aslib conference 'Translating and the Computer'*. London.
- Mitkov, R. and Ha, L.A. 2003. "Computer-aided generation of multiple-choice tests". *Proceedings of the HLT/NAACL 2003 Workshop on Building educational applications using Natural Language Processing*, 17-22. Edmonton, Canada.
- Mitkov, R., An, L.A. and Karamanis, N. 2006. "A computer-aided environment for generating multiple-choice test items". *Journal of Natural Language Engineering*, 12 (2), 177-194.
- Mitkov, R., Ha, L.A., Varga, A. and L. Rello. 2009. "Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation". *Proceedings of the EACL'2009 workshop on geometric models of natural language semantics*, 49-56. Athens, Greece
- Mitkov, R., Feherova A., Cotella, M., and Silvestre Baquero, A. (forthcoming) "Translation Memory systems have a long way to go"
- Orasan, C., Evans, R., and Dornescu, I. (2013) Text Simplification for People with Autistic Spectrum Disorders. In Tufis, D., Rus, V. and Forascu, C. (eds.), *Towards Multilingual Europe 2020: A Romanian Perspective*, Romanian Academy Publishing House, Bucharest, 2013, ISBN: 978-973-27-2282-4, pp. 287-312.
- Orasan, C., Evans, R. and Ruslan Mitkov, R. 2018. "Intelligent Text Processing to Help Readers with Autism". *Intelligent Natural Language Processing: Trends and Applications*. Springer.
- Pekar V. and Mitkov R. 2007. "New Generation Translation Memory: Content-Sensitive Matching". *Proceedings of the 40th Anniversary Congress of the Swiss Association of Translators, Terminologists and Interpreters*. Bern.
- Timonera, K. and R. Mitkov. 2015. Improving Translation Memory Matching through Clause Splitting. *Proceedings of the RANLP'2015 workshop 'Natural Language Processing for Translation Memories'*. Hissar, Bulgaria.

NLP-based assessment of reading efficiency in early grade children

Vito Pirrelli

Institute for Computational Linguistics
National Research Council, Italy
vito.pirrelli@ilc.cnr.it

Abstract

Assessing reading skills is a laborious and time-consuming task, which requires monitoring a variety of interlocked abilities, ranging from accurate word rendering, reading fluency and lexical access, to linguistic comprehension, and interpretation, management and inference of complex events in working memory. No existing software, to our knowledge, is able to cover and integrate reading performance monitoring, instant feedback, personalised potentiation and intelligent decision support to teachers and speech therapists, assessment of response to intervention. NLP and ICT technologies can make such an ambitious platform an achievable target.

Reading is not just the ability to assign the correct pronunciation to a sequence of written symbols making up a word (or word decoding), but the joint product of decoding and deep linguistic comprehension (Gough and Tunmer, 1986; Hoover and Gough, 1990). Effective linguistic comprehension relies on language skills such as semantic and syntactic awareness. Both decoding and linguistic comprehension are necessary for reading comprehension, and neither is by itself sufficient (Hoover and Gough, 1990). However, current protocols for reading assessment measure decoding (reading accuracy and speed) and reading comprehension separately (Cornoldi and Colpo, 2012; Shinn and Shinn, 2002; Wagner et al., 2009). This does not allow evaluation of reading efficiency (Cappa et al., 2016), defined as the ability to fully understand connected texts by minimising reading time, a cognitive ability that lies at the roots of students' academic achievement (García-Madruga et al., 2014; Speece et al., 2010).

Better support to children with reading difficulties requires substantial advances in our understanding of the basic mechanisms involved in learning to read connected texts, as well as better modelling of the dynamic interaction of these mechanisms and their impact on linguistic comprehension in natural reading conditions. All these requirements call for bigger and better data to be collected in naturalistic tasks, in different environments and through multiple modalities. Here I describe an on-going, self-funded project of the CNR Institute of Computational Linguistics in Pisa, which intends to leverage the full potential of ICT and NLP technology to put in place a ubiquitous infrastructure with a simple tablet as terminal equipment. Early graders at school can read a one or two page text displayed on a tablet touchscreen, either silently or aloud. Children are asked to slide their finger across the words as they read, to guide directional tracking. After reading, the child is prompted with a few multiple-answer questions on text content. Questions are presented on the tablet one at a time, while the text remains displayed on the screen for the child to be able to retrieve relevant information. In the process, the tablet keeps track of time-aligned multimodal data: voice recording, finger sliding time, time of reading, time of question answering, and number of correct answers. Data are recorded, stored locally, sent to a server through an internet connection, and processed remotely by a battery of cloud-based services, analysing data automatically to produce a detailed quantitative signature of each reading session. A server-based database aggregates anonymised data to make them available for specialists. Also individual's longitudinal profiles are stored, for them be queried and inspected upon authorised access.

The project will avail itself of sophisticated Natural Language Processing (NLP) techniques aimed at the automatic modelling and assessment of text complexity, with a view to providing an estimation of text readability, and the development of advanced readability measures (Collins-Thompson, 2014; Dell'Orletta et al., 2011). A children's speech recognition system (Cosi, 2015; Gerosa et al., 2007) will be able to check if a specific read word is rendered correctly, and offer an overall accuracy score for text decoding. Text annotation and formatting tools will also help teachers select and deliver text stimuli with controlled and gradually increasing levels of linguistic difficulty, thus supporting more targeted potentiation. We expect this to improve response to treatment, reduce downtime between successive intervention steps, minimise repetition of overlearned tasks, and increase motivation.

Keywords: NLP-based methods, reading efficiency, early years

References

- Cappa, C., Giulivi, S., and Muzio, C. (2016). L'efficienza di lettura: l'integrazione dell'abilità di comprensione del testo con quella della velocità di lettura. In *Proceedings of the XXV AIRIPA National Conference*.
- Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.
- Cornoldi, C. and Colpo, G. (2012). *Nuove Prove di Lettura MT - Sec. di I grado*. Giunti O.S.
- Cosi, P. (2015). A kaldi-dnn-based asr system for italian. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, pages 1–5. IEEE.
- Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011). Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83. Association for Computational Linguistics.
- García-Madruga, J., Vila, J., Gómez-Veiga, I., Duque, G., and Elosúa, M. (2014). Executive processes, reading comprehension and academic achievement in 3th grade primary students. *Learning and individual differences*, 35:41–48.
- Gerosa, M., Giuliani, D., and Brugnara, F. (2007). Acoustic variability and automatic recognition of children's speech. *Speech Communication*, 49(10-11):847–860.
- Gough, P. B. and Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and special education*, 7(1):6–10.
- Hoover, W. A. and Gough, P. B. (1990). The simple view of reading. *Reading and writing*, 2(2):127–160.
- Shinn, M. R. and Shinn, M. M. (2002). *AIMSweb® training workbook*. Eden Prairie, MN: Edformation.
- Speece, D. L., Ritchey, K. D., Silverman, R., Schatschneider, C., Walker, C. Y., and Andrusik, K. N. (2010). Identifying children in middle childhood who are at risk for reading problems. *School psychology review*, 39(2):258.
- Wagner, R. K., Torgesen, J. K., Rashotte, C. A., and Pearson, N. A. (2009). *TOSREC: Test of Silent Reading Efficiency and Comprehension*. MHS Assessments.

Figurative language processing: A developmental and NLP Perspective

Mila Vulchanova

Norwegian University of Science &
Technology

mila.vulchanova@ntnu.no

Valentin Vulchanov

Norwegian University of Science &
Technology

valentin.vulchanov@ntnu.no

Abstract

It is now common to employ evidence from human behaviour (e.g., child development) for the creation of computational models of this behaviour with a variety of applications (e.g., in developmental robotics). In this paper we address research in the comprehension and processing of figurative (non-literal) language in highly verbal individuals with autism in comparison with age- and language level-matched neuro-typical individuals and discuss critically what factors might account for the observed problems. Based on this evidence we try to outline the strategies used by human language users in understanding non-literal/non-compositional expressions and proceed to identifying possible solutions for automated language systems in the domain of idiomatic expressions.

1. Introduction

Figurative or non-literal language is a pervasive phenomenon in every-day human communication. It covers a wide range of expressions, such as idioms, metaphors, irony and jokes, hyperbole, indirect requests, as well as other stereotyped expressions, such as clichés. A recent study investigating the incidence of non-literal expressions in e-mails written by young people found that 94,30% of the e-mails included at least one non-literal statement and participants used on average 2,90 non-literal expressions per e-mail (Whalen, Pexman & Gill, 2009). Unlike literal language, where the interpretation depends on computing the meaning of each of the expression constituents, figurative expressions seem to require additional operations in order to arrive at the intended meaning. Furthermore, the competences and skills associated with figurative language mastery take much longer to develop than word knowledge (vocabulary) or core grammar. In typical language development, children appear to demonstrate appreciation for figurative expressions, such as idioms, at some point in the school years (Nippold, 1998; 2006; Nippold & Duthie, 2003; Cain et al., 2009, Levorato & Cacciari, 1995). Interestingly, this ability patterns in a way similar to the emergence of dimensionality in language competences and skills, as established recently in a large-scale cohort study covering pre-school to early school ages (LARRC, 2015). In that study, three clear dimensions of language competence (vocabulary, grammar and discourse) are first distinguished around third grade at school. It is still unclear, however, whether the developmental trajectory displays a linear trend over time (Nippold, 1998, 2006) or a quadratic trend peaking before adolescence, with less change afterwards (Kempler et al., 1999; Vulchanova et al., 2011; Laval & Bernicot, 2002). More intriguingly, figurative language skills, by taking longer to acquire, also manifest vulnerability both in developmental deficits and across the life-span. Research in typical ageing demonstrates that older adults produce fewer idioms and benefit more from cueing than younger speakers (Conner et al., 2011). Findings from acquired deficits, such as aphasia demonstrate impaired idiom comprehension (Milburn, Warren & Dickey, 2018). Problems with figurative language have been systematically documented in developmental deficits, such as autism spectrum disorder (ASD) (Volden & Phillips, 2010; Ramberg, Ehlers, Nydén, Johansson and Gillberg, 2011). Recent studies establish failure to understand pragmatic, non-literal aspects of language, such as metaphors, idioms and other forms of figurative language, even when structural language may appear to be intact (Gold and Faust, 2010; for a

Keywords: Figurative language, NLP, autism

comprehensive critical review of converging evidence from existing research see Vulchanova et al., 2015).

From a developmental and cognitive perspective the question then is what makes figurative (non-literal) language more challenging and open to such vulnerability. Given this vulnerability and complexity, a second related question is what can explain the high prevalence of non-literal language in discourse. Below we briefly present current accounts of figurative language processing and the main factors that impact on figurative language interpretation.

2. Accounts of figurative language processing

Unlike literal language, figurative language is non-transparent. It requires the hearer to go beyond the meaning of the individual constituent words in order to decode the speaker's intended meaning. Idioms, for instance, have lost their original semantic motivation, and need to be stored as multi-word expressions, very much like the way speakers store lexical items in long-term memory for the purposes of access and retrieval. Thus, expressions, such as *hit the sack* (literally «go to bed») or *kick the bucket* (literally «die») cannot be decoded solely based on co-composing the meaning of the verb (*hit*, *kick*) and its complement (*sack*, *bucket*), and the relationship between the head verb and its complement is not the same as between that same verb and an argument this verb subcategorises for (Nunberg et al., 1994). Yet, many idioms allow for partial analysability, in that one of its components functions according to its typical collocational environment, while it is the other constituent which requires a metaphorical interpretation, e.g., «the question» in *pop the question*. It has also been suggested that even in completely non-transparent expressions, such as *kick the bucket*, the head verb «kick» preserves its aspectual features preventing the acceptability of sentences, such as «??John lay kicking the bucket due to his chronic illness» or using an adverb which is inconsistent with the punctual feature of *kick* (Glucksberg, 1991; Hamblin & Gibbs, 1999). These properties of idioms have given rise to two types of accounts, non-compositional accounts, which acknowledge the need to store and retrieve idioms as multi-word chunks, and compositional accounts, which focus on the possibility of individual constituents to affect the interpretation or usage of the idiom.

2.1. Non-compositional accounts

A variety of accounts assume that idioms share similar properties, and are processed as, lexical items or multi-chunk words. Such approaches build on observations that the meaning of idioms is not a function of the meaning of their individual constituents and that language users need to acquire and store these expressions very much like words. In the linguistics tradition, this non decomposability issue has been addressed in numerous studies highlighting the impossibility to generate the expression following the rules of phrase structure and further to modify or change the structure (e.g., derive passives, insert modifiers etc. (Chomsky, 1980; Nunberg et al., 1994; Jackendoff, 2002). Non-compositional approaches differ on a range of parameters and specific assumptions. Thus, some approaches adopt single step processing, while others assume step-wise processing. According to the *standard pragmatic approach*, the first step in processing involves activating the literal meanings associated with the constituent words in the expression, only arriving at the intended figurative non-literal interpretation at the second step (Grice, 1975). In contrast, the *direct access model* assumes that there is no need for the initial activation of the literal meaning(s) (Gibbs, 1994). Instead, the figurative meaning is retrieved directly, following cues from the linguistic and other (e.g., communicative) context of the expression. The lexical nature of idioms has been recognised also in the *lexical representation hypothesis* (Swinney & Cutler, 1979), where idioms are assumed to be stored as lexical items, which can be retrieved fast, at the same time engaging in a second parallel process of lexical decomposition.

2.2. Compositional accounts

In contrast, compositional approaches rest on the assumption that idiom processing involves, at least at some level, de-composition. Thus, Hamblin & Gibbs (1999) insist on idiom decomposability, suggesting that idiom interpretation depends on identifying the individual constituents of the expression. Such approaches recognize that the processing and understanding of idioms cannot be reduced to lexical access or lexical retrieval only (Cacciari and Tabossi, 1988; Gibbs, 1992; Vega-

Moreno, 2001). A similar, albeit technically different, approach, is the *configuration hypothesis*. On this approach, idioms are represented in a distributed way and they are processed as complex expressions, very much like other instances of similar syntactic complexity (Cacciari and Tabossi, 1988).

2.3. The hybrid model

There is evidence in current research supporting both non-compositional and compositional approaches. The central question addressed in most studies is whether at all, in order to arrive at the target interpretation, speakers retrieve the literal aspects of constituents or by-pass this step, and retrieve the target non-literal interpretation. In a critical review of existing research and based on an experimental study, Titone & Connine (1999) propose a *hybrid model* for idiom processing where the main factor is idiom decomposability. Thus, idioms are assumed to function simultaneously as arbitrary associations between phonological form and meaning (like words), and compositional phrases. Processing of idiomatic expressions will then depend on the degree to which the idiom is inherently analyzable into constituent parts. It deserves mention here that compositionality of interpretation and the extent to which there is a need for literal processing have been brought to bear in models of the processing of other types of figurative language, for instance verbal irony and jokes (Dews & Winner, 1999).

3. Factors in figurative language processing

It becomes clear that the processing and interpretation of idioms depends on a number of key properties. Following Nunberg et al. (1994), three main factors have been identified, compositionality, conventionality and transparency. *Compositionality* applies to the degree to which the expression can be analysed into constituent parts or put differently, whether the expression reflects any syntactic and semantic structure. For instance, for an idiom comprising a head verb and its object (*spill the beans*), the question is whether this verb phrase structure is accessible to the speaker, and is it likely that speakers are using it for the processing of the idiom. *Conventionality* applies to the degree to which the expression has become lexically encoded and part of the lexical inventory of the language at hand. Conventionalisation is a natural process in language evolution and reflects the societal forces in language practice, whereby certain forms become adopted by the community of speakers by virtue of social agreement (de Saussure, 1964). *Transparency* applies to the extent to which the semantic motivation of the expression is evident. Even though these properties can be stated independently, they are not easily operationalisable, and quite often overlap. For instance, the difference between compositionality and transparency is often obscured by semantic judgements on whether the target interpretation is easily accessible or not. Furthermore, other criteria have been shown to play a role. For instance, *familiarity*, as established in population norming studies has a documented effect on processing latencies, as well as accuracy. Since language learning and use largely depend on frequency of exposure to linguistic input, frequency can be assumed to play a role in the processing of figurative expressions. Idiom familiarity and the frequency of constituent words interact, as evidenced in studies where the magnitude of the idiom familiarity effect seems to be diluted when the idiom contains low-frequency words as constituents (Cronk, Lima & Schweigert, 1993). Importantly, both the frequency of the individual constituents and their collocational frequency would play a role in how fast speakers access the target meaning, whether they retrieve literal meanings at all, and whether this process is marked by a competition between target figurative and literal interpretation. We address this question in more detail in 5. Below.

Conventionality/novelty often correlates highly with familiarity and frequency. However, there is no consensus on what test can be used as an objective measure of familiarity and how it can be operationalised (cf. Thibodeau, Sikos & Durgin, 2017 for a discussion). While some authors have used subjective measures, such as e.g., «the perceived experience with the metaphor» (Blasko & Connine, 1993), others suggest that frequency (measured in web corpora) can be used as an objective measure due to its high correlation with familiarity (Thibodeau & Durgin, 2011). In addition, conventionality and familiarity often do not yield clear independent effects in experimental research (Dulcinatti, Mazzarella, Pouscoulous & Rodd, 2014). Thus, a central methodological issue in research on figurative language is how to operationalize the factors that play a role in idiom processing and how to

establish objective measures to be used in experimental designs or modeling. Needless to say, a final crucial factor in idiom processing, is the presence of biasing *context*. Numerous studies have found effects of context which might bias for the target figurative interpretation of the expression or not.

4. Findings from research in ASD

The diversity and complexity of factors involved in the processing and comprehension of figurative language may be specifically challenging in developmental disorders, such as autism. Problems in this domain are well-attested (Tager-Flusberg, 2006; Volden & Phillips, 2010; Vulchanova, Saldaña, Chahboun & Vulchanov, 2015), however, their source remains largely controversial. In a series of specifically designed studies, we investigated performance on figurative language tasks (both idioms and metaphors) in highly verbal individuals with autism in comparison with IQ- and language ability-matched neuro-typical individuals. The participants in those studies came from two age groups, 10-12 (children) and young adults in the range 16 – 22 years analysed in a cross-sectional design. The two age ranges and the cross-sectional design were included specifically to establish possible developmental trajectories in both controls and experimental group.

Our main findings can be summed up in the following way. The main problems encountered by the participants with autism were primarily reflected in significantly greater reaction latencies in comparison to controls. The participants with autism performed at adequate levels of accuracy, though still displaying poorer responses in comparison to controls. Another major finding is the different developmental trajectories between the experimental groups and controls: young adult participants with autism performed at the level of control children, but better than children with autism, as evidenced by main effects of Age and Group in our results. We also have evidence of different underlying strategies in the processing of figurative language and in text comprehension.

A main finding in that research is that young adults with autism are less accurate than adults without autism. A valid question then is what types of errors are they making. The results in Chahboun et al. (2016b) suggest that the responses they provide are more literal. In this study a difference in degree of literalness was observed in response accuracy. The model revealed a main effect of group (control/ASD) ($\chi^2(1, 26) = 5.22, p = .022$), with more literal responses by participants with autism, and a marginally significant difference in accuracy between Age (children/young adults) ($\chi^2(1, 26) = 3.51, p = .06$). Furthermore, a two-way interaction between age and group was observed ($\chi^2(1, 26) = 4.89, p = .02$). Multiple comparisons with Tukey contrasts revealed that this interaction was due to a significant difference between control young adults and young adults with autism ($p = .015$), where the young adults with autism were converging on more literal responses than their typically developing peers. Thus, the younger participants and participants with autism in our study interpreted the stimuli more often literally than older participants and controls. These data provide support for other findings in research on young children and individuals with autism documenting a tendency for literal interpretation (Mitchell, Saltmarsh & Russell, 1997). Data from the same study further suggest that younger participants and participants with autism have specific problems with the idioms with greater decomposability, but no similar problems were observed with novel decomposable metaphors of literal expressions. This comes to suggest that idiom decomposability interferes in certain ways with processing and interpretation, unlike other decomposable expressions. These data find support in a recent ERP study of idiom comprehension in Chinese, whereby decomposable expressions, both idioms and free (literal) expressions elicited significantly greater ERP responses than the non-decomposable idioms.

5. Figurative language and NLP

On the backdrop of factors involved in the processing and comprehension of non-literal language in typical individuals, and the problems observed in highly verbal individuals with autism, a possible approach needs to look at what features of the expression might trigger literal (compositional) strategies, thus procrastinating the target figurative interpretation. We aim to outline under what conditions this is more likely to happen.

The central factor that needs to be considered is idiom decomposability and to what extent properties of the constituent words might trigger competition between a literal and target figurative

meaning. Some studies have used the notion of *semantic plausibility* suggesting that, in the absence of a biasing context, both a literal and a figurative interpretations are equally plausible. Thus, for instance the idiom *pull someone's leg* is equally plausible in a direct literal way, and figuratively. Other expressions do not easily yield such interpretations. For instance, *I am a bit under the weather today* cannot possibly make sense on a literal interpretation. It is to be expected that only the semantically plausible idioms may trigger competition between the literal and the figurative meaning, whereas the less plausible ones will directly cue the target figurative meaning, which is also the only plausible interpretation to be accessed. Another approach would be to assess the semantic similarity or closeness between the two available interpretations, the literal and the idiomatic (Milburn, 2017). This type of approach holds promise in circumventing issues arising from the need to categorise expressions according to decomposability, conventionality, transparency and other parameters. On this approach, frequency can be easily added in the equation to assess the collocational probability of one part of the expression co-occurring with the other part in comparison to collocational alternatives, e.g., the same word co-occurring with other lexical items.

To give a concrete example of an idiom like *kick the bucket*, one can estimate the probability of the noun phrase “the bucket” co-occurring with the head verb “kick” against the probability of the same verb co-occurring with other possible fillers of the complement position. The main measure can be **cloze probability** of dependent constituent, and we can assume that degree of activation of possible candidates for the verb complement position will depend on the ratio of cloze probabilities. In the case of idioms, this may be the ratio between the NP filler in the idiom and the most frequent literal filler of the argument position expressed in formula like $ClProbMax\ NP(VP) / CP\ NP(idiom)$. For instance, counts of these in the case of *kick the bucket* reveal that the most frequently occurring filler in the context of kick is “the ball” with native speaker cloze probability counts at 54%, and estimates from on-line corpora at 14.58%. Concerning the collocational frequency of “the bucket” as argument filler in the idiom, counts vary depending on corpus. A search in CoCA gives a value of 5.7 for **ball/bucket** suggesting that ball is by far more frequent after *kick*. Since corpora produce different results, cloze probabilities can be estimated in norming studies with native speakers, especially given the high correlation between cloze probabilities measured in sentence completion tasks with native speakers and in on-line (web) corpora (Hammerås, 2017).

We can then assume that the likelihood that literal interpretations may be activated can be measured as the value of the ratio between the cloze probabilities of the two argument filler candidates, and bigger values (according to the formula above) will lead to greater competition between the literal and figurative meaning as a result of more likely literal activation of head word/verb. This type of approach can be tested experimentally in a controlled design with carefully selected stimuli.

6. Conclusions

In this paper we have addressed issues arising from the factors which impact on the processing of figurative language against common assumptions and accounts. Based on evidence from behavioural research with idiom and metaphor processing in autism and typical individuals, we have proposed an approach which captures the common problems encountered by special populations, and often children, in the processing of non-literal language, at same time offering a solution to how to operationalize idiom processing in a measurable and meaningful way, consistent with how the human brain may be handling the task. Such an approach is testable, and as such, can be useful for computational modeling of natural language processes.

References

- Blasko, D. G., & Connine, C. M. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(2), 295-308.
- Cacciari, C. & Tabossi, P. (1988). The comprehension of idioms. *J Mem Lang.*, 27:668–83

- Cain, K., Towse, A. S., and Knight, R. S. (2009). The development of idiom comprehension: an investigation of semantic and contextual processing skills. *J. Exp. Child Psychol.* 102, 280–298. doi:10.1016/j.jecp.2008.08.001
- Chahboun, S., Vulchanov, V., Saldaña, D., Eshuis, H. & Vulchanova, M. (2016b). Can You Play with Fire and Not Hurt Yourself? A Comparative Study in Figurative Language Comprehension between Individuals with and without Autism Spectrum Disorder. *PLoS ONE* 11(12): e0168571. <https://doi.org/10.1371/journal.pone.0168571>.
- Chomsky, N. (1980). *Rules and representations*. New York: Columbia University Press.
- Conner, P. S., Hyun, J., O'Connor Wells, B., Anema, I., Goral, M., Monéreau-Merry, M., Rubino, D., Kuckuk, R. & Obler, L. K. (2011). Age-related differences in idiom production in adulthood. *Clinical Linguistics & Phonetics*, 25(10), 899-912.
- Cronk, B., Lima, S. & Schweigert, W. (1993). Idioms in sentences: Effects of frequency, literalness, and familiarity. *Journal of Psycholinguistic Research*, 22, 59. <https://doi.org/10.1007/BF01068157>
- de Saussure, F. (1964). *Course in General Linguistics*, 2nd impression. London: Peter Owen.
- Dews, S. & Winner, E. (1999). Obligatory processing of literal and nonliteral meanings in verbal irony. *Journal of Pragmatics* 31, 1579-1599
- Dulcinati, G., Mazzarella, D., Pouscoulous, N. & Rodd, J. (2014). Processing metaphor: The role of conventionality, familiarity and dominance. *UCL Working Papers in Linguistics*, 26, 72-88.
- Gibbs, R. W. (1992). Categorization and metaphor understanding. *Psychol. Rev.* 99, 572–577. doi: 10.1037//0033-295x.99.3.572
- Gibbs R. (1994). *The poetics of mind: Figurative thought, language and understanding*. Cambridge: Cambridge University Press.
- Glucksberg, S. (1991). Beyond literal meanings: The psychology of allusion. *Psychological Science* 2: 146-152.
- Gold, R. & Faust, M. (2010). Right hemisphere dysfunction and metaphor comprehension in young adults with Asperger syndrome. *J. Autism Dev. Disord.* 40, 800–811. doi: 10.1007/s10803-009-0930-1
- Grice P. (1975) Logic and conversation. In Cole P, Morgan J (eds.). *Syntax and Semantics 3: Speech Acts*. New York: Academic Pressp, 41-58.
- Hamblin, J. & Gibbs, R. (1999). Why You Can't Kick the Bucket as You Slowly Die : Verbs in Idiom Comprehension. *Journal of Psycholinguistic research*, 28(1): 25-39.
- Hammerås, M. L. (2017). Probing sensitivity to argument structure in two proficiency level groups - an exploratory study with Norwegian learners of English. MA thesis, NTNU. (<https://www.ntnu.edu/documents/1535402/35615794/Hammer%C3%A5s2017/0c92037d-7784-4bf5-9d76-53bb1872e861>)
- Jackendoff, R. (2002). What's in the lexicon? In: Nooteboom S, Weerman F, Wijnen F (eds) *Storage and computation in the language faculty*. Dordrecht: Kluwer, 23–58.
- Kempler, D., Van Lancker, D., Marchman, V., and Bates, E. (1999). Idiom Comprehension in children and adults with unilateral damage. *Dev. Neuropsychol.* 15, 327-349. doi:10.1080/87565649909540753
- LARRC. (2015), The dimensionality of language ability in young children. *Child Development*, 86: 1948–1965. doi: 10.1111/cdev.12450

- Laval, V., and Bernicot, J. (2002). Tu es dans la lune: understanding idioms in French speaking children and adults. *Pragmatics* 12, 399–413.
- Lavorato, M. C. & Cacciari, C. (1995). The effects of different tasks on the comprehension and production of idioms in children. *J. Exp. Child Psychol.* 60, 261–283. doi: 10.1006/jecp.1995.1041
- Milburn, E. (2018). *The effects of meaning dominance and meaning relatedness on ambiguity resolution: Idioms and ambiguous words*. Doctoral Dissertation, University of Pittsburgh.
- Milburn, E., Warren, T. & Dickey, M. W. (2018). Idiom comprehension in aphasia: Literal interference and abstract representation. *Journal of Neurolinguistics* (<https://doi.org/10.1016/j.jneuroling.2018.02.002>)
- Mitchell, P., Saltmarsh, R. & Russell, H. (1997). Overly Literal Interpretations of Speech in Autism : Understanding That Messages Arise from Minds. *The Journal of Child Psychology and Psychiatry*, 38(6), 685–91.
- Nippold, M. (1998). *Later Language Development: The School-Age and Adolescent Years*. 2nd Edn. Austin, TX: Pro-Ed.
- Nippold, M. A. (2006). “Language development in school-age children, adolescents and adults,” in *Encyclopedia of Language and Linguistics* (Vol.6), 2nd Edn. Ed K. Brown (Oxford, UK: Elsevier Publishing), 368–372.
- Nippold, M. A., and Duthie, J. K. (2003). Mental imagery and idiom comprehension: a comparison of school-age children and adults. *J. Speech Lang. Hear. Res.* 46, 788–799. doi:10.1044/1092-4388(2003/062)
- Nunberg, G., Sag, I., & Wasow, T. (1994.) “Idioms”. *Language*, 70, 3, 491-538.
- Ramberg, C., Ehlers, S., Nydén, A., Johansson, M. & Gillberg, C. (2011). Language and pragmatic functions in school-age children on the autism spectrum. *International Journal of Language & Communication Disorders*, 31, 387-413.
- Swinney D. & Cutler A. (1979). The Access and Processing of Idiomatic Expressions. *Journal of Verbal Learning and Verbal Behavior*, 18, 523–34.
- Tager-Flusberg, H. (2006). Defining language phenotypes in autism. *Clinical Neuroscience Research*, 6, 219-224.
- Thibodeau, P. H., & Durgin, F. H. (2011). Metaphor aptness and conventionality: A processing fluency account. *Metaphor and Symbol*, 26(3), 206-226.
- Thibodeau, P. H., Sikos, L., & Durgin, F. H. (2017). Are subjective ratings of metaphors a red herring? The big two dimensions of metaphoric sentences. *Behavior Research Methods*, 1-14.
- Titone D. & Connine C. (1999). On the compositional and noncompositional nature of idiomatic expressions. *J Pragmat*, 31(12), 1655–74.
- Vega-Moreno, R. E. (2001). Representing and processing idioms. *UCLWPL* 13, 70–109.
- Volden, J. & Phillips, L. (2010). Measuring Pragmatic Language in Speakers with Autism Spectrum Disorders: Comparing the Children’s Communication Checklist 2 and the Test of Pragmatic Language. *American Journal of Speech-language Pathology*, 19 (3), 204-212.
- Vulchanova, M., Vulchanov, V. & Stankova, M. (2011). Idiom comprehension in the first language : a developmental. *Vigo International Journal of Applied Linguistics*, 8, 207–34.
- Vulchanova M., Saldaña D., Chahboun S. & Vulchanov V. (2015). Figurative language processing in atypical populations : the ASD perspective. *Frontiers in Human Neuroscience*, 9(24), 1–11.

Whalen, J. M., Pexman, P. M & Gill, A. J. (2009) “Should Be Fun—Not!” Incidence and Marking of Nonliteral Language in E-Mail. *Journal of Language and Social Psychology*, 28, 3, 263-280.

Abstractive Text Summarization with Application to Bulgarian News Articles

Nikola Taushanov

Faculty of Mathematics and Informatics
Sofia University St. Kliment Ohridski
nktaushanov@gmail.com

Ivan Koychev

Faculty of Mathematics and Informatics
Sofia University St. Kliment Ohridski
koychev@fmi.uni-sofia.bg

Preslav Nakov

Qatar Computing Research Institute
HBKU
pnakov@qf.org.qa

Abstract

With the development of the Internet, a huge amount of information is available every day. Therefore, text summarization has become critical part of our first access to the information. There are two major approaches for automatic text summarization: abstractive and extractive. In this work, we apply abstractive summarization algorithms on a corpus of Bulgarian news articles. In particular, we compare selected algorithms of both techniques and we show results which provide evidence that the selected state-of-the-art algorithms for abstractive text summarization perform better than the extractive ones for articles in Bulgarian. For the purpose of our experiments we collected a new dataset consisting of around 70,000 news articles and their topics. For research purposes we are also sharing the tools to easily collect and process such datasets.

1. Introduction

Text summarization is the task of creating a shorter version of a given text that retains the most important pieces of information. There are two major approaches for automatic text summarization: abstractive and extractive. The latter selects different parts (sentences) of the text in order to construct the summary. On the other hand, the abstractive summarization is considered closer to the way people approach the problem: they first analyze and understand the input and then generate the content of the summary.

Recent studies (Nallapati et al., 2016) have shown that abstractive summarization methods perform better than extractive ones, but it is clear that the two approaches have different problems, which still need to be solved. Abstractive summaries are often unable to provide accurate factual details and they also tend to repeat words or sentences as shown in (See et al., 2017). Extractive summaries on the other hand have problems related to the fact that sentences cannot easily be separated from the context, especially when they contain references to others which are not extracted.

A great amount of work has been done on applying, evaluating, and improving these models in English, but not a lot of research exists for other languages. This document shows results of applying effective methods in both abstractive and extractive text summarization on a big corpus of news articles in Bulgarian. It should also serve as a starting point for future research in this language.

In the rest of this work, Section 2 provides more context on the related work and how ours fits in it. Section 3 has more details on the models which will be used. The experiments and the results of applying them are shown in Section 4, and in Section 5 we provide qualitative analysis on the produced output. Section 6 concludes our work and gives direction for future developments.

2. Related work

The majority of the research made in the past in the area of text summarization focuses on extractive methods. Their goal is to extract important sentences and use them to form summaries. Earlier research

is based on features of the sentences such as position in the text, frequency of the words or sentences mostly based on TF-IDF (Edmundson, 1969; Baxendale, 1958). Some of the best results were achieved when the text is represented as a graph, in which each sentence is a node and the edges and their weights are based on similarity metrics. The problem then shifts to finding the most important or most central node in the graph. The node degree is used in (Freeman, 1978) and eigenvector centrality in (Bonacich, 1972). Variation of the eigenvector centrality is used for PageRank in (Page et al., 1999). The same graph algorithm but a different distance function adapted for sentences is used in (Erkan and Radev, 2004) and in (Mihalcea and Tarau, 2004). An algorithm similar to PageRank, which uses absorbing Markov chains to encourage diversity among the top ranked nodes, is proposed in (Zhu et al., 2007).

With the recent development of large computing resources, the enormous amount of available data online, and the research and advancements made in the area of deep neural networks, the focus falls back to abstractive summarization. Initially, for the problem of machine translation, some of the first works which apply encoder-decoder networks are (Cho et al., 2014b; Cho et al., 2014a). In (Sutskever et al., 2014) they show promising results by using Long Short-Term Memory (LSTM) units (Hochreiter and Schmidhuber, 1997) and reversing the input sequence. In (Schuster and Paliwal, 1997) bidirectional recurrent neural networks (RNNs) are used for the first time, to the best of our knowledge. (Bahdanau et al., 2014) show that using attention in bidirectional RNNs improves the encoder-decoder performance even further.

In (Rush et al., 2015), inspired by the machine translation models, they use neural attention model for abstractive text summarization. In (Nallapati et al., 2016; See et al., 2017) switching generator-pointers are applied in order to solve the common problems of inaccurately reproducing details, inability to use out-of-vocabulary words and repetition of words or sentences.

In this work, we use a model architecture similar to (Bahdanau et al., 2014) but we also apply multi-layer bidirectional RNNs as in (Vinyals et al., 2014). The solution also addresses the very common problem of covariant shifting in deep neural networks using layer normalization (Ba et al., 2016). We apply this model on a novel corpus in Bulgarian and compare the results with extractive models which implement TextRank with a few different similarity metrics.

3. Implemented Methods for Summarization

In this section, we describe the selected extractive models and the proposed abstractive ones.

3.1. Extractive Summarization

The extractive summarization methods identify the most important parts of the text, extract them, and then use them to create a summary. Some of the best results in this area are observed in models which use a modification of the PageRank algorithm (Page et al., 1999). We have chosen to implement and apply TextRank (Mihalcea and Tarau, 2004) using two different similarity metrics.

It is a graph-based ranking algorithm for computing relative importance of vertices within a graph. Each vertex V_i is essentially a sentence from the input text. The weight w_{ij} of an edge between two nodes V_i and V_j in the graph is defined by the value of a similarity metric applied on their corresponding sentences. To build a weighted score WS for each node V_i , the algorithm uses the slightly modified PageRank formula (1), where d is the damping factor, with value between 0 and 1, used for modeling the probability of jumping from a given vertex to another random one. Each sentence is then ranked based on the score of its node.

$$WS(V_i) = (1 - d) + d \sum_{j \in In(V_i)} \frac{w_{ij}}{\sum_{v_k \in Out(V_j)} w_{jk}} WS(V_j) \quad (1)$$

In the first model - *TR*, we use a similarity metric which measures the content overlap of two sentences and it is the one proposed in the original work (Mihalcea and Tarau, 2004). Given two sentences S_i and S_j , each represented as $S_i = s_1^i, s_2^i, \dots, s_{|S_i|}^i$, the similarity between them is defined in (2)

$$Similarity(S_i, S_j) = \frac{|\{s_k \mid s_k \in S_i \& s_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)} \quad (2)$$

Representing the sentences using the vector space model (Salton et al., 1975) creates TF-IDF weighted vectors for each of them. A cosine similarity on those vectors is used in the second model called TR-cosine. This is similar to the solution proposed in (Erkan and Radev, 2004) but without the binarization of the graph weights which they propose.

Regardless of the similarity metric, the algorithm scores each sentence, ranks them accordingly and picks the top scoring ones for the summary.

3.2. Abstractive Summarization

The state-of-the-art abstractive summarization models use sequence-to-sequence with attention networks in order to read the input text and then generate a summary word by word. The baseline model, which we have implemented and applied, is very similar to the one proposed in (Nallapati et al., 2016). It is the recurrent neural network with encoder-decoder architecture which is depicted in Figure 1. Each word of the text X_0, X_1, \dots, X_n , is first transformed using a word embeddings layer. It is then fed to the multi-layered bidirectional encoder. In comparison (Nallapati et al., 2016) uses a single layer.

Also, each cell in both the encoder and the decoder is implemented with a LSTM unit instead of Gated Recurrent Unit (Cho et al., 2014b). The last layer of the encoder is connected to the decoder using attention as in (Bahdanau et al., 2014) and calculated with (3) and (4)

$$e_i^t = v^T \tanh(W_h h_i + W_s s_t + b_{attn}) \tag{3}$$

$$a^t = \text{softmax}(e^t) \tag{4}$$

where v , W_h , W_s and b_{attn} are learnable parameters, h_i is the hidden state of the encoder at encoding step i , s_t is the decoder state and a^t is the attention vector at timestep t .

The result from the decoder is transformed back to a word using a projection layer. We will refer to this baseline model with *s2s-lstm*.

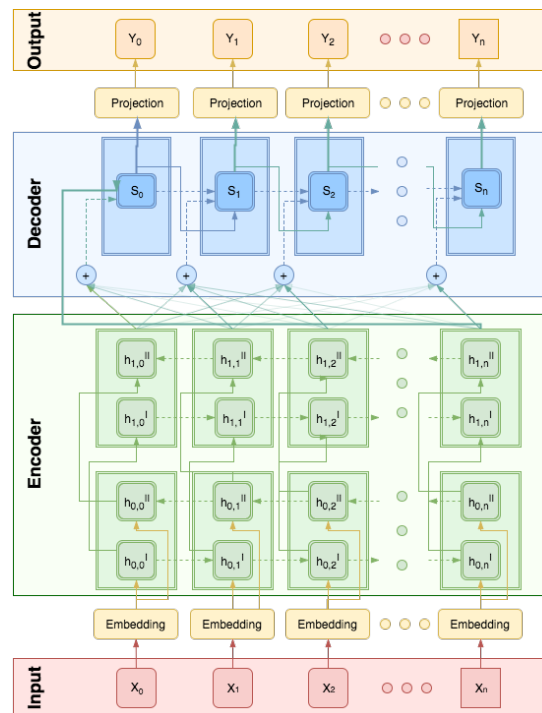


Figure 1: Encoder-decoder RNN with embeddings, multi-layered encoder and attention.

We will extend the baseline model with dropout (Srivastava et al., 2014) for better regularization of the network and call it *s2s-lstm+d*. The dropout factor is part of the hyper parameters of the model.

The final modification which we have made, which leads to model $s2s-lstm+d+ln$, is a network layer normalization (Ba et al., 2016). This technique is often used to solve the problem of covariant shifting in deep neural networks.

All of the selected models will use gradient clipping (Pascanu et al., 2012) to help with the exploding gradients, Xavier weights initialization (Glorot and Bengio, 2010) for an improved network initialization and Adagrad gradient descent (Duchi et al., 2011) for better learning. The loss function is a weighted cross-entropy for a sequence of logits with sampling (Jean et al., 2014) because the output of the network is a sequence of words and the size of the vocabulary is usually very big.

As usual for the sequence-to-sequence networks, the input fed into the decoder during the training starts with a special word for start of sentence and ends with a special word for end of sentence. During the decoding phase, each output of the decoder is used as input on the next step until an end of sentence is generated. The usual beam search approach for sequence-to-sequence networks is used, in order to find the best result (Cho et al., 2014a).

On the other hand, on each step of the training, the words from the actual summary are the input to the decoder. This approach showed better performance and faster training compared to the curriculum learning strategy proposed in (Bengio et al., 2015) which uses the word generated on the previous step for an input to the current one.

4. Experiments and results

4.1. Dataset

FocusNews: As of the time of writing, there is no big enough dataset in Bulgarian which could be used for abstractive text summarization. As part of this work, we created a big corpus which is suitable to run our experiments upon. An appropriate source for this was the FocusNews news agency website (FOCUS, 2018). It contains around 70000 articles at any given moment. Extracting articles for the period from January 2017 to September 2017 resulted in a corpus of 76300 news articles and their headlines. It contains texts with minimal length of 1 word, a maximum of 8854 and average of 20. The size of the vocabulary of words used is 200 000. This dataset will be split into a training set of size 65472, a validation set with 7271 and a test set with 3557 articles. The code for collecting and processing such datasets is available online¹.

Compared to some popular English corpora, where the data is anonymized and the names of towns, people, countries, etc. are replaced with tokens, in this corpus it is not. The articles are very close to those a person would read. Because all the articles and their headlines come from a reputable source, we could easily make the assumption that the headlines contain the most important information and could be used as short summaries of the articles.

DUC-2004: A dataset presented on the DUC 2004 (NIST, 2004) competition has been the default way to experiment and test automatic text summarization in various researches. It consists of a small number of English news articles on different topics with multiple human produced reference summaries for each of them. This dataset is small and unsuitable for training abstractive models but works well for extractive ones.

4.2. Evaluating the FocusNews dataset

FocusNews is a new corpus which has not been used in other research so far. We will compare the results of applying the same models on DUC-2004 and FocusNews using ROUGE-1, ROUGE-2 and ROUGE-L recall, precision and F-scores (Lin, 2004). In Table 1 we show that both datasets are of similar quality. As expected, the models show better scores when applied on DUC-2004, given the fact that its reference summaries are human prepared and well selected. When looking at ROUGE-1 and ROUGE-L, both models show better results for DUC-2004 than FocusNews, but for ROUGE-2 the results are very close to each other. In both datasets TR is better than $TR-cosine$. Despite the difference in the results, the numbers show that FocusNews is a suitable dataset for our experiments.

¹<https://github.com/nktaushanov/focusnews>

	ROUGE-1			ROUGE-2			ROUGE-L		
	R	P	F	R	P	F	R	P	F
DUC-2004: TR	0.292	0.291	0.292	0.046	0.046	0.046	0.256	0.255	0.255
DUC-2004: TR-cosine	0.256	0.255	0.255	0.038	0.038	0.038	0.247	0.247	0.247
FocusNews: TR	0.186	0.153	0.160	0.057	0.049	0.051	0.180	0.149	0.155
FocusNews: TR-cosine	0.147	0.121	0.126	0.044	0.037	0.038	0.144	0.118	0.122

Table 1: Comparing DUC-2004 and FocusNews with extractive summarization

4.2.1. Models evaluation and analysis

The three selected abstractive models have been tested with a variety of different hyper parameters. The best results from all tests were observed with the hyper parameters specified in Table 2.

	s2s-lstm	s2s-lstm+d	s2s-lstm+d+ln
Minimum learning rate	0.01	0.01	0.01
Batch size	50	200	50
Learning rate	0.15	0.15	0.15
Encoding layers	3	2	2
Encoding steps	120	120	120
Decoding steps	40	30	30
Minimum input length	2	2	2
Hidden state size	256	256	256
Embedding dimensions	128	128	128
Max gradient norm	2	2	2
Dropout keep probability	1.0	0.7	0.5
Num of loss samples	4096	4096	4096
Max article sentences	4	4	4
Min article sentences	2	2	2

Table 2: Hyper parameter of the abstractive models

The evaluation of the models on FocusNews presented in Table 3 shows a clear performance difference between the abstractive models *s2s-lstm* and *s2s-lstm+d* compared to *s2s-lstm+d+ln* which has layer normalization. The ROUGE-1, ROUGE-2 and ROUGE-L scores for the last one are almost twice higher compared to the first two models. It is clear that covariant shifting and better regularization make a huge difference in this task. In the case of extractive summarization, *TR* performs better than *TR-cosine* which means that the similarity metric of content overlap is better than the cosine distance of the vectorized sentences.

Comparing the best results from the extractive and abstractive algorithms, it looks like the latter perform better, although the numbers are not that far apart. In each of the recall metrics they produce almost the same results, but the extractive solution performs worse based on precision and F-score.

	ROUGE-1			ROUGE-2			ROUGE-L		
	R	P	F	R	P	F	R	P	F
s2s-lstm	0.102	0.109	0.103	0.012	0.013	0.012	0.102	0.109	0.103
s2s-lstm+d	0.093	0.103	0.096	0.014	0.016	0.015	0.093	0.103	0.096
s2s-lstm+d+ln	0.192	0.198	0.191	0.052	0.053	0.052	0.192	0.198	0.191
TR	0.186	0.153	0.160	0.057	0.049	0.051	0.180	0.149	0.155
TR-cosine	0.147	0.121	0.126	0.044	0.037	0.038	0.144	0.118	0.122

Table 3: Evaluation of all the models on FocusNews

5. Qualitative Analysis

Table 4 shows a couple of good summary examples generated from our best model *s2s-lstm+d+ln*. All of them are correct and do not have any syntactical or semantical problems.

In the first example we can observe how the generated summary is different from any of the sentences in the text but it still has the same meaning as the original. Instead of having the exact same words as in the article, it contains their synonyms - such as *ограничава* (*ogranichava*, "restricts") instead of "спира" (*spira*, "stops") and "в двете посоки" (*dvete posoki*, "in both directions") instead of "в двете платна" (*v dvete platna*, "in both lanes"). Looking closely at the original text and the generated one, it looks like the original sentence was transformed by omitting some of the words which were not important and replacing others with their synonyms.

In the second example, the generated summary has less details but is a good paraphrase. In the original one, the text says that the landfill of the town is on fire whereas in the generated - there is a fire in the area of the town. The latter is a bit more generic, but still preserves the important information. What is more interesting in this case is the fact that the generated summary has a reference to an earlier part of the sentence - "града" (*grada*, "the town") which refers to "пазарджик" (*Pazardzhik*) - the name of the town.

Both examples show impossible to achieve with extractive summarization situations which could often be observed in human made summaries. Unfortunately, the widely used scoring metrics cannot measure very well those solutions.

Article	пловдив . спира се движението на превозни средства в двете платна на бул . " александър стамболийски " в пловдив заради ремонт на водопроводната мрежа . това съобщиха за радио " фокус " – пловдив от оп " организация и контрол на транспорта " . затворен ще бъде участъкът от ул . " никола димков " до ул . " стефан стамболов " на 16.08.2017г /сряда/ от 9:00 ч до 16:00 часа . маршрутите на автобусни линии # 16 , 20 , 27 и 36 от вътрешноградския транспорт също се променят . цветана тончева
Original	пловдив : спират движението по бул . " александър стамболийски " в града заради ремонт на 16 август
Generated	пловдив : ограничава се движението в двете посоки на бул . " александър стамболийски " заради ремонт на водопроводната мрежа
Article	пазарджик . гори сметището на пазарджик , предаде репортер на агенция " фокус " . виждат се пламъци от пътя . задимен е пътят за селата капитан димитриево и дебръщица по посока пътя пазарджик-пещера .
Original	гори сметището на пазарджик
Generated	пазарджик : пожарът в района на града
Article	78-годишна жена от град сливен е станала жертва на телефонна измама . това съобщиха от областната дирекция на мвр – сливен . потърпевшата е била въввлечена в заблуждение , че помага на органите на реда при залавянето на престъпна група , занимаваща се с телефонни измами ...
Original	възрастна жена е станала жертва на телефонна измама
Generated	възрастна жена от града е станала жертва на телефонна измама

Table 4: Examples of good summaries generated from model *s2s-lstm+d+ln*

There is an example of an average summary in Table 5. Its content is ambiguous and has wrong details - "7-те задържани" (*sedemte zadarzhani*, "the 7 detained") instead of "14 обвиняеми" (14 *obvinyaemi*, "14 defendants"). It is still a correct sentence whose meaning is very close to the original one, but the detail mismatch is hard to ignore. This shows that abstractive approaches are generally better at paraphrasing and showing generic information but they are more error-prone when specific details are in place - numbers, places, people, etc.

Article	пазарджик . окръжният съд в пазарджик започна изслушването на компютърни и технически експертизи , извършени от две вещи лица по делото срещу 14 обвиняеми за разпространение идеите на идил , предаде репортер на радио ” фокус ” ...
Original	пазарджик : окръжният съд започна изслушването на компютърни и технически експертизи по делото срещу 14 обвиняеми за разпространение идеите на идил
Generated	пазарджик : окръжният съд започна изслушването на 7-те задържани по делото за разпространение идеите на идил

Table 5: Example of an average summary generated from model $s2s-lstm+d+ln$

Regardless of the good examples, there are lot of bad summaries in the output as well. The one in Table 6 shows a completely erroneous summary which makes no real sense. Other common issues which we observed were incorrect people names, ambiguities, word repetitions, etc.

Article	галерия видин . паметникът на благодарността в центъра на града , пострададал от вандалски акт , е почистен . това съобщиха от пресцентъра на община видин . ” това е възмутително и недопустимо деяние ” , заяви кметът на община видин огнян ценков по повод оскверняването на паметника до стамбол капия , поставен в знак на благодарност...
Original	видин : паметникът на благодарността в центъра на града , пострададал от вандалски акт , е почистен
Generated	видин : паметникът на бензина в центъра на града , пострададали от войните , е почистен

Table 6: Example of a bad summary generated from model $s2s-lstm+d+ln$

6. Conclusions

In this work, we experiment with both extractive and abstractive automatic text summarization and show that the latter performs better on articles in Bulgarian. To the best of our knowledge, no other work so far has applied abstractive summarization in this language. We also propose a novel benchmarked dataset in Bulgarian which is suitable for training and evaluation of abstractive models. Future research would focus on resolving the issues of inaccurate factual details and unnecessary repetition.

References

- Ba, L. J., Kiros, R., and Hinton, G. E. (2016). Layer normalization. *CoRR*, abs/1607.06450.
- Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.
- Baxendale, P. B. (1958). Machine-made index for technical literature: An experiment. *IBM J. Res. Dev.*, 2(4):354–361, October.
- Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. *CoRR*, abs/1506.03099.
- Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 2(1):113–120.
- Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014a). On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259.
- Cho, K., van Merriënboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., and Bengio, Y. (2014b). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078.

- Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, July.
- Edmundson, H. P. (1969). New methods in automatic extracting. *J. ACM*, 16(2):264–285, April.
- Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.*, 22:457–479.
- FOCUS. (2018). *FOCUS Information Agency*. <http://www.focus-news.net/>, Accessed: 2018-02-04.
- Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215 – 239.
- Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. and Titterton, M., Eds., *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May. PMLR.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.
- Jean, S., Cho, K., Memisevic, R., and Bengio, Y. (2014). On using very large target vocabulary for neural machine translation. *CoRR*, abs/1412.2007.
- Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Proc. ACL workshop on Text Summarization Branches Out*, page 10.
- Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into texts. In Lin, D. and Wu, D., Eds., *Proceedings of EMNLP 2004*, pages 404–411, Barcelona, Spain, July. Association for Computational Linguistics.
- Nallapati, R., Xiang, B., and Zhou, B. (2016). Sequence-to-sequence rnns for text summarization. *CoRR*, abs/1602.06023.
- NIST. (2004). *Document Understanding Conference DUC-2004*. <http://duc.nist.gov/duc2004/>, Accessed: 2018-02-04.
- Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report 1999-66, Stanford InfoLab, November. Previous number = SIDL-WP-1999-0120.
- Pascanu, R., Mikolov, T., and Bengio, Y. (2012). Understanding the exploding gradient problem. *CoRR*, abs/1211.5063.
- Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. *CoRR*, abs/1509.00685.
- Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November.
- Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, Nov.
- See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.
- Vinyals, O., Kaiser, L., Koo, T., Petrov, S., Sutskever, I., and Hinton, G. E. (2014). Grammar as a foreign language. *CoRR*, abs/1412.7449.
- Zhu, X., Goldberg, A., Gael, J. V., and Andrzejewski, D. (2007). Improving diversity in ranking using absorbing random walks. *HLT-NAACL*, pages 97–104.

Towards Lexical Meaning Formal Representation by virtue of the NL-DL Definition Transformation Method

Maria Gritz

Herzen State Pedagogical University of Russia,
Faculty of Philology
maria.gritz@yandex.ru

Abstract

The paper represents a part of an extensive study devoted to the issues of lexical meaning formal representation in OWL 2 DL notation. Both theoretical and methodological aspects of lexical meaning formalization within the framework of an ontology are observed in the paper. Model-theoretic semantics paradigm and Kripke model are considered to form a theoretical background for formalization of lexical meaning, whereas the NL-DL definition transformation method is investigated as a method designed to provide us with acceptable formal definitions in OWL 2 DL notation with natural language definitions given at the input. A brief critical study of the method has allowed to reveal particular problematic cases of the method application, which arise due to syntactic peculiarities of natural language definitions given at the input.

1. Introduction

The technology of lexical meaning formal representation is supposed to play the key role within the Semantic Web development since the latter is aimed to extend current Web search engines with applications able to conduct Web content and query analysis based on word meaning processing (Berners-Lee, 2001; Horrocks, 2008; Yu, 2014). In order to reason about the meaning a user renders by a sentence, an inference engine has to use a specific knowledge base – an ontology that represents a scope of formal definitions of domain terms written in a formal language (Horrocks et al., 2007; Ding, 2010). OWL 2 is currently used for that purpose as a formal language, which is expressive enough to give an extensive and accurate specification of lexical meanings of terms representing ontology classes (Hitzler et al., 2012). In the current study the OWL 2 DL extension corresponding to the Description Logic SROIQ is observed (Hitzler et al., 2009).

In OWL 2 DL notation formal definitions of domain terms are given in a form of class descriptions, which are currently derived from natural language texts by means of the transformation method which takes a parsed natural language definition at the input and produces a formal definition in OWL 2 DL corresponding SROIQ notation. The method of NL-DL definition transformation is subjected to a brief critical review, the problematic cases which arise during the method implementation are exemplified with DL-definitions of sociolinguistic terms. The review is preceded by a brief description of theoretical and methodological grounds for lexical meaning formal representation and SROIQ syntax toolkit used for that purpose.

2. Lexical meaning formalization within the model-theoretic semantics paradigm

OWL 2 DL, a DL compatible syntactic fragment of OWL 2, standardized by the World Wide Web Consortium in 2009, is provided semantics within the framework of model-theoretic semantics (Motik et al., 2012; Krötzsch et al., 2014). Model-theoretic semantics allows to ascribe an axiom defining meaning of a natural language expression a set-theoretic interpretation on a domain existing in a set of

possible worlds (Farrugia, 2003). The cornerstone of lexical meaning representation within the model-theoretic semantics paradigm was laid by R. Carnap. R. Carnap made a contribution by specifying the Frege's vague sense to reference dichotomy, which was used to distinguish a denoted object and a mode of reflection of the object in a description and to define the semantic difference between two expressions designating the same thing. The notion of intension that exhibits sense is opposed to extension denoting the scope of a lexeme's referents feasible in a domain, which is observed in a scope of states of affairs. The idea of possible states was used by R. Carnap to define an intension as a semantic feature that provides an identity of a lexeme's extension in all observed states of affairs (Carnap, 1947).

The notion of intension as proposed by R. Carnap was formalized within the framework of possible world semantics, which is considered as an extension of standard model-theoretic semantics devised for predicate logic by A. Tarsky to ascribe each formula a set-theoretic interpretation on a domain of reference (Menzel, 2017). Within the current study we use the four-part Kripke model $M \langle W, R, D, I \rangle$, where W is a non-empty set of possible worlds, R is a binary relation on W binding possible worlds $w \in W$ and $w' \in W$ as alternative realities, D is a non-empty object domain considered as a scope of observed objects bound with a number of n-ary relations and represented in a scope of possible worlds, and I is a function from a set of possible worlds to a set of all n-ary relations on a domain, which assigns each vocabulary unit a referent on D : $I_V: W \rightarrow D^n$ (Lindström, 2001; Fitting, 2015).

As long as a lexeme is considered to acquire a specific logical type in accordance with its reference, a lexeme denoting an n-ary relation on a domain should be represented by a predicate symbol in an atomic formula $P(x_1, \dots, x_n)$, whereas if a lexeme's extension is represented by a member of an n-ary relation on a domain, the lexeme is supposed to act as a constant symbol substituting a variable in an argument position (Trentelman, 2009). In order to describe an intension of a lexeme functionally, i.e. as a referential function of a lexeme that it takes in a set of possible worlds on a non-empty object domain, we have to take into account that a predicate is supposed to be mapped to a set of homogeneous n-ary relations in all possible worlds, whereas a constant symbol should denote the same entity in all possible worlds. On occasion we equate lexical meaning to intension of a lexeme, lexical meaning of a vocabulary unit should be understood as a function from a set of possible worlds to a set of all subsets of homogeneous n-ary relations on a domain: $Int_V: W \rightarrow 2^{D^n}$, this way of lexical meaning comprehension is extensively shared in theoretical papers on ontologies (Guarino, 1998; Guizzardi, 2005; Gritz, 2017).

An intension of a lexeme might be defined conceptually as proposed by R. Carnap: '*the general conditions which an object must fulfill in order to be denoted by (that) word*' (cit. ex. Gasparri and Marconi, 2016). Hence, in order to compile a formal definition of lexical meaning, the indispensable attributes of a lexeme's referent should be described by means of a formal language, which is understood as a set of strings formed by means of a finite set of syntactic rules and a set of symbols that make up an alphabet. In the current study the strings are given interpretation within the framework of model-theoretic semantics by means of the four-part Kripke model $M \langle W, R, D, I \rangle$ described above, therefore the strings should be formed by means of the first-order modal logic. Within the current study we concentrate our attention on intension and extension of predicate symbols, consequently a simplified signature of the first-order modal logic is implemented: functional and constant symbols are excluded from consideration. An extension is ascribed to a predicate symbol in accordance with the following set of rules defining truth values, where P is a predicate symbol and x is a variable, and $P(x_1, \dots, x_n)$ is an atomic formula. Atomic formulas are combined in a string with the connectives: $\neg, \wedge, \vee, \rightarrow$; the quantifiers: \exists, \forall ; and modal operators: \Box, \Diamond used to denote necessity and possibility, m and n are variable assignment functions mapping each variable to an entity on a domain which disagree only on the variable x in some cases:

$$M, w \models_m P(x_1, \dots, x_n) \Leftrightarrow (m(x_1), \dots, m(x_n)) \in I(P, w),$$

$$M, w \models_m \neg \alpha \Leftrightarrow \text{not } M, w \models_m \alpha,$$

$$M, w \models_m \alpha \wedge \beta \Leftrightarrow M, w \models_m \alpha \text{ and } M, w \models_m \beta,$$

$$M, w \models_m \alpha \vee \beta \Leftrightarrow M, w \models_m \alpha \text{ or } M, w \models_m \beta,$$

$$M, w \models_m \alpha \rightarrow \beta \Leftrightarrow \text{not } M, w \models_m \alpha \text{ or } M, w \models_m \beta,$$

- $M, w \models_m \forall x. \alpha \Leftrightarrow M, w \models_n \alpha$ for all x variants n of m ,
- $M, w \models_m \exists x. \alpha \Leftrightarrow M, w \models_n \alpha$ for some x variant n of m ,
- $M, w \models_m \diamond \alpha \Leftrightarrow M, w' \models_m \alpha$ for some w' such that $R(w, w')$,
- $M, w \models_m \Box \alpha \Leftrightarrow M, w' \models_m \alpha$ for all w' such that $R(w, w')$.

One should distinguish the statements $\diamond \alpha$, which are true under particular interpretation only in some of the possible worlds under consideration, and the statements $\Box \alpha$, which are true under particular interpretation in all the possible worlds under consideration. A borderline between these two kinds of statements was drawn by R. Carnap (1947, 1952) and by W. V. O. Quine (1951), who distinguished the statements which exhibit synthetic truth, that depends on a particular state of affairs, and analytic truth, or L-truth in terminology by R. Carnap, that holds in all possible worlds and depends on meanings of words.

For instance, referring to the John Smith's enterprise as to a single possible world we can make the following true statement revealing attributes a referent of the lexeme 'employee' possesses on a domain in a particular state of affairs:

$$\diamond \forall x(\text{Employee}(x) \rightarrow \text{Male}(x) \wedge \text{Engineer}(x)).$$

On the contrary, one can formulate a statement that describes referents of the lexeme 'employee' in the scope of all possible worlds one might imagine to exist in 2018 or in the scope of possible worlds where both people and sensible robots – androids are employed that can be conceived to take shape in a century:

$$\Box \forall x(\text{Employee}(x) \rightarrow \text{Person}(x)),$$

$$\Box \forall x(\text{Employee}(x) \rightarrow \text{Person}(x) \wedge \text{Android}(x)).$$

One could define all individuals who are employees in all existing possible worlds assigning values to variables within the following meaning postulate, which states the meaning of the lexeme 'employee':

$$\Box \forall x \exists y (\text{Employee}(x) \rightarrow \text{Person}(x) \wedge \text{has_Job}(x, y) \wedge \text{is_paid_for}(x, y)).$$

For this reason, within the practice of lexical meaning formal representation the statements of the type $\Box \alpha$ referred to as meaning postulates are proposed to define an intension of a predicate symbol representing a common noun, a verb, or an adjective within the framework of model-theoretic semantics (Carnap, 1952; Montague, 1973). Even though meaning postulates have been subjected to wide criticism (Quine, 1951; Katz, 1982), the point of implementing meaning postulates is still advocated (Horsey, 2000; Wechsler, 2015). One of the reasons for implementing meaning postulates is that they represent semantics ignoring cognitively problematic hierarchy of concepts arranged by compositional complexity and associated with individual words, which is proposed by decompositional approach to lexical semantics (Chierchia and McConnell-Ginet, 2000). We suppose that meaning postulates are able to serve as a basis for formal definitions development provided that the basic problems of lexical meaning formal representation have been solved:

- a correlation between a data unit retrieved by virtue of NL text analysis and a type of a formal language symbol should be found,
- the data retrieved by virtue of NL text analysis should be expressed by means of the strings formed in accordance with syntactic rules of a formal language applied for the purpose of formalization.

3. Formal representation of lexical meaning within an ontology

An ontology is a knowledge base which provides a formal specification of a vocabulary that represents a scope of entities of a domain together with entity bounding n-ary relations defined on the domain which is observed in a set of possible worlds (Gruber, 1993; Guarino, 1998). The knowledge base is comprised of assertional axioms, which describe individuals; terminological and relational axioms, which describe features of classes and object properties accordingly, all axioms being written by means of a particular formal language. Hence, within an ontology structure three basic types of units should be distinguished: individuals (also referred to as instances), which represent single entities of a domain;

classes (also referred to as concepts), which represent subsets of entities of a domain and are instantiated by members of the subsets; object properties¹ (also referred to as roles), which denote binary relations that bind single entities of a domain (Krötzsch et al., 2014).

Assertional, terminological and relational axioms, which compose an ontology as a knowledge base, are currently written by means of a standard formal language known as OWL 2 DL, which exploits the expressive power of the Description Logic titled as SROIQ (Lehmann, 2010). The signature of SROIQ contains non-logical symbols forming a triple $\langle O, \mathcal{C}, \mathcal{Q} \rangle$. Each vocabulary unit is assigned an interpretation function I , which maps each individual $a \in O$ to an entity on a domain $a^I \in \Delta^I$, which maps each concept $C \in \mathcal{C}$ to a subset of a domain $C^I \subseteq \Delta^I$, which maps each role $Q \in \mathcal{Q}$ to a binary relation on a domain $Q^I \subseteq \Delta^I \times \Delta^I$ (Leinberger et al., 2016). The signature includes a top-concept \top , which can be instantiated by every individual; bottom concept \perp , which denotes an empty set; nominals $\{a\}$, which denote sets with a single member; and a universal relation U , which is associated with a universal binary relation on a domain $\Delta^I \times \Delta^I$ (Horrocks et al., 2006; Krötzsch et al., 2014). Assertional axioms define features of individuals by relating them to each other by means of individual equality (inequality) assertions: $a \equiv b$ ($a \neq b$), to concepts by means of concept assertions: $a:C$, to roles by means of role assertions: $(a, b):Q$, $(a, b):\neg Q$. Terminological axioms describe concept inclusion: $C \sqsubseteq D$, whenever C is subsumed by D , and concept equivalence: $C \equiv D$, whenever C and D share the same instances (Krötzsch et al., 2014). Relational axioms are not considered in this paper. Terminological axioms of both types might be expanded to form complex class descriptions by means of specific concept constructors. The concept constructors allowed by SROIQ, and consequently by OWL 2 DL, to define ontology class features within terminological axioms are given in the Table 1 (Horrocks et al., 2006; Hitzler et al., 2012; Motik et al., 2012).

Concept Constructor	OWL 2 DL Syntax	SROIQ Syntax	Semantics
Complement	<code><owl:complementOf rdf:resource="#C"/></code>	$\neg C$	$\Delta^I \setminus C^I$
Intersection	<code><owl:intersectionOf rdf:parseType="Collection"> <owl:Class rdf:resource="#C"/> <owl:Class rdf:resource="#D"/> </owl:intersectionOf></code>	$C \sqcap D$	$C^I \cap D^I$
Union	<code><owl:unionOf rdf:parseType="Collection"> <owl:Class rdf:resource="#C"/> <owl:Class rdf:resource="#D"/> </owl:unionOf></code>	$C \sqcup D$	$C^I \cup D^I$
Universal restriction	<code><owl:Restriction> <owl:onProperty rdf:resource="#Q"/> <owl:allValuesFrom rdf:resource="#C"/> </owl:Restriction></code>	$\forall Q. C$	$\left\{ x \in \Delta^I \mid \forall y. (x, y) \in Q^I \rightarrow y \in C^I \right\}$
Existential restriction	<code><owl:Restriction> <owl:onProperty rdf:resource="#Q"/> <owl:someValuesFrom rdf:resource="#C"/> </owl:Restriction></code>	$\exists Q. C$	$\left\{ x \in \Delta^I \mid \exists y. (x, y) \in Q^I \wedge y \in C^I \right\}$
Qualified cardinality restriction (at-least restriction)	<code><owl:Restriction> <owl:onProperty rdf:resource="#Q"/> <owl:onClass rdf:resource="#C"/> <owl:minQualifiedCardinality rdf:datatype=</code>	$\geq nQ. C$	$\left\{ x \in \Delta^I \mid \left \left\{ y \mid (x, y) \in Q^I \wedge y \in C^I \right\} \right \geq n \right\}$

¹ Datatype properties, which assign an individual a data value, are not considered in this paper.

	"&xsd;nonNegativeInteger">n </owl:minQualifiedCardinality> </owl:Restriction>		
Qualified cardinality restriction (at-most restriction)	<owl:Restriction> <owl:onProperty rdf:resource="#Q"/> <owl:onClass rdf:resource="#C"/> <owl:maxQualifiedCardinality rdf:datatype="&xsd;nonNegativeInteger">n </owl:maxQualifiedCardinality> </owl:Restriction>	$\leq nQ.C$	$\left\{ x \in \Delta^I \mid \left \left\{ y \mid \begin{array}{l} (x,y) \in Q^I \\ \wedge y \in C^I \end{array} \right\} \right \leq n \right\}$
Local reflexivity	<owl:Restriction> <owl:onProperty rdf:resource="#Q"/> <owl:hasSelf rdf:datatype="&xsd;boolean">true </owl:hasSelf> </owl:Restriction>	$\exists Q.Self$	$\{x \in \Delta^I \mid (x, x) \in Q^I\}$
Enumeration	<owl:oneOf rdf:parseType="Collection"> <rdf:Description rdf:resource="#o ₁ " /> ... <rdf:Description rdf:resource="#o _n " /> </owl:oneOf>	$\{o_1\} \sqcup, \dots, \sqcup \{o_n\}$	$\{o_1^I, \dots, o_n^I\} \subseteq \Delta^I$

Table 1: Concept constructors used to enable class defining axioms formation by means of the syntax of OWL 2 DL and the syntax of SROIQ

As long as OWL 2 DL is considered to be a fragment of the first-order logic, the semantics of OWL 2 DL statements is regarded within the framework of the four-part Kripke model $M \langle W, R, D, I \rangle$ discussed above. It should be considered that whereas assertional axioms are supposed to state something true about relations on a domain at least in a single possible world, i.e. to state synthetic truth, terminological axioms are presumed to state something true about relations on a domain in every possible world under consideration, i.e. to state analytic truth. Therefore we put forward a hypothesis that terminological axioms defining ontology class features form meaning postulates referred to as class descriptions that reveal lexical meanings of the terms representing ontology classes. Since we define an ontology class as a formal concept in the following way: ‘ $\langle A, B \rangle$ is a formal concept if and only if A contains just objects sharing all attributes from B and B contains just attributes shared by all objects from A ’ (Belohlávek, 2008: 7), class descriptions should include all the attributes from B , i.e. all features of the class, which distinguish it from any other class in the ontology.

Class descriptions are derived in the process of ontology learning. Ontology learning techniques allow to retrieve ontology units from structured data (databases), semi-structured data (HTML or XML documents, Wordnet), and unstructured data (unannotated text documents) (Biemann, 2005). Specific ontology learning techniques are implemented to define classes and object properties in an unannotated natural language text: term and synonym extraction techniques, conceptual clustering (Cimiano et al., 2009), association discovery algorithms (Mädche and Staab, 2000), dependency relation analysis (Schutz and Buitelaar, 2005), and noun phrase analysis (Hearst, 1992).

4. An outline of the NL-DL definition transformation method

In order to generate ontology class descriptions systematically, LExO approach has been proposed to derive terminological axioms through transformation of syntactically parsed natural language definitions of class representing lexemes into OWL corresponding DL-statements (Völker et al., 2007; Völker et al., 2008). The method underlying LExO as well as later developments such as ACE and TEDEI (Mathews and Kumar, 2017) includes three main steps of natural language material processing:

syntactic parsing, omission/concatenation and transformation (Azevedo et al., 2014). LExO has been devised to transform natural language sentences into SHOIN axioms, which are transferrable into OWL 1 DL (Völker et al., 2008), while TEDEI has introduced the rules for OWL 2 DL axioms formation (Mathews and Kumar, 2017). Since OWL 2 syntax has been developed on the basis of OWL 1 syntax, the set of transformation rules proposed by LExO has been enhanced by TEDEI recommendations developed specifically for OWL 2 DL (see Table 2), omission and concatenation rules proposed by ACE are also taken into account.

For the purpose of natural language text syntactic analysis an off-the-shelf parser is used. For instance, Völker et al. (2007) have proposed to use the MINIPAR parser, following Azevedo et al. (2014) we use the Stanford parser, the version accessible online². It is common practice to associate all noun phrases containing common nouns with ontology classes, whereas verb phrases are supposed to introduce object properties. Following LExO approach we omit subjective adjectives attached by common nouns, intersective adjectives are considered to represent ontology classes (see rule 4), whereas privative adjectives are supposed to indicate complement (see rule 5). Following LExO and ACE the prepositions attaching noun phrases as prepositional complements are subjected to concatenation (see rules 11, 12), determiners should be omitted, nominal collocations used as terms, adverbs, and modal verbs undergo concatenation (Völker et al., 2008; Mathews and Kumar, 2017). We refer to TEDEI to obtain lexical and grammatical indicators of specific OWL 2 DL constructors: qualified cardinality restriction (see rule 13); universal and existential restriction and local reflexivity (see rules 8, 11, 12) (Mathews and Kumar, 2017). The set of transformation rules provides a roadmap for mapping phrase structures to DL constructs and a scheme for development of unfold terminological axioms according to the principle of compositionality. The terminological axioms recognized as DL-definitions state concept equivalence between an atomic concept and a complex description which defines a set of referents denoted by the atomic concept. Therefore in order to provide an account on lexical meaning of a term, a DL-definition should be coherent in terms of set theory based interpretation on a domain of reference.

	Transformation rule	NL syntax	DL syntax
(1)	Copula	NP_0 is/are NP_1	$NP_0 \equiv NP_1$
(2)	Conjunction	NP_0 and NP_1	$NP_0 \sqcap NP_1$
(3)	Disjunction	NP_0 or NP_1	$NP_0 \sqcup NP_1$
(4)	Intersective adjective	$Adj_0 NP_0$	$Adj_0 \sqcap NP_0$
(5)	Privative adjective	$Adj_0 NP_0$	$\neg NP_0$
(6)	Negation (not)	not $V_0 NP_0$	$\neg \exists V_0. NP_0$
(7)	Negation (without)	NP_0 without $NP(\text{pcomp-n})_1$	$NP_0 \sqcap \neg \exists \text{with}. NP_1$
(8)	Transitive verb phrase	$V_0 NP(\text{obj})_0$	$\exists V_0. NP_0 / \forall V_0. NP_0 / \exists V_0. \text{Self}$
(9)	Relative clause	$NP_0 C(\text{rel}) VP_0$	$NP_0 \sqcap VP_0$
(10)	Participle	$NP_0 VP(\text{vrel})_0$	$NP_0 \sqcap VP_0$
(11)	Verb with prepositional complement	$V_0 \text{Prep}_0 NP(\text{pcomp-n})_0$	$\exists V_0_Prep_0. NP_0 / \forall V_0_Prep_0. NP_0$
(12)	Noun with prepositional complement	$NP_0 \text{Prep}_0 NP(\text{pcomp-n})_1$	$NP_0 \sqcap \exists NP_0_Prep_0. NP_1 / NP_0 \sqcap \forall NP_0_Prep_0. NP_1$
(13)	Number restriction	$V_0 \text{Num} NP(\text{obj})_0$	$\geq nV_0. NP_0 / \leq nV_0. NP_0$

Table 2: The basic transformation rules used for the purpose of DL-definitions formation

5. Problematic cases of the NL-DL definition transformation method application

The data for the NL-DL definition transformation method critical study has been derived from a Dictionary of Linguistics and Phonetics (Crystal, 2008) and Routledge Dictionary of Language and Linguistics (Bussmann, 1996). Pairs of alternative natural language definitions of 50 randomly selected sociolinguistic terms retrieved from the two dictionaries were parsed and transformed into DL-definitions of ontology class representing terms using the described method of transformation. The left

² <http://nlp.stanford.edu:8080/parser/index.jsp>

part of a DL-definition includes an ontology class bearing a role of definiendum, whereas the right part of the definition contains syntactically bound definiens – distinctive ontology classes and object properties. Whenever ontology classes act as definiens of the boundaries of a subset denoted by an ontology class being described, the classes undergo intersection, union, and complement. Whenever an object property acts as definiens of the boundaries of a subset denoted by an ontology class being described, the object property is ascribed a range with a universal, existential, or cardinality restriction imposed on it. Whether ontology classes or object properties act as definiens, certain problems arise in the process of transformation mainly due to the fact that natural language syntax is undoubtedly far more expressive than the syntax of SROIQ, and natural language syntax peculiarities need to be taken into account.

One problem we have discovered is the formation of intersections between disjoint ontology classes, which arises as a result of relative clause transformation being conducted in case a relative pronoun, which is coreferential with the modified noun, acts not as a nominal subject, but as a nominal modifier in relation to a predicate of the relative clause. For instance, within the NL-definition of the term ‘archistratum’ a relative pronoun, which is coreferential with the modified noun ‘variety’, acts as a nominal modifier of the relative clause’s predicate, whereas the noun ‘community’ takes the role of a nominal subject. An attempt to obtain a formal class description from this NL syntactic material results in an inadmissible DL-axiom describing the intersection of disjoint sets of privileged varieties of a language and communities (see Table 4). However, this is not the case in a DL-definition of the term ‘divergence’ since the modified nominal collocation ‘process of dialect change’ denotes a set of events, whereas the relative clause ‘in which the dialects become less like each other’ describes a subset of the set of events that could be defined as a process of dialect change. Consequently, if a noun or a nominal collocation modified by a relative clause designates a set of objects, not a set of events, a relative pronoun should act as a nominal subject in relation to a predicate of the relative clause for the NL-definition to be transferrable into DL-definition by virtue of the NL-DL transformation method. A good example of a suitable NL-definition is a definition of the term ‘standard’: ‘*Standard is a prestige variety of language used within a speech community, which cuts across regional differences and provides unified means of communication and an institutionalized norm, which can be used in the mass media and teaching*’, where in both relative clauses the relative pronouns are found in the syntactic role of a nominal subject.

Structurally the right part of a DL-definition might be represented as a chain of classes undergoing intersection. Each class obtains a description, which involves other classes and object properties, some of the classes’ names occur twice or more times in different links of the chain, the term ‘language’ in DL-definitions of terms ‘non-native variety’, ‘interference’, and ‘adstratum’ is among them. The issue of coreference between the recurred names should be resolved, otherwise nonsensical DL-statements emerge, one of them advocates that non-native varieties emerge in societies where speakers do not have a mother tongue at all (see Table 4). Since determiners are subjected to omission during the NL-DL transformation, in order obtain a correct DL-definition of a term, one should expand the DL-definition with information on whether the same or different subsets of a named set are bound in different links of the chain. In order to improve otherwise improper DL-definitions, one should annotate recurred names or use a local reflexivity constructor to characterize an object property as reflexive or irreflexive in case the same class name is used to characterize both domain and range of the object property (see Table 3).

DL-definitions with annotation of recurred names	DL-definitions with a local reflexivity constructor introduced
<p><i>Adstratum</i> $\equiv \text{Scope} \sqcap \exists \text{scope_of. (Feature)}$ $\sqcap \exists \text{features_in. Language}_0$ $\sqcap \exists \text{have_resulted_from. (Contact)}$ $\sqcap \exists \text{contact_with. (Neighbouring Language}_0))$</p>	<p><i>Divergence</i> $\equiv \text{Process} \sqcap \exists \text{process_of. (Dialect Change)} \sqcap \text{(Dialect)}$ $\sqcap \exists \text{become_less_like. Dialect}$ $\sqcap \neg \exists \text{become_less_like. Self}$</p>

<p><i>Interference</i> $\equiv Scope \sqcap \exists scope_of. (Error \sqcap (Speaker \sqcap \exists introduces_into. \mathbf{Language}_0 \sqcap \exists introduces_as_a_result_of. (Contact \sqcap \exists contact_with. \mathbf{Language}_1)))$)</p>	<p><i>Isolect</i> $\equiv Linguistic \sqcap (Variety \sqcap \exists differs_minimally_from. Variety \sqcap \neg \exists differs_minimally_from. Self)$)</p>
<p><i>Language shift</i> $\equiv (Gradual \sqcup Sudden) \sqcap (Move \sqcap \exists move_from. (Use \sqcap \exists use_of. \mathbf{Language}_0) \sqcap \exists move_to. (Use \sqcap \exists use_of. \mathbf{Language}_1) \sqcap \exists move_by. (Individual \sqcup Group))$)</p>	<p><i>Convergence</i> $\equiv Process \sqcap \exists process_of. (Dialect \sqcap Change) \sqcap (Dialect \sqcap \exists become_more_like. Dialect \sqcap \neg \exists become_more_like. Self)$)</p>

Table 3: The improvements proposed to enhance otherwise improper DL-definitions

Another problem arises as soon as we adopt the rule which states the conversion of a privative adjective ‘former’ into complement since the adjective denotes the absence of an attribute in the present and at the same time states its presence in the past. For this reason, the DL-definition of the term ‘creole’ fails to express an essential attribute of creole languages – the fact that all creole languages evolve from pidgins. Even the possibility to define adjectives as an intersective on a reasonable basis does not make things easier since many of them have wide meanings, the adjectives ‘cultured’ and ‘intellectual’ derived from the definition of the term ‘archistratum’ are good examples. As a result, the inclusion of the ontology classes represented by intersective adjectives in a taxonomy might be a challenging task.

The method also ignores the cases of an adjective acquiring a function of a predicate and representing an object property. A definition of the term ‘variable’: ‘Variables are the units in a language which are most subject to social or stylistic variation, and thus most susceptible to change in the long term’, clearly illustrates the use of two adjectives ‘subject’ and ‘susceptible’ in the role of a predicate. Both adjectives attach noun phrases as prepositional complements, yet the proposed list of transformation rules is limited to solutions for formal representation of noun phrases and verb phrases with prepositional complements.

NL-definitions	DL-definitions
Adstratum is a scope of features in a language which have resulted from contact with a neighbouring language.	<i>Adstratum</i> $\equiv Scope \sqcap \exists scope_of. (Feature \sqcap \exists features_in. Language \sqcap \exists have_resulted_from. (Contact \sqcap \exists contact_with. (Neighbouring \sqcap Language)))$)
Archistratum is a privileged variety of a language from which a community draws its cultured or intellectual vocabulary.	<i>Archistratum</i> $\equiv (Privileged \sqcap Variety \sqcap \exists variety_of. Language) \sqcap (Community \sqcap \exists draws. ((Cultured \sqcup Intellectual) \sqcap Vocabulary) \sqcap \exists draws_from. Language)$)
Change from below is the scope of the alterations that people make in their speech below the level of their conscious awareness.	<i>Change from below</i> $\equiv Scope \sqcap \exists scope_of. (Alteration \sqcap (People \sqcap \exists make_in. Speech \sqcap \exists make_below. (Level \sqcap \exists level_of. (Conscious \sqcap Awareness))))$)
Convergence is a process of dialect change in which the dialects become more like each other.	<i>Convergence</i> $\equiv Process \sqcap \exists process_of. (Dialect \sqcap Change) \sqcap (Dialect \sqcap \exists become_more_like. Dialect)$)
Creole is a former pidgin whose functional and grammatical limitations and simplification have been eliminated and which now functions as a full-fledged, standardized native language.	<i>Creole</i> $\equiv \neg Pidgin \sqcap \forall have_been_eliminated. (((Functional \sqcap Grammatical) \sqcap Limitation) \sqcap Simplification) \sqcap \exists now_functions_as. ((Full - fledged \sqcap Standardized) \sqcap Native_language)$)

Divergence is a process of dialect change in which the dialects become less like each other.	<i>Divergence</i> $\equiv Process \sqcap \exists process_of. (Dialect \sqcap Change)$ $\sqcap (Dialect \sqcap \exists become_less_like. Dialect)$
Interference is a scope of the errors a speaker introduces into one language as a result of contact with another language	<i>Interference</i> $\equiv Scope \sqcap \exists scope_of. (Error \sqcap (Speaker$ $\sqcap \exists introduces_into. Language$ $\sqcap \exists introduces_as_a_result_of. (Contact$ $\sqcap \exists contact_with. Language)))$
Isolect is a linguistic variety which differs minimally from another variety.	<i>Isolect</i> $\equiv Linguistic \sqcap (Variety$ $\sqcap \exists differs_minimally_from. Variety)$
Language shift is the gradual or sudden move from the use of one language to another, either by an individual or by a group.	<i>Language_shift</i> $\equiv (Gradual \sqcup Sudden) \sqcap (Move$ $\sqcap \exists move_from. (Use \sqcap \exists use_of. Language)$ $\sqcap \exists move_to. (Use \sqcap \exists use_of. Language)$ $\sqcap \exists move_by. (Individual \sqcup Group))$
Network is the set of linguistic interactions that a speaker has with others.	<i>Network</i> $\equiv Set \sqcap \exists set_of. ((Linguistic \sqcap Interaction)$ $\sqcap Speaker \sqcap \exists has_with. Speaker)$
Non-native variety is a variety of a language which has emerged in a speech community in which most speakers do not have the language as a mother tongue.	<i>Non – native_variety</i> $\equiv (Variety \sqcap \exists variety_of. Language)$ $\sqcap \exists has_emerged_in. (Speech_community$ $\sqcap (Speaker \sqcap \neg \exists have. (Language$ $\sqcap \exists language_as. Mother_tongue)))$

Table 4: The samples of unacceptable definitions obtained by means of the NL-DL definition transformation method

6. Conclusion

The brief critical study of the NL-DL definition transformation method has revealed the fact that the method proposes a reasonable solution to the basic problems of lexical meaning formal representation that have been outlined in the current article. First of all, the correlation between syntactic categories and types of OWL 2 symbols and conforming DL symbols has been set. Secondly, the transformation rules propose a working algorithm which allows to transfer the natural language phrase combinations revealed via parsing techniques into DL-axioms which compose ontology class descriptions. Hence, the DL-NL definition transformation method should be considered as a theoretically appropriate solution to the problem of lexical meaning formal representation within the framework of an ontology.

On the other hand, the method obviously does not involve the formal analysis of syntactic dependencies connecting phrase units, which leads to the problem of semantic correlation between the definiendum and definiens in a DL-definition. The extensive practice of omission of determiners resulting in the unresolved issue of coreference between recurred names poses additional challenges towards formation of DL-definitions that could cover the whole scope of class representing term's referents on a domain.

References

- Azevedo, R., Freitas, F., Rocha, R., Menezes, J., Rodrigues, C., and Gomes, M. (2014). Representing Knowledge in DL ALC from Text. *Procedia Computer Science*, (35):176–185.
- Belohlávek, R. (2008). *Introduction to Formal Concept Analysis*. Olomouc, Palacky University.
- Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5): 35–43.

- Biemann, C. (2005). *Ontology Learning from Text: a Survey of Methods*. LDV-Forum 2005, 20(2):75–93.
- Bussmann, H. (1996). *Routledge Dictionary of Language and Linguistics*. Translated from German and edited by Trauth, G. and Kazzazi, K. London, New York: Routledge.
- Carnap, R. (1947). *Meaning and Necessity. A Study in Semantics and Modal Logic*. USA, IL, Chicago: University of Chicago Press.
- Carnap, R. (1952). Meaning Postulates. *Philosophical Studies*, 3(5):65–73.
- Chierchia, G. and McConnell-Ginet, S. (2000). *Meaning and Grammar: An Introduction to Semantics*, 2nd ed. USA, MA, Cambridge: MIT Press.
- Cimiano, P., Mädche, A., Staab, S., and Völker, J. (2009). Ontology Learning. In Staab, S. and Studer R., Eds., *Handbook on Ontologies. International Handbooks on Information Systems*. Berlin, Heidelberg: Springer, 245–267.
- Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics*. 6th ed. Oxford: Blackwell.
- Ding, Y. (2010). Semantic Web: Who is Who in the Field – A Bibliometric Analysis. *Journal of Information Science*, 36(3):335–356.
- Farrugia, J. (2003). *Model-Theoretic Semantics for the Web*. <http://www.w3.org/2003.org/cdrom/papers/refereed/p277/p277-farrugia.html>.
- Fitting, M. (2015). Intensional Logic. In Zalta E. N., Ed., *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/sum2015/entries/logic-intensional/>.
- Gasparri, L. and Marconi, D. (2016). Word Meaning. In Zalta E. N., Ed., *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/spr2016/entries/word-meaning/>.
- Gritz, M. (2017). An Ontology as a Medium of Lexical Meaning Formal Representation. *Foreign Languages in Tertiary Education*, 2(41):60–71.
- Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2):199–220.
- Guarino, N. (1998). Formal Ontology and Information Systems. In *Formal Ontology in Information Systems. Proceedings of the FOIS'98*. Trento, Italy, June 6 – 8, 1998. Amsterdam: IOS Press, 3–15.
- Guizzardi, G. (2005). *Ontological Foundations for Structural Conceptual Models*. PhD thesis. Centre for Telematics and Information Technology, University of Twente. The Netherlands, Enschede.
- Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING 1992)*. Nantes, France. Vol. 2, 539–545.
- Hitzler, P., Krötzsch, M., and Rudolph S. (2009). *Knowledge Representation for the Semantic Web. Part 1: OWL 2. Knowledge Representation and Reasoning for the Semantic Web – OWL 2 Rules*. A tutorial at KI 2009, Paderborn, Germany, September 15, 2009. <http://semantic-web-book.org/w/images/b/b0/KI09-OWL-Rules-1.pdf>.
- Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., and Rudolph, S., Eds. (2012). *OWL 2 Web Ontology Language Primer*. W3C Recommendation. 2nd ed. <https://www.w3.org/TR/owl2-primer/>.
- Horrocks, I. (2008). Ontologies and the Semantic Web. *Communications of the ACM*, 51(12):58–67.
- Horrocks, I., Kutz, O., and Sattler, U. (2006). The Even More Irresistible SROIQ. In *Proceedings of the 10th International Conference on Principles of Knowledge Representation and Reasoning (KR 2006)*. Lake District, UK, June 2 – 5, 2006. AAAI Press, 57–67.
- Horrocks, I., Patel-Schneider, P. F., McGuinness, D. L., and Welty, C. A. (2007). OWL: a Description Logic Based Ontology Language for the Semantic Web. In Calvanese, D., McGuinness, D. L.,

- Nardi, D., Patel-Schneider, P. F., Eds., *The Description Logic Handbook: Theory, Implementation, and Applications*. 2nd ed. Cambridge: Cambridge University Press, ch. 14, 458–487.
- Horsey, R. (2000). Meaning Postulates and Deference. *UCL Working Papers in Linguistics*, 12:45–64.
- Katz, J. J. (1982). Common Sense in Semantics. *Notre Dame Journal of Formal Logic*, 23(2):174–218.
- Krötzsch, M., Simancík, F., and Horrocks, I. (2014). *A Description Logic Primer*. In Lehmann, J. and Völker, J., Eds., *Perspectives on Ontology Learning. Studies on the Semantic Web. Vol. 18*. Amsterdam: IOS Press, ch. 1, 3–19.
- Lehmann, J. (2010). *Learning OWL Class Expressions*. PhD thesis in Computer Science. University of Leipzig.
- Leinberger, M., Lämmel, R., and Staab, S. (2016). *λDL: Syntax and Semantics*. Preliminary Report. University of Koblenz-Landau.
- Lindström, S. (2001). Quine’s Interpretation Problem and the Early Development of Possible Worlds Semantics. In Carlson, E. and Sliwinski, R., Eds., *Omnium-gatherum. Philosophical Essays Dedicated to Jan Österberg on the Occasion of his Sixtieth Birthday*. Uppsala philosophical studies 50. Uppsala: Uppsala University, Department of Philosophy, 187–213.
- Mädche, A. and Staab, S. (2000). Discovering Conceptual Relations from Text. In *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI)*, 321–325.
- Mathews, K. and Kumar, P. (2017). Extracting Ontological Knowledge from Textual Descriptions through Grammar-based Transformation. In *Proceedings of the Knowledge Capture Conference (K-CAP 2017)*. Austin, TX, USA, December 4 – 6, 2017. USA, New York: ACM, article 21.
- Menzel, C. (2017). Possible Worlds. In Zalta E. N., Ed., *The Stanford Encyclopedia of Philosophy* <https://plato.stanford.edu/archives/win2017/entries/possible-worlds/>.
- Montague, R. (1973). The Proper Treatment of Quantification in Ordinary English. In Kulas J., Fetzer, J. H., and Rankin, T. L., Eds., *Philosophy, Language, and Artificial Intelligence. Studies in Cognitive Systems*. The Netherlands, Dordrecht: Springer, Vol. 2:141 – 162.
- Motik, B., Patel-Schneider, P. F., and Grau, B. C., Eds. (2012). *OWL 2 Web Ontology Language Direct Semantics*. W3C Recommendation. 2nd ed. <https://www.w3.org/TR/owl2-direct-semantics/>.
- Quine, W.V.O. (1951). Two Dogmas of Empiricism, *Philosophical Review*, 60(1):20–43.
- Schutz, A. and Buitelaar, P. (2005). RelExt: A Tool for Relation Extraction from Text in Ontology Extension. In *Proceedings of the 4th International Semantic Web Conference (ISWC 2005)*. Galway, Ireland, November 6 – 10, 2005. 593–606.
- Trentelman, K. (2009). *Survey of Knowledge Representation and Reasoning Systems*. Australia, Edinburgh: Defence Science and Technology Organisation (DSTO).
- Völker, J., Haase P., and Hitzler, P. (2008). Learning Expressive Ontologies. In Buitelaar, P. and Cimiano, P., Eds., *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*. The Netherlands, Amsterdam: IOS Press, 45–69.
- Völker, J., Hitzler, P., and Cimiano, P. (2007). Acquisition of OWL DL Axioms from Lexical Resources. In *Proceedings of the 4th European Semantic Web Conference (ESWC’07)*. Innsbruck, Austria, June 3 – 7, 2007. Berlin, Heidelberg: Springer, 670–685.
- Wechsler, S. (2015). *Word Meaning and Syntax: Approaches to the Interface*. 1st ed. Oxford: Oxford University Press.
- Yu, L. (2014). *A Developer’s Guide to the Semantic Web*. 2nd ed. Berlin, Heidelberg: Springer.

Narrow Productivity, Competition, and Blocking in Word Formation

Junya Morita

Kinjo Gakuin University

College of Humanities

morita@kinjo-u.ac.jp

Abstract

The present study explores the productivity of word formation processes in English, focusing on word composition by suffixes such as *-ize* (e.g. *transcendentalize*), *-(a)(t)ion* (*territorization*), and *-al* (*realizational*). An optimal productivity measure for affixation is identified, which makes best use of hapax legomena in a large-scale corpus and attaches great importance to the base forms of an affix. This measure is then applied to the data collected from a large corpus to compute the productivity values of twelve kinds of affixes. The detailed investigation reveals that (i) the high productivity rate of an affix demonstrates a creative aspect of the affix, giving full support to the idea of “generative” morphology, (ii) productivity is gradient; very high, fairly high, and low productivity of affixes are recognizable, and (iii) this is necessarily reflected in determining the word form of a derivative (cf. *territorization*); competition is carried out to decide which affix is selected for a given base form (*territorize*) and the “losers” (*-ment/-al*) are blocked out.

1. Introduction

Productivity – the potentiality of creating new lexemes – is one of the central themes of morphological studies and attempts have been made to establish a refined and accurate productivity measure (Baayen and Renouf, 1996; Plag, 1999). The productive devices of word formation create new items whenever the necessity arises, and the regular selection of a productive affix from the competing affixes enables us to describe the lexicon elegantly. The aim of this study is to measure the productivity of word formation patterns based on a large parsed corpus of natural language and show its theoretical implications. The outline of this study is as follows: after pointing out some previous approaches to productivity measurement and their problems, we propose a new measure in §2. Section 3 applies this measure to the data obtained from a large-scale corpus to calculate the productivity of twelve kinds of suffixations. Section 4 concentrates on the results of the research and their theoretical implications.

2. Productivity Measure

2.1. Previous Studies and Their Problems

Speakers always have the capacity to make up new words and word formation devices are crucially engaged in generating new forms (Aronoff, 1976). It is important to recognize what sort of devices speakers have at their command and to what extent each device can produce new items. It is thus necessary to identify the productivity of each device; the extent to which a word formation device can give rise to new words (Lieber, 2010: 59).

There have been three major approaches to quantifying productivity. The first way is to simply count the total number of the relevant complex words listed in a large dictionary; *The Oxford English Dictionary* contains more *in-* negatives than *non-* negatives, and so the former prefixation is judged to be more productive than the latter. This approach has the problem of neglecting the fact that transparent derivatives are unlikely to be listed in dictionaries (Plag, 1999: 98). The second way of quantifying productivity is to count the number of the relevant complex words which are used in a given material. This way also has the drawback of taking no notice of the fact that there exists an affix which no longer yields new words although the related words are still used; the nominal suffix *-th* has

Keywords: affixation, productivity, hapax, generative morphology

no capability of making new words even though a set of nouns in *-th* still exist (Plag, 1999: 22-23).

The last productivity measure attaches great importance to hapax legomena – token frequency 1 – of a large-scale corpus (Baayen and Renouf, 1996; Plag, 1999). This is based on the view that the capacity of an affix to create new forms crucially involves the degree to which the affix produces words of very low frequency (Hay, 2003). Baayen and Renouf (1996: 73) propose a productivity measure: *Productivity* (P)= n_i/N , where n_i is the number of hapaxes and N is the total number of tokens. For instance, as seen in Table 1, when the token number of *-ize* is 20865 and the hapax number of *-ize* is 80, the productivity value of *-ize* is 0.0038.

suffix	hapaxes (n_i)	tokens (N)	types (V)	productivity (P)
-ate	69	41561	481	0.0017
-ify	18	7236	88	0.0025
-ize	80	20865	347	0.0038

Table 1: The productivity of *-ate*, *-ify*, and *-ize* in the Cobuild corpus. (cf. Plag, 1999: 111)

This corpus-based model also has weaknesses: (i) the existence of extremely highly frequent derivatives (e.g. *realize*, 5506 tokens in the British National Corpus (BNC)) excessively lowers the productivity value of an affix, preventing us from finding the real value of productivity, and (ii) to calculate the productivity value of the whole affixation process would be of little significance to morphological theory, as will be discussed shortly.

2.2. A Proposal

We basically accept the corpus-based productivity measure proposed by Baayen and Renouf, 1996. In adopting it, however, we will revise it in two respects. First, for the productivity formula given above, the total number of types (but not tokens) is placed in the denominator. This is derived from the view that the productivity of a particular process is reflected in the type frequency of the process (Goldberg, 1995: 134-139). Given the observation of *-ize* suffixation in Table 1, the revised productivity formula “ $P=n_i/V$ (V : types)” is applied, giving a productivity value of 0.2305. In this new measure, the productivity of *-ize* is defined as the potentiality of creating 80 kinds of new words when 347 kinds of *-ize* derivatives are used. This is in sharp contrast to Baayen and Renouf’s framework, where the productivity of *-ize* is defined as the potentiality of coining 80 kinds of new words when *-ize* words are used 20865 times.¹

The second point is that productivity should be measured in a small and specific domain. An affix has a tendency to combine with certain affixes (Jespersen, 1949: 449). Based on this fact, Aronoff (1976: 36) claims that “... there is no absolute way to say that one word formation rule is more productive than another. Rather, one must take into account the morphology of the base.” Valuing this base-motivated derivation, we propose the “narrow productivity” of an affix: the rate of its deriving new words for a given base form. For example, of a total of 106 types in $[[X-al]-ize]$ collected, 42 *-ize* hapaxes are detected in BNC, giving a productivity value of 0.396 for the *X-al* base. In the next section, we will morphologically classify the bases of each affixation and calculate the productivity value of affixation for each base.

3. A Research

3.1. Target and Methodology

We calculate the productivity values of (i) five nominal suffixes (*-(a)(t)ion/-al/-ment/ -ity/-ness*), (ii) four adjectival suffixes (*-al/-less/-ic/-ical*), and (iii) three verbal suffixes (*-ize/-ify/-ate*) for the total number of twenty kinds of base forms (e.g. *X-ize/X-ify*). In order to collect the hapaxes ending in these suffixes, we look for them in BNC, a 100-million-word corpus. By repeatedly using the “wild card” function of a research engine, the frequency of complex words ending in the above suffixes is checked

¹ Baayen and Lieber (1991: 810-811) refer to several possible methods to gauge productivity, one of which is essentially equivalent to our productivity measure.

to find the ones which occur only once in the corpus.² As for ascertaining the total number of types of the derivatives concerned, we make a list of those which are included in Lehnert's *Reverse Dictionary of Present-Day English* and attested in BNC.³

3.2. Result

The results of the research are put into tabular forms in terms of (i) noun-forming suffixes, (ii) adjective-forming suffixes, (iii) verb-forming suffixes, and (iv) a suffix which is sensitive to a particular prefix.

3.2.1. Noun-forming Suffixes

Table 2 shows the total number of types, the total number of hapaxes, and the productivity values concerning the deverbal nominal suffixes *-(a)(t)ion*, *-ment*, and *-al* for the bases of *X-ize*, *X-ify*, and *X-ate*. The number of types and that of hapaxes of *[[X-ize]-(a)(t)ion]* nouns are 252 and 47 respectively, giving its productivity value of 0.187. By contrast, since *[[X-ize]-ment]* or *[[X-ize]-al]* nouns are not found in BNC, the suffixes *-ment* and *-al* take a productivity value 0 for these base forms.

base	1. -(a)(t)ion			2. -ment/ 3. -al		
	V	n ₁	P	V	n ₁	P
X-ize	252	47	0.187	0	0	0
X-ify	84	31	0.369	0	0	0
X-ate	460	30	0.065	0	0	0
X (simple)	391	7	0.018			

e.g. 1. *Mongolization, humidification, fantastication, spoilation*

Table 2: The productivity of deverbal-noun forming suffixes.

Table 3 displays the type number, hapax number, and productivity values concerning the deadjectival nominal suffixes *-ity* and *-ness* for the bases of *X-able*, *X-al*, *X-ic*, *X-ile*, *X-ar*, *X-ous*, and *X-ive*. We find that the productivity of *-ness* is slightly lower than that of *-ity* for the base *X-al*, although overall *[[X-al]-ity]* types are much more than overall *[[X-al]-ness]* types.

base	4.-ity			5.-ness		
	V	n ₁	P	V	n ₁	P
X-able	386	177	0.459	19	4	0.211
X-al	118	52	0.441	17	5	0.294
X-ic	35	19	0.543	2	1	0.500
X-ile	26	6	0.231	1	1	1.000
X-ar	24	17	0.708	0	0	0
X-ous	69	10	0.145	170	58	0.341
X-ive	29	7	0.241	88	26	0.295

² For this hapax-finding I am indebted to the research engine of <http://view.byu.edu/reg3.asp?c=aybfyfml>.

³ For the purpose of deciding which suffix is selected in creating a new word, we redefine a hapax as a word which occurs only once and whose rival formation is of zero frequency. Thus, when the frequency of *etymologic* is 1 and that of *etymological* is 29, *etymologic* is not regarded as hapax.

X (simple)	33	3	0.091	321	29	0.090
------------	----	---	-------	-----	----	-------

e.g. 4. *allowability, ornamentality, ellipticity*, 5. *mathematicalness, patheticness, rifeness*

Table 3: The productivity of deadjectival-noun forming suffixes.

3.2.2. Adjective-forming Suffixes

Table 4 exhibits the overall types, overall hapaxes, and productivity values concerning the denominal adjectival suffixes *-al*, *-less*, *-ic* and *-ical* for the bases of *X-(a)(t)ion*, *X-ment*, *X-oid*, *X-(o)logy*, and *X-ist*.

base	6.-al			7.-less			8.-ic			9.-ical		
	V	n ₁	P	V	n ₁	P	V	n ₁	P	V	n ₁	P
X-(a)(t)ion	220	86	0.391	11	5	0.455	0	0	0	0	0	0
X-ment	31	8	0.258	0	0	0	0	0	0	0	0	0
X-oid	17	3	0.176	0	0	0	0	0	0	0	0	0
X-(o)logy	0	0	0	0	0	0	32	0	0	159	69	0.434
X-ist	0	0	0	1	1	1.000	137	43	0.314	12	2	0.167

e.g. 6. *gradational, managerial, sphenoidal* 8. *historistic* 9. *malacological*

Table 4: The productivity of denominal-adjective forming suffixes.

3.2.3. Verb-forming Suffixes

Table 5 presents the overall types, overall hapaxes, and productivity values concerning the deadjectival verbal suffixes *-ize*, *-ify*, and *-ate* for the bases of *X-al*, *X-(i)an*, *X-ic*, and *X-ive*.

base	10.-ize			11.-ify			12.-ate		
	V	n ₁	P	V	n ₁	P	V	n ₁	P
X-al	106	42	0.396	0	0	0	0	0	0
X-(i)an	13	9	0.692	0	0	0	0	0	0
X-ic	14	5	0.357	0	0	0	0	0	0
X-ive	2	1	0.500	0	0	0	0	0	0

e.g. 10. *musicalize, pedestrianize, poeticize, comprehensivize*

Table 5: The productivity of deadjectival-verb forming suffixes.

3.2.4. A Suffix Sensitive to a Particular Prefix

Table 6 displays the overall types, overall hapaxes, and productivity values concerning the deverbal nominal suffixes *-ment*, *-(a)(t)ion*, and *-al* for the base beginning with *en-*.

base	-ment			-(a)(t)ion			-al		
	V	n ₁	P	V	n ₁	P	V	n ₁	P
en-X	59	12	0.203	0	0	0	0	0	0

e.g. *enfacement*

Table 6: The productivity of deverbal-noun forming suffixes.

4. Theoretical Implications

4.1. Generalizations

Three generalizations are drawn from the results of the present investigation. To begin with, a large number of words are created by the processes concerned. A total of 774 relevant hapaxes are found in BNC, which pushes up the related productivity values. The high rate of productivity elucidates a creative aspect of each productive suffix within a specific limited domain.

Secondly, productivity is a gradient and relative concept, and accordingly it is not defined in terms of a clear-cut binary opposition of “productive” and “unproductive.” Three cases of suffix selection are recognizable: (i) a suffix is automatically selected for a base form, (ii) a suffix is preferentially selected for a base form, and (iii) there is little preference in a choice between competing suffixes. The first case is well represented in Table 2. The nominalizer *-(a)(t)ion* is inevitably selected for the base form of *X-ify*; the *P* of *-(a)(t)ion* for nominalizing *X-ify* verbs is 0.369, while the corresponding *P* of *-ment/-al* is 0. Similarly, as demonstrated in Table 4, *-ical* is chosen in deriving adjectives from *X-(o)logy* nouns; the *P* of *-ical* for adjektivizing *X-(o)logy* nouns is 0.434, while the comparable *P* of *-all/-less/-ic* is 0. It should be noted here that there is a considerable difference in the degree of productivity between hapax-oriented productivity and type-oriented productivity; although the former defines *-ic* affixation to *X-(o)logy* as unproductive, while the latter may predict this process as fairly productive, since 32 word types in *X-(o)logic* are discerned in BNC.

The second case is well illustrated in Table 3. The nominalizer *-ity* has a strong affinity with the base forms of *X-able(-ible)/X-al/X-ic/X-ile/X-ar*; *-ity* is generally used in deriving abstract nouns from adjectives ending in these suffixes.⁴ We also find in Table 3 that the use of *-ness* takes precedence over the use of *-ity* to nominalize *X-ous* adjectives, as evidenced by the difference in narrow productivity between both suffixes. The *-ity* or *-ness* affixation to *X-ive* adjectives, displayed in Table 3, is an example of the third case. There is little preference in a choice between these nominalizers for the *X-ive* base; the *P* of *-ity* for nominalizing *X-ive* adjectives is 0.241 and the corresponding *P* of *-ness* is 0.295.

The final generalization that can be made is that suffixation to simple words is not productive. As indicated in Tables 2 and 3, the productivity values of *-(a)(t)ion*, *-ity*, and *-ness* for nominalizing simple words are, respectively, 0.018, 0.091, and 0.090; there is no suffix which productively combines with monomorphemic bases.

4.2. Generative Morphology and the Simplified Lexicon

The generalizations sketched above substantiate the theory of “generative” morphology and its system of lexical insertion. The point of the first one was that a considerable number of hapaxes in BNC confirm a productive facet of each relevant suffix within a narrow domain. This demonstrates the thesis of generative morphology: regular complex words are constantly generated by word formation. It should be noted that experimental evidence implies that while highly frequent words are stored and easily accessible, infrequent ones are generally created by some rule (Hay, 2003: 77-81). Therefore, hapax legomena in a large corpus provide a reliable and objective indicator of a high level of productivity—the potentiality of creating new words.

Importantly, complex words produced by a productive process are not stored at all but composed by rule as needed; words such as *Mongolization* and *managemental* are aptly coined in a particular context. Jackendoff (2002: 155-159) specifically states that inventive kind of word formation like this gets involved in working memory, where items are essentially composed by free combinatory rules. It is this kind of word coinage that crucially contributes to constructing the simplified and elegant lexicon; the succinct lexicon is obtained by reducing the number of listed items

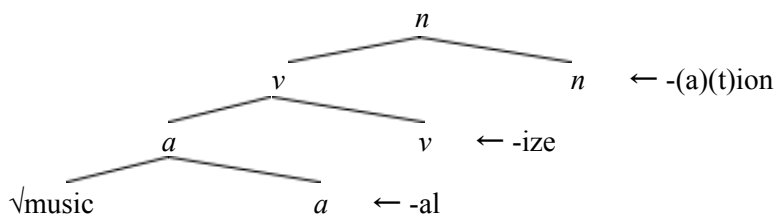
⁴ A case where the number of word types is extremely low may be problematic for our productivity measure; the *P* of *-ness* affixation to *X-ic* is very high, although there are only two types of *[[X-ic]-ness]* forms attested. In relation to this, Baayen and Lieber (1991: 818-819) suggest “the global productivity *P**”: the *P** of an affixation rule is defined in terms of its coordinates in the *P-V* interaction region, with productivity (*P*) on the horizontal axis and types (*V*) on the vertical axis; a productive affix occupies a central position in the region. According to this method, we may define a domain where productivity measurement is possible, sending the case in question outside the domain.

and generating unregistered items through word formation processes.⁵

4.3. Competition and Blocking

In a current theory of antilexicalism, Distributed Morphology (DM) (Embick, 2010), major word formation is situated within syntax: to construct a highly constrained grammar by severely restricting the morphological component and its relation with syntax, DM attributes the core properties of word construction to its syntactic structure while consigning the role of its formal make-up to the morphological module. In the DM framework, the syntactic structures of multiply affixed words are uniformly constructed by the merging of a root and category-defining heads, as shown diagrammatically in (1) (cf. Embick, 2010: 94-96).⁶

(1) *musicalization*



The second and third generalizations discussed in Section 4.1 imply that a word-form of each complex word is productively realized by Competition and Blocking. In the postsyntactic morphological module, specification of an adjectival, verbal, or nominal form is made by the addition of an appropriate affix to the base, and competition is carried out to decide which affix is chosen for a given base form in deriving a new item. In the case of nominalization, *-(a)(t)ion* wins out over its competitors for the base forms ending in *-ify*, *-ize*, and *-ate* (cf. Table 2). According to Embick’s framework, the lexical entries of *-(a)(t)ion* are formalized as “ $n \leftrightarrow -(a)(t)ion$ /LIST 4[^] LIST 4={Roots, ... v^{-ify} , v^{-ize} , v^{-ate} ...}, where v^{-ify} , for instance, stands for *-ify* final verbs.” The rival nominalizers *-ment* and *-al* are then prevented from joining to these base forms, following the narrowly defined blocking principle: competition takes place only between single morphemes (Embick and Marantz, 2008: 7); *-(a)(t)ion* competes with *-ment/-al* for an insertable nominalizer and the losers (*-ment/-al*) are blocked out. The same obtains for the verbalization of *musical* in (1) (cf. Table 5).

Another case of the second generalization is that a suffix takes precedence over its rival suffix to categorize a certain base form; *-ity* has priority to *-ness* for the *X-al* base, whereas *-ness* is prior to *-ity* for the *X-ous* base (cf. Table 3). In this case, the prior suffixal affinity is regulated in the related lexical entries, whereas the bases of the rival suffix are marked item-by-item in its entries. Thus, the lexical entries of *-ity* are “ $n \leftrightarrow -ity$ /LIST 7[^] LIST 7={Roots, [X-ous]₁, [X-ous]₂, [X-ous]₃, ... a^{-al} , ...},” while those of *-ness* are “ $n \leftrightarrow -ness$ /LIST 8[^] LIST 8={Roots, [X-al]₁, [X-al]₂, [X-al]₃, ... a^{-ous} , ...}.” The productive use of *-ity* for the *X-al* base generally blocks the addition of *-ness* to the base, although a set of *X-al* words which *-ness* takes are specified one by one in its entries.

Finally, the unpredictable bases of a suffix are specified item-by-item in its entries. Suffixation to simple words is unproductive (generalization 3) and hence which suffix preferentially combines with a given monomorphemic base is totally unpredictable. The simple base forms (roots) are then specified as in “ $n \leftrightarrow -ity$ /LIST 7[^] LIST 7={√civil, √null, √odd, √sane, ...}.” Another good example of this case is *-th* suffixation. Our BNC research shows that *-th* is almost always added to monomorphemic words and it takes a productivity value of 0 for this base form; ten *-th* derivatives are detected in BNC: *breadth* (575 tokens), *death* (19889), *depth* (2990), *length* (7049), *strength* (6946), *truth* (7930), *untruth* (35), *warmth* (1957), *width* (1141), *youth* (5308). Its entries can therefore be

⁵ See Stemberger and MacWhinney, 1988; Frauenfelder and Schreuder, 1992 for the related psycholinguistic experiments.

⁶ Root (√) is defined as bound morpheme that becomes the core of a word.

something like this: “ $n \leftrightarrow -th$ /LIST 9[^] _ LIST 9={√broad, √dead, √deep, √long, √strong, ...}.”⁷

5. Conclusion

We have proposed a new productivity measure—narrow productivity—for affixation, which crucially depends on hapax derivatives and their base forms, and then conducted an in-depth analysis of twelve kinds of derivatives identified in BNC to calculate the productivity values of relevant affixes. The results have disclosed some intriguing properties of affixation: a creative and generative aspect, competition and blocking of rival affixes, and their consequent implications for the systematic materialization of a word form. The proposed productivity measure and its consequences are expected to obtain further support by extensive research of a variety of affixes.

Acknowledgement

I would like to express my gratitude to three anonymous reviewers for their valuable comments and suggestions on an earlier draft of this paper. This work is partly supported by a Grant-in-Aid for Scientific Research (C) (No. 26370462) from the Japan Society for the Promotion of Science.

References

- Aronoff, M. (1976). *Word Formation in Generative Grammar*. Cambridge, MA: MIT Press.
- Baayen, H. and Lieber, R. (1991). Productivity and English Derivation: A Corpus-based Study. *Linguistics*, 29:801-843.
- Baayen, H. and Renouf, A. (1996). Chronicling *the Times*: Productive Lexical Innovations in an English Newspaper. *Language*, 72:69-96.
- Embick, D. (2010). *Localism versus Globalism in Morphology and Phonology*. Cambridge, MA: MIT Press.
- Embick, D. and Marantz, A. (2008). Architecture and Blocking. *Linguistic Inquiry*, 39:1-53.
- Frauenfelder, U. H. and Schreuder, R. (1992). Constraining Psycholinguistic Models of Morphological Processing and Representation: the Role of Productivity. *Yearbook of Morphology*, 1991:165-183.
- Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.
- Hay, J. (2003). *Causes and Consequences of Word Structure*. New York: Routledge.
- Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.
- Jespersen, O. (1949). *A Modern English Grammar on Historical Principles VI*. London: George Allen and Unwin.
- Lehnert, M. (1971). *Reverse Dictionary of Present-Day English*. Leipzig: VEB Verlag Enzyklopädie.
- Lieber, R. (2010). *Introducing Morphology*. Cambridge: Cambridge University Press.
- Plag, I. (1999). *Morphological Productivity: Structural Constraints in English Derivation*. Berlin: Mouton de Gruyter.
- Stemberger, J. P. and MacWhinney, B. (1988). Are Inflected Forms Stored in the Lexicon? In Hammond, M. and Noonan, M., Eds., *Theoretical Morphology*, pages 101-116. San Diego: Academic Press.

⁷ Two types of blocking should be recognized: (i) the most productive affix for a given class of base blocks the attachment of rival affixes to the base form (class-blocking); *-ment* and *-al* are prevented from joining to $[X-ize]_V$, and (ii) when an affix is added to a given item to derive a word, the corresponding derivatives are pre-empted (item-blocking); *-ity* and *-ness* are blocked from combining with \sqrt{long} .

Knowledge and Rule-Based Diacritic Restoration in Serbian

Cvetana Krstev

University of Belgrade
Faculty of Philology
cvetana@matf.bg.ac.rs

Ranka Stanković

University of Belgrade
Faculty of Mining and Geology
ranka@rgf.rs

Duško Vitas

University of Belgrade
Faculty of Mathematics
vitas@matf.bg.ac.rs

Abstract

In this paper we present a procedure for the restoration of diacritics in Serbian texts written using the degraded Latin alphabet. The procedure relies on the comprehensive lexical resources for Serbian: the morphological electronic dictionaries, the Corpus of Contemporary Serbian and local grammars. Dictionaries are used to identify possible candidates for the restoration, while the data obtained from SrpKor and local grammars assists in making a decision between several candidates in cases of ambiguity. The evaluation results reveal that, depending on the text, accuracy ranges from 95.03% to 99.36%, while the precision (average 98.93%) is always higher than the recall (average 94.94%).

1. Motivation

In Serbia, the use of Cyrillic alphabet is prescribed by law (Zakon, 2010: article 1), while the use of Latin alphabet is permitted in special situations (traffic signs, street names, etc.). However, due to historical and other reasons Latin alphabet is widely used.¹ One of the reasons is that the Cyrillic alphabet had poor support in the digital world before Unicode was fully implemented. But the Serbian version of the Latin alphabet does not coincide fully with the English version; some letters are not used – *q*, *w*, *x* and *y* – while other letters, the ones with diacritics – *š*, *đ*, *č*, *ć* and *ž* – are not represented in the ISO 8859-1 Latin 1 encoding scheme, the most used 8-bit superset of ASCII. These circumstances hindered the use of the Serbian variant of the Latin alphabet in some applications in the past and forced a search for other solutions. One of these solutions was to drop diacritics (use *s* instead of *š*, *z* instead of *ž*, and *c* instead of both *č* and *ć*) and replace *đ* with the digraph *dj*, which squeezed the Serbian Latin alphabet to ASCII.²

The full implementation of Unicode rendered these solutions unnecessary. And once the degraded use of the Latin alphabet seemed to belong to the past, the new social media applications, especially those relying on short messages, such as Twitter and SMS, revived it. These messages are usually not only short but also written very quickly and it is easier to use the basic Latin alphabet.³ Besides that, texts without diacritic marks, completely or partially, can emerge as a result of an inadequate transformation from PDF into raw text, as well as a result of a poor OCR.

In the past, some interesting Serbian texts were prepared using the degraded Latin alphabet, that could be used as a corpus material.⁴ Many short messages are produced every day and they are a source

¹Note that the Law on the usage of Language and Script is presented on the web site that collects all Serbian laws and regulations in Latin script http://www.paragraf.rs/propisi/zakon_o_sluzbenoj_upotrebi_jezika_i_pisama.html

²Such use of the Latin alphabet is in the computer jargon sometimes called *ošišana latinica* ‘shaved Latin’ or *ćelava latinica* ‘bold Latin’. One should note that, in the past, applications were not as user-friendly as they are today. Consequently, even when support for the Serbian Latin alphabet existed many users did not know how to use it.

³Note also that SMS messages cost more when characters beyond ASCII are used, since less characters can then be used in one message.

⁴One example: Web site <http://www.yuope.com/people/nena/Zabeleske/> contains a literary works from a number of Serbian writers, all written in the degraded Latin script.

for various research. However, all this cannot be done if the degraded Latin alphabet were not restored to the regular Serbian Latin alphabet.

This paper is organized as follows: in Section 2. we discuss some related work on the same and similar problems, in Section 3. we detail the lexical resources which serve as a base for our solution presented in Section 4. In Section 5., the results of the evaluation are discussed. Finally, in Section 6. we conclude with some remarks and hints for the use of similar procedures for solving similar problems.

2. Related Work

The problem of the degraded alphabet occurs in many languages and in various forms. One of its forms is the omission of the diacritics. Therefore, the solutions to this problem are named “diacritic restoration” or “diacritization”. A lot of work has been dedicated to solving this problem for many languages that use the Latin alphabet, including Croatian (Šantić et al., 2009), French (Yarowsky, 1999), Hungarian (Novák and Siklósi, 2015; Acs and Halmi, 2016), Lithuanian (Kapočiūtė-Dzikienė et al., 2017), Romanian (Tufiş and Ceaşu, 2008; Iftene and Trandabat, 2009; Petrică et al., 2014), Slovak (Hládek et al., 2013), Spanish (Atserias et al., 2012; Francom and Hulden, 2013), Turkish (Adali and Eryiğit, 2014) and Vietnamese (Pham et al., 2017). To the best of our knowledge no work was reported for Serbian, besides a solution aiming at several South Slavic languages (Ljubešić et al., 2016).⁵ Besides for the Latin script, a similar problem arises for the Arabic script (Alghamdi et al., 2010; Belinkov and Glass, 2015), and in that case it is sometimes named “the vowel restoration” since the diacritics are mainly used to represent the vowels and gemination. Finally, although many of the cited works claim that their systems are language independent, several authors presented specifically language-independent solutions or solutions that can be applied to several related languages. For instance, the solution reported in (Iftene and Trandabat, 2009) is aiming at the law-resourced languages, while those described in (Haertel et al., 2010) and (Ljubešić et al., 2016) were designed for the Semiotic languages and the South Slavic languages, respectively.

The incentive to develop a system for the diacritic (vowel) restoration is obvious – there is a need to obtain a correct text written according to the norms of a certain language. However, as we mentioned in Section 1., recently this problem has received the some attention due to the large masses of texts produced for many languages in the form of short messages using the “ASCII” Latin script (Acs and Halmi, 2016; Adali and Eryiğit, 2014; Ljubešić et al., 2016) that need further processing, for instance by a text-to-speech system (Petrică et al., 2014; Ungurean et al., 2008). Actually, our work on the diacritization in Serbian was spurred by a need to correct and normalize Twitter messages (Mladenović et al., 2017).

Concerning methods used, few systems are primarily rule-based (like (El-Sadany and Hashish, 1988) for Arabic verbs) or knowledge-based (for instance, (Tufiş and Ceaşu, 2008) for Romanian). The main drawback to these approaches is that they rely on lexicons and other NLP resources (e.g. POS taggers) that may not be available and/or would not cover the non-standard word forms (that are usually found in social media messages) (Ljubešić et al., 2016).

The diacritization problem is seen by some authors as a spelling-check problem (Atserias et al., 2012), a disambiguation problem (Yarowsky, 1999), a classification problem (Acs and Halmi, 2016; Adali and Eryiğit, 2014), or a machine-translation problem (Novák and Siklósi, 2015; Ljubešić et al., 2016; Pham et al., 2017). These problems are solved through various statistical approaches which can be grouped into two main categories: character-based and word-based. The attractiveness of the character-based approaches lies in the fact that they are language independent and do not need extensive resources (Alghamdi et al., 2010; Kapočiūtė-Dzikienė et al., 2017). The word-based methods are usually language-dependent as they rely on at least some language resources, while using some kind of a language model: n -grams (of words) (Atserias et al., 2012), Hidden-Markov-Models (HMM) (Gal, 2002; Ibraheem, 2017) or neural networks (Belinkov and Glass, 2015; Pham et al., 2017).

In this paper we discuss a rule-based and a knowledge based approach for restoring diacritics in a Serbian text written in the Latin script. An advantage of the rule/knowledge-based systems is that their

⁵There is a web page offering the solution for this problem <http://www.slovomajstor.com/>; however, the author(s) and the methods are not disclosed.

work is transparent and their results can more easily be explained and corrected, should that be necessary. Moreover, as we had a rich lexicon and other NLP tools at our disposal, the main reason for not using a rule/knowledge-based system was no longer relevant. Besides that, our work is inspired by the idea to develop a system that could be adapted to solve several related problems (as will be mentioned in Section 6.).

3. Textual Resources

The input of our procedure for restoring the diacritics is a Serbian text that does not use the diacritics. Once this text is tokenized, it consists of two types of simple words: (a) Words that will not be affected by our procedure, because they contain neither letters *c*, *s* and *z* that only use the diacritics nor the digraph *dj*, for instance *majka* ‘mother’; these words will be denoted W_a . (b) Words that will be affected by our procedure because they contain either one or more letters that sometimes contain the diacritics or the digraph *dj*, even if the diacritic is actually not missing or a sequence *dj* represents the consonant cluster. For instance, both *zvono* ‘bell’, and *zvaka* (representing *žvaka* ‘bubble-gum’) and *podjednak* ‘equal’ and *takodje* (representing *takođe* ‘also’) will be included among the words of the type W_b .

Our procedure is based on the following basic ideas:

1. For each word W_b we intent to offer all possible Serbian words that use diacritics, including the original word if it exists in the language. For instance, if $W_b = liscem$ then our procedure should identify the following candidates: *lišcem* ‘face (diminutive, instrumental case)’, *lišćem* ‘foliage (instrumental case)’, and the original word *liscem* ‘male fox (instrumental case)’. Potential word forms *lisćem*, *lišćem*, *liščem* would not be considered since they do not exist in the Serbian language. Simultaneously, if $W_b = lucice$, our procedure would identify *lučice* ‘port (diminutive, genitive case)’ and *lučiće* ‘to separate (future tense, 3rd person)’, but not the original word since it does not exist in the Serbian language. If a W_b word exists in Serbian and no words with diacritics can be derived from it, no further actions are performed on it. For all these words we refer to the morphological dictionary of Serbian.
2. For each word W_b all the possible candidates ($W_{b1}, W_{b2}, \dots, W_{bn}$) should be ranked according to the possibility of their occurrence in a text. In the case of the examples given above, *lišćem* is more frequent than both *liscem* and *lišcem* (the latter two have approximately the same frequency), while *lučice* is more frequent than *lučiće*. For these statistics we refer to the Corpus of Contemporary Serbian.
3. For each word W_b that has more then one possible candidate W_{bi} our procedure uses heuristics, lexicons and rules to choose the right one (more details in Section 4.).
4. For each word W_b that has no candidate at all our procedure does not offer any solution at this time.

In order to put the idea (1) into practice we have transformed the Serbian morphological dictionary (SMD)⁶ into the appropriate format. Namely, we have extracted from SMD of the simple forms all those with the diacritic marks (we will name them W_c words) as well as those that are of the type W_b . We have transformed all words of the type W_c to words of the type W_b by stripping the diacritics and replacing *d* with the digraph. At the same time, we removed all unnecessary information and recorded the original form. After that, we collated all identical word forms of the type W_b into one entry. For example, the original dictionary entries for the example given above were transformed in the following way:

```
liscem, lisac.N+Zool:ms6v    liscem, .X+CR=liscem
lišćem, lišće.N+Conc:ns6q    liscem, .X+CR=lišćem    liscem, .X+CR=liscem_lišćem_lišćem
lišcem, lišce.N+Dem:ns6q    liscem, .X+CR=lišcem
```

⁶Serbian Morphological Dictionaries cover both simple- and multi-word units (MWU). Dictionaries of simple-word units have more than 5 million grammatical forms generated from more than 141,000 lemmas, while the dictionaries of MWUs have more than 18,000 entries. To each grammatical form, whether it is a simple- or a multi word unit, various morphological, syntactic, semantic and other information is assigned (Krstev, 2008).

The obtained dictionary entry suggests that the word form $W_b = liscem$ represents three Serbian word forms $W_{b1} = liscem$, $W_{b2} = lišćem$, $W_{b3} = lišcem$. We will dub a dictionary of such entries SMD_DR. The same procedure was applied both to the dictionary of simple words and to the dictionary of multi-word units (MWU).

Our dictionary is case-sensitive. In SMD, all common words are written in lower-case letters, and they match the text words in a case-insensitive manner. In SMD, the initial letters of all proper names are upper-case, while all letters of acronyms are written in upper-case, and they match the text words in a case-sensitive manner. The same applies to our dictionary SMD_DR. This approach should allow us to cover all possibilities without introducing any incorrect candidates. Take, for instance, the word forms *liže* ‘to lick (present tense, 3rd person singular)’ and *Lize* ‘Lisa (feminine name, the genitive case)’:

```
liže, lizati.V+Imperf:Psz   lize, .X+CR=liže
Lize, Liza.N+NProp:fs2v   Lize, .X+CR=Lize   Lize, Liza.X+CORR=Lize_liže
```

If the word form $W_b = lize$ is written in lower-case, only one solution is offered – *liže*, if it is written in upper-case it can represent either a proper name or a verb form.

The resulting SMD_DR has 943,804 entries, 95% of which have only one candidate, while 4.5% have two candidates. The maximum number of candidates is 8 with one such entry:

```
Celice, .N+CR=čeliče_Celiće_Ćeliće_Čeliće_čeliče_ćelice_celice_celiće
```

A word form $W_b = Celice$ can represent a vocative of the surname *Čelik*, the accusative plural of the surnames *Celić*, *Ćelić* and *Čelić*, and various forms of the verbs *čeličiti* ‘to steel, to harden’ and *celiti* ‘to heal’ and the nouns *ćelica* ‘bald spot (diminutive)’ and *celica* ‘ground’.

In order to implement the idea (2) we enhanced the dictionary SMD_DR with additional information from 100 million word excerpt collected in the Corpus of Contemporary Serbian (SrpKor).⁷ From the list containing tokens and their respectable frequencies we calculated the total number of tokens (*totalNumTokens*).⁸ The frequency calculation was different for dictionary entries with initial upper-case and for those that were entirely in lower-case. In the first case, only occurrences with the initial letter in upper-case were taken in the account, while in the second case, all occurrences were counted, regardless of the case.

$$relFreq = Round\left(\frac{freq \cdot 10000000}{totalNumTokens + 0.5}, 0\right)$$

Relative frequency for 0 was 0, for 1–10 was 1, for 11–21 was 2, 22–32 was 3,... Maximal absolute frequency was 3,706,356 and corresponding relative 340,596. For MWUs frequencies were not calculated.

Occurrence in the corpus	Number of candidates					
	1	2	3	4	5	>6
no occurrence in the corpus	742,941	24,004	1,509	47	9	2
%	82.82	57.92	42.09	10.22	9.68	12.50
at least one occurs in the corpus, but not all		9,325	1,274	261	66	12
%		22.50	35.54	56.74	70.97	75.00
all occur in the corpus	154,136	8,115	802	152	18	2
%	17.18	19.58	22.37	33.04	19.35	12.5
Total	897,077	41,444	3,585	460	93	16

Table 1: Distribution of entries in SMD_DR according to the number of candidates and their occurrence in the Corpus of Contemporary Serbian

The data in Table 1 may may create an impression that the problem of diacritics restoration is not very severe – only 5% of entries from SMD_DR offer more than one correction, and many of offered

⁷<http://www.korpus.matf.bg.ac.rs/prezentacija/korpus.html>

⁸Tokens are defined as the strings containing letters of the Serbian Latin alphabet only.

candidates do not occur in SrpKor at all, and can thus be routinely rejected (in many applications). However, a changed perspective may help us understand that this issue is actually a very pressing one. There are 147 entries with two candidates and 4 entries with 3 candidates, and all of them occur in SrpKor with a high frequency ($refFreq \geq 100$). A prominent example is $W_b = reci$, where each candidate occurs in a 100 million word corpus with a frequency higher than 2,000:

`reci, .X+CR=reci(237)_reči(2607)_reči(1448)`

All forms are frequent and highly ambiguous: *reci* can be a form of the nouns *reka* ‘river’ and *redak* ‘line’ and of the verb *reći* ‘to say’, while *reči* is a form of the noun *reč* ‘word’.

4. The Procedure for Diacritic Restoration

In order to implement idea (listed as 3. in Section 3.) we have developed a set of rules and implemented them as cascades of the Finite-State Transducers (FSTs) in the corpus processing system Unitex⁹. The regular SMD is applied to texts that need to be corrected. Each word form is assigned a dictionary interpretation. This means that a W_b word form can either be left without an interpretation (the case of *lize* given in Section 3.) or some interpretation can be assigned to it (the case of *reci* – it can be either a form of the nouns *reka* or *redak*, as the dictionary suggests, or it has to be corrected).

We developed two working cascades: the first one retrieves data from the lexical resources – the dictionaries and the n -gram lists – and at the same time resolves straightforward cases, while the second one assigns one solution to each W_b word in a text by applying the set of rules, as given in Figure 1.

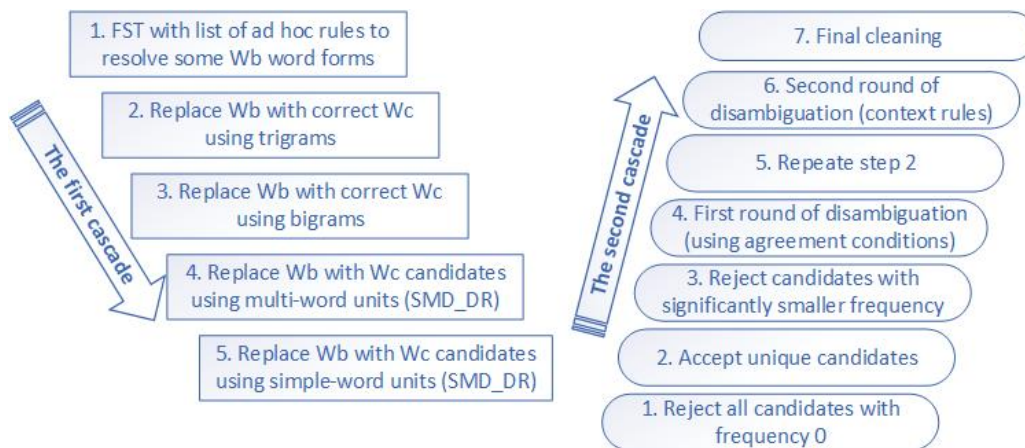


Figure 1: The two cascades of the Finite-State Transducers (FSTs)

The steps in the first cascade are:

1. This FST implements a list of *ad hoc* rules that are able to resolve some W_b word forms by analysing their context. For example, the word form *sto* can be a numeral ‘hundred’ or a noun ‘table’, or a W_b form of *što*, a relative and interrogative pronoun. The rules to confirm *sto* as correct are: (a) if *sto* is followed by word forms *puta* or *posto* as in the conventional phrases ‘hundred times’ and ‘hundred percent’ (in the later case *posto* is confirmed as well); (b) if it is followed by a numeral (e.g. *sto hiljada* ‘hundred thousand’); (c) if it is followed by an adjective-noun phrase in the genitive case plural, as required by the Serbian agreement rules. The application of the rules (b) and (c) is possible because the SMD dictionaries were applied to the text before its correction. Naturally, they will work only if the context words are not the W_b words in need of correction.
2. From the list of the most frequent trigrams obtained from SrpKor we have chosen 30 most frequent ones that contain at least one W_b word and replaced it with a correct W_c word. Two rules of this type are: *sto se tice* → *što se tiče* ‘concerning’ and *na taj nacin* → *na taj način* ‘in that way’.

⁹Unitex is a lexically-based corpus processing suite that has a strong support for finite-state processing – unitexgramlab.org

3. From the list of the most frequent bigrams obtained from SrpKor we have chosen 50 most frequent ones that contain at least one W_b word and replaced it with a correct W_c word. Two rules of this type are: *sto ce* → *što će* ‘that will’ and *je takodje* → *je takođe* ‘is also’.¹⁰
4. A SMD_DR of the multi-word units is consulted and a W_b multi-word form is replaced by the list of candidates assigned to the corresponding entry. For instance, *klucna rec* → *ključna reč(0)* ‘key word’. In the case of MWUs, a W_b word form has in most cases only one candidate; for that reason, the frequencies were not calculated for them and are currently not used by the second cascade.
5. A SMD_DR of the simple-word units is consulted and a W_b simple-word form is replaced by the list of candidates assigned to the corresponding entry. The W_b simple words are replaced if they (a) do not appear as such in SMD (the unknown words) or (b) they do appear in SMD and may be the correct choice.

After the application of the first cascade a new version of a text is obtained in which almost all W_b words are either confirmed, or corrected or have a list of possible solutions assigned to them. However, all this operations are not done without a trace; on the contrary, all modifications, as well as their type, are visible in the text, because that is important for the work of the second cascade. After the application of the first cascade a text is transformed (see Table 2).

the source text	the annotated text
Jer je imao dovoljno vremena da spreči zlocin cak i nakon reci kojima je podstrekivao sina.	Jer je imao dovoljno vremena da 5a_(spreči(302)_spreči(0)) 5a_(zločin(456)) 3_(čak i) 1_(nakon reči) kojima je podstrekivao 5b_(sina(518)_šina(54)).
U novinama vise nije bilo ni reci o ratnoj steti.	U novinama 5b_(vise(35)_više(17628)) 1_(nije bilo ni reči o) 4_(ratnoj šteti(0)).

Table 2: Two sentences without the diacritics (left), and their annotated version after the application of the first cascade (right). Numbers correspond to the steps in the cascade that perform the annotation.

The role of the second cascade is to produce a clean text with the diacritics restored. For that purpose, dictionaries (SMD) are applied to an annotated text. As a result, all words in the text (W_a , W_b and W_c) obtain one or more dictionary interpretation. Those W_b words that are not words in Serbian (or are not in SMD), of course, do not get a dictionary interpretation. Words of the type W_c are those that resulted from the work of the first cascade. The steps of the second cascade are the following:

1. All candidates that do not occur in SerKor (frequency is 0) are rejected, if there is an alternative candidate occurring in SerKor (frequency higher than 0), e.g. *1_5b_(zemplja(1518)_žemlja(0))* → *1_5b_(zemplja(1518))* (*zemplja* ‘earth/ground’, *žemlja* ‘bread roll (Ijekavian)’).
2. This FST accepts unique candidates (only one candidate exists for a W_b word according to SMD), e.g. *5a_(ništa(3531))* → *ništa* ‘nothing’.
3. This FST rejects candidates which have significantly lower frequency of occurrence than some alternative candidate. This FST is subject to changes depending on user’s views what “significantly smaller” means: ten times less, hundred times less, less than hundred/greater than hundred, etc. For instance, *5b_(tacno(1)_tačno(1219))* → *5b_(tačno(1219))* (*tačno* ‘right/correct’, *tacno* ‘tray (the vocative case’).
4. This FST performs the first round of disambiguation. It looks in the unambiguous context (either W_a words or the previously disambiguated words) of the list of possible candidates that can confirm

¹⁰The size of lists of bigrams and trigrams was chosen rather arbitrarily in belief that the ambiguity problem will not occur among the most frequent n -grams. This decision has to be reconsidered in future.

at least one of the candidates and reject others. Possibilities are: (a) an adjective that is unambiguous is followed by a list of candidates among which is a noun that agrees with the adjective in the gender, the number and the case, for instance, *žrtvene 5b_(jarče(2)_jarce(1))* → *žrtvene 5b_(jarce(1))* ‘lit. sacrificial goat; scapegoat’; (b) a noun that is unambiguous is preceded by a list of candidates among which there is an adjective that agrees with the noun in the gender, the number and the case, e.g., *5b_(čelo(212)_celo(181)) popodne* → *5b_(celo(181)) popodne* ‘whole afternoon’; (c) a preposition that is unambiguous is followed by a list of candidates among which there is an adjective, a noun or a pronoun in the case required by the preposition, e.g., *Iz 5b_(reci(237)_reči(2607)_reči(1448))* → *Iz 5b_(reči(2607))*. If among the candidates at least one is confirmed by the context, those that are not confirmed are rejected, while all that are confirmed are accepted.

5. The step 2) is repeated, because, as a result of the intervening steps, some candidates may have become unique.
6. The second round of disambiguation takes into account the context with regard to: (a) some specific candidate lists, like *reci(237)_reči(2607)_reči(1448)* or *nišu(820)_nisu(9675)*, or some more general cases: (b) the reflexive particle *se* followed by a form of a reflexive verb; (c) the negative form of the auxiliary verb *neće* followed by an impersonal verb form (such as the infinitive); (d) the particle *ne* followed by a personal verb form. Examples of these decisions are: (a) *1_5b_(nišu(820)_nisu(9675))* → *Nišu* (since an upper-case initial letter is required in the middle of a sentence, it most probably refers to *Niš*, a city in Serbia); (b) *(da) se 5a_(suši(30)_šuš(1))*; → *(da) se suši* ‘(to) dry itself’; (c) *(da se) neće 5b_(obuci(49)_obući(30)_obući(5))* → *(da se) neće obući* ‘not (to) dress oneself’; (d) *ne 5b_(tući(67)_tući(12)_tuci(1))* (me) → *ne tuci* (me) ‘do not beat (me)’.
7. In the last step, apart from the final cleaning (such as the deletion of the duplicates) the last decisions are made in order for the resulting text to be completely resolved: (a) in 5b cases (among candidates one is without diacritics), the one without diacritics is chosen; (b) in 5a cases (among candidates all are W_c words) the one with the highest frequency is chosen.

5. Evaluation

In order to estimate the extent of the problem as well as how various rules contributed to its solution in Table 3 we present the data that correspond to the annotation and disambiguation of steps of the first and the second cascade when applied to two sample texts, one belonging to Ekavian and the other to Ijekavian pronunciation. The samples have similar size, the Ekavian has 2,024 word tokens, Ijekavian 1,930 word tokens. One cannot assume that a number of changes by the first and the second cascade should sum up to equal totals, since the second cascade sometimes resolves several annotations together, while, on other hand, other annotations are addressed in several steps. Also, the disambiguation steps, both in the first and in the second cascade, sometimes resolve more than one W_b word. The importance of the dictionaries is justified by the fact that in both texts, neither of which is very long, there were unique candidates with 0 frequency in SrpKor (5 in the Ekavian text and 7 in the Ijekavian text).

The first cascade			The second cascade		
Step	Text Ek	Text Ijk	Step	Text Ek	Text Ijk
1 (disambiguation)	8	12	1 ($W_c \notin$ SrpKor)	23	20
2 (trigrams)	1	1	2 (W_c is unique)	210	311
3 (bigrams)	9	16	3 (frequency)	55	64
4 (MWU)	3	0	4 (disambiguation 1)	12	5
5a ($W_b \notin$ SMD)	201	257	5 (W_c is unique)	96	78
5b ($W_b \in$ SMD)	138	145	6 (disambiguation 2)	2	4
			7 (final cleaning)	64	91

Table 3: The contribution of each step in cascades to the diacritic restoration.

For evaluation we used a set of 65 correctly typed documents of different length, type and domain. These documents are new, that is they are not part of the SrpKor used for the calculation of frequencies. We prepared all documents for evaluation by stripping the diacritics, and than applying the restoration procedure.¹¹ First, we aligned sentences of the source and restored files. After that, words were aligned and compared. For each document we counted the number of words that were correctly restored TP (true positive), number of words that were rightly not restored TN (true negative), number of erroneously restored words FP (false positive), number of words without expected restoration FN (false negative). In Table 4 the first two rows contain absolute and relative values that were calculated by taking into account only different word tokens (types), while the second two rows contain results that were obtained by counting all the occurrences.

Calculating	Total	W_b	W_a	TP	TN	FP	FN
by types	836,764	549,978	286,786	214,249	318,232	3,198	14,299
relative %	1.000	65.727	34.273	38.956	57.863	0.581	2.600
by tokens	3,493,785	1,682,680	1,811,105	575,618	1,070,181	6,211	30670
relative % (W_b)	1.000	48.162	51.838	34.208	63.600	0.369	1.823
relative % (Total)				16.475	30.631	0.178	0.879

Table 4: The basic data about the evaluation set, and evaluation results on the whole set.

We calculated the standard measures: the precision, the recall, the accuracy and F_1 for each document. The average, the maximum and the minimum values of these measures as observed in this set are presented in Table 5. Note that all these measures were calculated only for the words that were candidates for restoration (W_b words). The overview of the fluctuation of four measures for a few selected documents is given in Figure 2. It can be seen that, except in the case of one specific document, the precision is always significantly higher than the recall.

Calculating		Precision	Recall	Accuracy	F-measure		
types	average %	98.529	93.744	96.819	96.077		
	maximum %	99.677	96.142	98.097	97.671		
	minimum %	91.611	89.534	94.416	93.005	Correct	Incorrect
tokens	average %	98.933	94.941	97.808	96.896	98.944	1.056
	maximum %	99.958	98.428	99.358	99.187	99.647	2.247
	minimum %	95.752	88.467	95.029	92.855	97.753	0.353

Table 5: Precision $P = tp/(tp + fp)$, Recall $R = tp/(tp + fn)$, Accuracy $Acc = (tp + tn)/(tp + tn + fp + fn)$, F-measure $F_1 = 2PR/(P + R)$; correct words $((T - (fp + fn))/T)$, incorrect words $(fp + fn)/T$.

We compared results that were obtained in our sample of 65 texts with those published for Croatian in (Šantić et al., 2009) and Croatian and Serbian in (Ljubešić et al., 2016). In order to compare our results with those presented in (Šantić et al., 2009: 317) we calculated the percentages of correct and incorrect words in the restored text (the two last columns in Table 5) and we observed that, on the average, we obtained a slightly higher number of correct words (98.94 vs. 98.81). In order to compare our results with those discussed in (Ljubešić et al., 2016: 3615) we calculated percentages of false positives and false negatives relative to the total number of tokens (not just W_b) and we observed that our precision (0.18 vs. 0.12) and recall (0.88 vs. 0.41) are slightly lower.¹² In future we plan to apply our procedure and some language independent tools to the same set of texts (both standard and non-standard) in order to test and evaluate different approaches in the proper manner.

¹¹The same approach was previously employed by many authors who were concerned with similar problems: (Šantić et al., 2009: 316), (Tufiş and Ceaşu, 2008: 6), (Iftene and Trandabat, 2009: 39), (Francom and Hulden, 2013: 3).

¹²In both cases we compared only results that authors in (Šantić et al., 2009) and (Ljubešić et al., 2016) obtained on a fully diacriticized text and/or on a standard text.

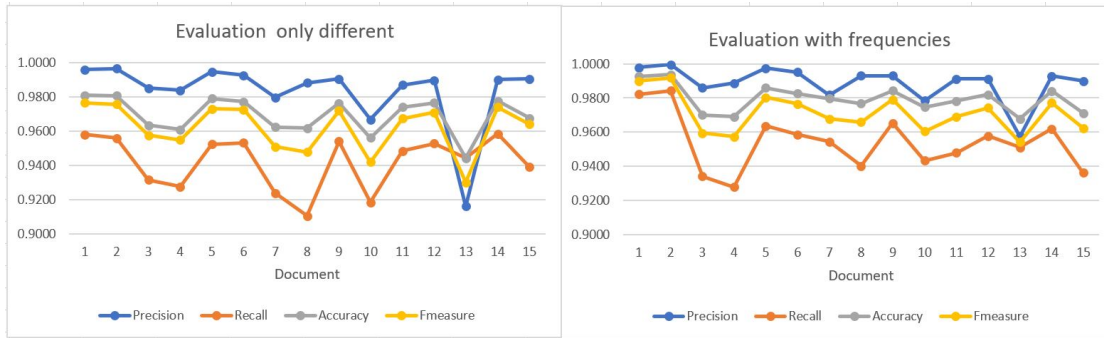


Figure 2: The evaluation report for a subset of documents: left – calculation on the W_b types; right – calculation on the W_b tokens

6. Concluding Remarks and Future Work

In this paper, we have shown that the problem of the diacritic restoration can be successfully solved by using a rule-based approach that relies on the lexical resources. This solution exhibits the advantage of transparency which is usually characteristic of such methods. Namely, any successful or unsuccessful change of the text can be easily explained and that can lead to the improvement of the solution. Our approach has some additional advantages. First, the improvement of the lexical resources, longer and more reliable lists of bigrams and trigrams, the enrichment of the e-dictionaries, particularly dictionaries of MWUs, will contribute to system’s better results. Second, the system is highly modular, which means that the order of steps can be easily changed and some steps removed or replaced for particular purposes. There are some disadvantages as well. First of all, considerable time was invested in its development. Also, its performance time does not make this solution applicable for interactive applications (e.g. mobile devices). Second, the extensive use of dictionaries implies that the procedure works only on the standard and the reasonably correct texts (missing only the diacritics) This means that its performance would be less impressive on non-standard texts, such as the Twitter posts. For non-standard texts, the procedure has to be used in conjunction with the other tools that deal with abbreviations, non-standard spelling, foreign words, etc., as we have already done (Mladenović et al., 2017).

The solution discussed in this paper can be adapted for solving some similar problems. First, the same solution can be applied to the texts that are missing diacritics only partially. In that case, only dictionaries SMD_DR have to be modified. For instance, for the example from Section 3. the dictionary entries would be, not only

```
lišcem, .X+CR=lišcem_lišćem_lišćem,
```

but also

```
lišćem, .X+CR=lišćem_lišćem
lišćem, .X+CR=lišćem.
```

Next, very promising experiments have already been conducted aiming to correct the texts obtained by OCR and to transform Serbian texts from one variant to another (from Ekavian to Ijekavian and vice versa, e.g. *lepa devojka* (Ek) ‘beautiful girl’ ↔ *lijepa djevojka* (Ijk)).

Finally, our goal is also to experiment with the hybrid solutions that would use explicit language models in the candidate selection phase.

Acknowledgements

This research was partially supported by the Serbian Ministry of Education and Science under the grants #III 47003 and 178006.

References

- Acs, J. and Halmi, J. (2016). Hunaccent: Small Footprint Diacritic Restoration for Social Media. In Utka, A., Vaičėnienė, J., and Butkienė, R., Eds., *Normalisation and Analysis of Social Media Texts (NormSoMe) Workshop Programme*, pages 1–4.
- Adali, K. and Eryiğit, G. (2014). Vowel and diacritic restoration for social media texts. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 53–61.
- Alghamdi, M., Muzaffar, Z., and Alhakami, H. (2010). Automatic restoration of Arabic diacritics: a simple, purely statistical approach. *Arabian Journal for Science and Engineering*, 35(2C):125–135.
- Atserias, J., Fort, M. F., Nazar, R., and Renau, I. (2012). Spell Checking in Spanish: The Case of Diacritic Accents. In *LREC*, pages 737–742.
- Belinkov, Y. and Glass, J. R. (2015). Arabic Diacritization with Recurrent Neural Networks. In *EMNLP*, pages 2281–2285.
- El-Sadany, T. and Hashish, M. (1988). Semi-automatic vowelization of Arabic verbs. In *10th NC Conference, Jeddah, Saudi Arabia*.
- Francom, J. and Hulden, M. (2013). Diacritic error detection and restoration via part-of-speech tags. In *Proceedings of the 6th Language and Technology Conference*.
- Gal, Y. (2002). An HMM Approach to Vowel Restoration in Arabic and Hebrew. In *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages, SEMITIC '02*, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Haertel, R. A., McClanahan, P., and Ringger, E. K. (2010). Automatic Diacritization for Low-resource Languages Using a Hybrid Word and Consonant CMM. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10*, pages 519–527, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Hládek, D., Staš, J., and Juhár, J. (2013). Diacritics Restoration in the Slovak Texts Using Hidden Markov Model. In *Language and Technology Conference*, pages 29–40. Springer.
- Ibraheem, A. O. (2017). *A Concept Paper on a New Model for Automatic Diacritic Restoration*.
- Iftene, A. and Trandabat, D. (2009). Recovering diacritics using Wikipedia and Google. In *Knowledge engineering: Principles and techniques, Proceedings of the international conference on knowledge engineering KEPT2009*, pages 37–40.
- Kapočiūtė-Dzikiene, J., Davidsonas, A., and Vidugirienė, A. (2017). Character-Based Machine Learning vs. Language Modeling for Diacritics Restoration. *Information Technology And Control*, 46(4):508–520.
- Krstev, C. (2008). *Processing of Serbian – Automata, Texts and Electronic dictionaries*. Faculty of Philology, University of Belgrade.
- Ljubešić, N., Erjavec, T., and Fišer, D. (2016). Corpus-based diacritic restoration for South Slavic languages. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.
- Mladenović, M., Krstev, C., Mitrović, J., and Stanković, R. (2017). Using Lexical Resources for Irony and Sarcasm Classification. In *Proceedings of the 8th Balkan Conference in Informatics, BCI '17*, pages 13:1–13:8, New York, NY, USA. ACM.
- Novák, A. and Siklósi, B. (2015). Automatic Diacritics Restoration for Hungarian. Association for Computational Linguistics.
- Petrică, L., Cucu, H., Buzo, A., and Burileanu, C. (2014). A Robust Diacritics Restoration System Using Unreliable Raw Text Data. In *Spoken Language Technologies for Under-Resourced Languages*.
- Pham, T.-H., Pham, X.-K., and Le-Hong, P. (2017). On the Use of Machine Translation-Based Approaches for Vietnamese Diacritic Restoration. *arXiv preprint arXiv:1709.07104*.
- Šantić, N., Šnajder, J., and Bašić, B. D. (2009). Automatic Diacritics Restoration in Croatian Texts. *INFUTURE2009: Digital Resources and Knowledge Sharing*, pages 309–318.

- Tufiş, D. and Ceaşu, A. (2008). DIAC+: A professional diacritics recovering system. *Proceedings of LREC 2008*.
- Ungurean, C., Burileanu, D., Popescu, V., Negrescu, C., and Dervis, A. (2008). Automatic diacritic restoration for a TTS-based e-mail reader application. *UPB Scientific Bulletin, Series C*, 70(4):3–12.
- Yarowsky, D. (1999). A comparison of corpus-based techniques for restoring accents in Spanish and French text. In *Natural language processing using very large corpora*, pages 99–120. Springer.
- Zakon, Ed. (2010). *Zakon o službenoj upotrebi jezika i pisma*[Law on Official Usage of Language and Script]. Službeni glasnik Republike Srbije.

Perfect Bulgarian Hyphenation, or how not to stutter at end-of-line

Anton Zinoviev

Sofia University “St. Clement of Ochrid”
Institute of Information and Communication Technologies
at Bulgarian Academy of Sciences,
anton@lml.bas.bg

Abstract

What is Perfect Bulgarian Hyphenation? We know that it has to be based somehow on the syllables and on the morphology but considering that these two factors often contradict each other, how exactly are we going to combine them? And speaking about syllables, what are they and how are we going to determine them? Also, how are we going to find the morphemes in the words? Don't we have to develop an electronic derivational dictionary of the Bulgarian language? Isn't all this going to be forbiddingly difficult?

1. Foreword

What heartless man is not going to sympathise with an intelligent speaker whose stuttering distracts the listeners, the thoughts behind his words remaining unheard? Demosthenes had to train in a cave with pebbles in his mouth and a sword over his shoulder. He had to make his speeches more apprehensible, this was a matter of life and death. In this paper we shall see that in the written language, too, there can be distracting things. In the spoken language the speech therapists fight the stuttering and in the written language the professional printers do the same. For example, they know that it is preferable not to use boldface. They also know that irregular white space is distracting, so it has to be eliminated by proper hyphenation. Likewise, the hyphenation should be done in such a way that while the eyes of the reader are moving from a line to the next one, his expectations about what follows are not deceived.

Indeed, it is this striving for clarity what has made the English hyphenation so complex. One peculiarity of the English language is that the pronunciation of a vowel depends on whether in its morpheme it belongs to an open or to a closed syllable. For example, the vowel *e* of the morpheme *hyphen* of the word *hyphenation* is part of the closed syllable *phen*. Consequently, the vowel *e* is pronounced as if it is in a closed syllable even though its syllable in the word *hyphenation* is the open syllable *phe*. If we hyphenate this word as *hyphe-nate* we will confuse the reader because while his eyes are still moving from *hyphe-* to the next line, he will expect that *e* is part of an open syllable with pronunciation as in *me* or *bee*. Therefore, the English hyphenation prefers not to change the apparent closeness of the syllables. This explains cases like *collect-ing*, *mod-el*, *sec-ond*, *trav-el*. It also explains the hyphenation of homonyms with equal spelling but different pronunciation, such as *prog-ress* and *pro-gress*, *rec-ord* and *re-cord*, *eve-ning* and *even-ing*.

2. Bulgarian Affairs

As for the Bulgarian hyphenation, it has always been governed by the same two factors as the hyphenation of most other languages—the syllables and the morphology. As in most languages, the case when a syllable boundary coincides with a morpheme boundary is clear. Uncertainties arise only when the syllables and the morphology specify different positions for word division. For a relatively long period the Bulgarian hyphenation was done intuitively and according to the existing tradition. Formal rules existed only about the most important cases. It seems the earliest attempts to formulate extensive exact rules about the Bulgarian hyphenation were from 1945 (Andreychin, 1945; Hadzhov and Minkov, 1945).

Keywords: hyphenation, morphology, syllables, phonology, Bulgarian

Unsurprisingly, these rules turned out to be complex. Due to this complexity, many mistakes were made. Especially the hyphenation in the newspapers was more or less arbitrary. So, instead of fixing the newspapers, someone decided that it would be easier to fix the hyphenation rules. In result, in 1983 the Institute for Bulgarian language published new hyphenation rules (Georgieva and others, 1983). These rules broke completely with the existing tradition. The morphology was no longer a factor and the syllables were proclaimed as the main ruling factor. However, the syllables were not to be determined according to the most convenient pronunciation, but rather by some non deterministic mechanistic rules.

Because the morphology was ignored, there were some absurd hyphenations, such as *авток-луб* (*avtok-lub* 'moto c-lub') and *вакуу-апарат* (*vacuu-maparat* 'vacuu-m apparatus'). In many cases the mechanistic hyphenation rules were too permissive. For example for the word *агентство* (*agentstvo* 'agency') we could have *аген-тство* (*agen-tstvo*), *агент-ство* (*agent-stvo*), *агентс-тво* (*agents-tvo*), *агентст-во* (*agentst-vo*), all at once. This was so despite that the pronunciation of the parts *агентст-* (*agentst-*) and *-тство* (*-tstvo*) was clearly impossible. In other cases the rules were too restrictive. For example the hyphenation *на-дро-бя* (*na-dro-bya* 'crumble') was forbidden despite that it seems that this is the most natural syllable division (and in addition, it is in agreement with the morphology).

However bad the new rules were in some aspects, they were good about the following: they were exact, unambiguous and they were easy to implement in software. The earliest mathematical analysis of the new Bulgarian hyphenation rules was by Noncheva (1988). She proposed a mathematical formalisation of the hyphenation rules in a table of 22 rows. In the same year, Belogay (1988) proposed an alternative formalisation with only 9 rules. Belogay proved that his rules were consistent and that they formed a minimal set. The rules of Belogay had a negative character—every hyphenation which was not forbidden by a rule was permitted. The work of Belogay was not limited to merely a mathematical analysis of the Bulgarian hyphenation rules. In his paper he published a short algorithm in Pascal which implemented these rules. It didn't take long for this algorithm to be used in various text processing software. The algorithm of Belogay was famous for many years. Even as late as 1997 the author of one book about \TeX (Vasilev, 1997) didn't care to give any explanations but simply wrote about "the algorithm of Belogay" as something well known to the reader.

The earliest implementations of the Bulgarian hyphenation in \TeX did not rely on the internal hyphenation algorithm of \TeX . Instead, an external tool implementing the algorithm of Belogay was used to insert soft hyphens in all Bulgarian words. The first usable Bulgarian hyphenation patterns for \TeX were developed by Georgi Boshnakov in 1994. In order to solve the encoding problem, Boshnakov had developed \TeX fonts supporting the MIK encoding (the prevalent encoding at that time in Bulgaria). This allowed him to introduce a fully working implementation only a few months after $\LaTeX 2_\epsilon$ became the official \LaTeX version. Later Boshnakov modified his work with the Babel system. The hyphenation patterns of Boshnakov did their job well enough, so for almost quarter a century after their initial creation, they remained the only Bulgarian hyphenation patterns in the standard distributions of \TeX and \CTAN .

The algorithm of Belogay and the hyphenation patterns of Boshnakov adhered to the official hyphenation rules of 1983. Nevertheless, the new rules were not universally accepted. Even today, the traditional rules by Andreychin (1945) are mentioned in various places in Internet. They are also included in some grammar books (Pashov, 1989; Stoyanov, 1993).

In 1995 Atanas Topalov defended a Masters thesis in the Faculty of Mathematics and Informatics at Sofia University titled "Algorithms and software about text processing" (Topalov, 1995). One of the main topics in his thesis was the Bulgarian hyphenation. Topalov criticised vehemently the official hyphenation rules and their total disregard of the morphology. He wrote:

If we look at the history of the problems of the hyphenation, we will discover something very strange. Instead of the expected involvement with the depths and aspiration for more admissible and satisfactory style, we can find a growing tendency for simplification. One unpleasant discovery is that the development of the hyphenation software stays firmly on the principle "let us do the easiest thing". The earliest works which have been studied are from 1978. It turned out that they present the best approach concerning the automated hyphenation.

In 1999 in a paper about the automated Bulgarian hyphenation, Koeva (1999) published a list of

hyphenation patterns that could be used as a basis for automated hyphenation. In 2004 with the help of Stoyan Mihov she formalised these rules with regular relations and rewriting rules. They were implemented in a software product named ItaEst which provided Bulgarian hyphenation and grammar checking for various software products of Microsoft Corp. and Apple Inc. The hyphenation rules of Koeva were more permissive than the official rules. For example they permitted cases such as се-стра (*se-stra* 'sister'), ай-сберг (*ay-sberg* 'iceberg') and материа-лна (*materia-lna* 'physical').

In 2000 Anton Zinoviev created new hyphenation patterns for T_EX. In 2001 Radostin Radnev used the hyphenation patterns of Zinoviev in his free grammar dictionary of Bulgarian. From there the work of Zinoviev propagated to OpenOffice, LibreOffice and various online dictionaries, including bg.wiktionary.org and rechnik.chitanka.info. However Zinoviev didn't bother to make his work officially available in the various T_EX distributions and CTAN.

The hyphenation patterns of Zinoviev were more restrictive than the official rules. For example in consonant sequences like тст (*tst*) in братство (*bratstvo* 'brotherhood'), the two equal consonants т (*t*) were always separated, so that братст-во (*bratst-vo*) was forbidden. The hyphenation was forbidden after a sonorant consonant following an obstruent consonant. For example отм-ра (*otm-ra* 'die out') was forbidden but от-мра (*ot-mra*) was permitted. Also, the hyphenation separated a pair of two kindred, one voiced and one voiceless consonants. For example субп-родукт (*subp-rodunkt* 'subproduct') was forbidden and суб-продукт (*sub-produkt*) was permitted.

Eventually, in 2012 the Institute for Bulgarian language published revised hyphenation rules (Murdarov and others, 2012). The new rules are even more liberal than the rules of 1983. While absurdities such as авток-луб (*avtok-lub*) and вакуу-апарат (*vakuu-maparat*) remain valid, the main advantage of the new rules is that the natural hyphenations авто-клуб (*avto-klub*) and вакуум-апарат (*vakuum-apatat*) are now also valid. It seems that the linguists at the Institute for Bulgarian Language have recognised that good hyphenation is a complex matter. They no longer attempt to invent universal rules about everything. Instead, they provide some very permissive rules while the good application of these rules is leaved to the discretion and the experience of the printers and the developers of hyphenation software.

The present work was carried out on the initiative of the leader of Bulgarian localisation team of the browsers Mozilla and Firefox. In 2017 he contacted me with an inquiry about the best automated Bulgarian hyphenation. Since the new official hyphenation rules were so permissive, I told myself: "Great, it seems I will be free to implement the hyphenation in any way I see fit or I deem appropriate. Good or bad, there will be fair chances that my implementation will be in compliance with the official rules. If I want to make a computer implementation of the *Perfect Bulgarian Hyphenation*, my hands will be untied and I will be free to act." So, to act I decided.

3. Plan

Evidently, we are coming to the real question: what is *Perfect Bulgarian Hyphenation*? We know that it has to be based somehow on the syllables and on the morphology but considering that these two factors often contradict each other, how exactly are we going to combine them? And speaking about syllables, what are they and how are we going to determine them? Also, how are we going to find the morphemes in the words? Don't we have to develop an electronic derivational dictionary of the Bulgarian language? Isn't all this going to be forbiddingly difficult? Let us delay no more and move right to the answers.

4. Combining Syllables with Morphology

Let us recall that when the English printers decided to hyphenate *beam-ing*, *draw-ing*, *stew-ing*, etc., they did so in order to make the reading easier. Then some grammarian noticed these specific cases and proclaimed the general rule that in all present participles we have to hyphenate before the ending *-ing*. This was a generalisation made by someone who admired general grammatical rules but didn't really understand the real objective of the hyphenation. In result, now we have deceiving cases, such as *hat-ing*.

Therefore, whatever rules about the hyphenation we invent, we should never lose sight of its main objective. And this objective is to make the reading smoother and easier. When our eyes are moving from

one line to the next, we know only the first part of the hyphenated word and we have only expectations about what follows. Any unnecessary ambiguity is bad. Creating wrong expectations and fooling them is worse. Bad hyphenation certainly is capable to confuse and to disturb the readers.

Ignoring completely the morphology is one good way to deal with the conflict between syllables and morphology. We the people are adaptive creatures. We can get used to any rules as long as they are not too unreasonable. And hyphenating according to the syllables of the word is certainly not a totally unreasonable way to hyphenate.

Nevertheless, there are cases when hyphenation according to the morphology creates less confusion. We already saw one such case: we have to respect the constituents of a compound word. Divisions such as авток-луб (*avtok-lub*) and вакуу-мапарат (*vakuu-maparat*) are extremely irritating.

Second in severity (but far more numerous, so also important) is the case with the word prefixes. While the eyes of the reader still look at the first part of the word, the rest is unknown. At this point, it is very important not to deceive the expectations. For example, when the reader sees на- (*na-*) at the end of the line, he will expect that this is the prefix на- (*na-*) with semantics 'achieve a state after accumulation'. This expectation will be fooled if this wasn't really a prefix, but a deceiving hyphenation of the word на-диграя (*na-digraya* 'outplay') where the real prefix is not на- (*na-*) but над- (*nad-*) with semantics 'attain more than'. Even more confusing is the case when we see над- (*nad-*) at the end of line and this wasn't really the prefix над- (*nad-*) but a deceiving hyphenation of the word над-ремя (*nad-remya* 'have dozed enough') where the real prefix is not над- (*nad-*) but на- (*na-*). Such hyphenations distract the reader and make the reading more difficult.

The traditional Bulgarian hyphenation rules (Andreychin, 1945; Hadzhov and Minkov, 1945) prescribed that in all cases the prefixes should be respected. In some cases the hyphenation was able to differentiate between two homonyms. For example пре-дреша (*pre-dresha* 'change clothes') but пред-реша (*pred-resha* 'predetermine') or прес-пите (*pres-pite* 'the snow-drifts') but пре-спите (*pre-spite* 'sleep for overnight'). On the other hand, the requirement to respect the suffixes was significantly relaxed: they should be preserved only when this doesn't create impossible syllables. Indeed, a hyphenation хлеб-ар (*hleb-ar* 'baker') is completely unwarranted despite that хлеб (*hleb*) is the root, -ар (*-ar*) is the suffix and both morphemes are productive in the modern language.

How can we explain the different treatment of the prefixes and the suffixes? If we try to shout rhythmically, syllable by syllable the word хлебар (*hlebar*), then the shouting хлеб-ар (*hleb-ar*) will be very unnatural and strained. However, if we do the same experiment with the prefixes, we will find something unexpected: it is quite possible (even if somewhat inconvenient) to rhythmically shout над-и-гра-я (*nad-i-gra-ya*), под-у-ча (*pod-u-cha*), etc. Despite that in normal speech these are not the syllables in these words, it is possible, nevertheless, to divide the words in this way. Clearly, the different treatment of the prefixes and the suffixes in the traditional Bulgarian hyphenation was not an arbitrary decision but it had to do with something different about the phonology of the prefixes and the suffixes.

The glottal stop (ʔ) is this different thing. Many languages use the glottal stop as a regular consonant. In the Bulgarian language it is not phoneme,¹ however it is readily inserted at the beginning of words starting with a vowel. For example, if we try to pronounce the word уча (*ucha* 'learn'), then there are good chances that in reality we will pronounce ʔуча (*ʔucha*). Now, notice that there are words starting with a prefix or with a root, but there are no words starting with a suffix. Therefore, there are plenty of cases when Bulgarians will add a glottal stop in front of a prefix or in front of a root starting with a vowel. So the Bulgarians are used to treat the prefixes об- (*ob-*) and ʔоб- (*ʔob-*) as identical, the roots уча (*ucha*) and ʔуча (*ʔucha*) as identical and so on. In fact, untrained Bulgarians won't even notice the difference. On the contrary, since a word never starts with a suffix, there are no cases when Bulgarians would insert a glottal stop before a suffix. If someone decides to pronounce a glottal stop before a suffix, this will be something very noticeable to any Bulgarian.

Therefore, we can formulate the following rule: in the traditional Bulgarian hyphenation, the morphology was subordinated to the syllables. We would never divide a word according to the morphology

¹One curiosity is the negative particle ʔъ-ʔъ (*ʔǎ-ʔǎ*) which is the only Bulgarian word using the glottal stop as a phoneme. Despite that this word is very common, I have no idea how to write it with Cyrillic letters.

if this creates impossible syllables. On the other hand, we were not insisting about using the most natural syllable division. If it was possible to preserve the morphology by insertion of a glottal stop, then we would do so in order to preserve the morphology. For example divisions such as над-играя (*nad-igraya*) and под-уча (*pod-ucha*) in reality were над-?играя (*nad-?igraya*) and под-?уча (*pod-?ucha*).

Should we respect the prefixes in the *Perfect Bulgarian Hyphenation*? Unfortunately, it seems impossible to formulate a clearly cut rule about this. Remember that our goal should be to make the reading easier by creating the right expectations in the reader while his eyes are moving from one line to the next. Cases as над-играя (*nad-igraya* 'outplay') and под-уча (*pod-ucha* 'to prompt') seem most clear because the most natural syllable divisions на-диграя (*na-digraya*) or по-дуча (*po-ducha*) create deceiving impression about the prefix. Somewhat less clear are cases as раз-ора (*raz-ora* 'plough up') whose derivation is productive in the modern language. To a more literate reader a prefix раз- (*raz-*) will provide more useful information than a meaningless syllable ра- (*ra-*). Therefore, to such a reader this hyphenation will be helpful. Is this going to be so with a less literate reader? I don't know.² And perhaps there is no need to preserve the prefix in cases when the derivation is from an obscure root (под-ема (*pod-ema* 'take on')) or when the root is clear but nevertheless the prefix is not felt as a prefix (раз-ум (*raz-um* 'intelligence')).

One clear counterexample is the word отявлен (*otyavlen* 'downright') where the morphology boundary and the syllable boundary coincide. Therefore, the preferable hyphenation of this word is от-явлен (*ot-yavlen*) even though the Cyrillic letter я (*ya*) becomes merged with the following vowel а (*a*) and creates the false impression of incorrect syllable division. Indeed, the letter я (*ya*) in this word signifies the semivowel й (*y*). If, on the other hand, the syllable boundary were о-тявлен (*o-tyavlen*), then the letter я (*ya*) would no longer signify a semivowel but a palatalisation of the preceding sound т (*t*).

The traditional Bulgarian hyphenation tried to respect the suffixes but only when this would create no conflict with the syllables. Should we do the same in the *Perfect Bulgarian Hyphenation*? There are three cases to consider.

First, it is not appropriate to follow the morphology when the suffix starts with a vowel. This would contradict the whole Bulgarian hyphenation tradition where the morphology has a subordinate role with respect to the syllables. For example хлеб-ар (*hleb-ar*) is unwarranted.³

Second, when a suffix starts with one consonant, for example -ка (*-ka*), then the morpheme boundary is a possible syllable boundary. Therefore, even if we disregard the morphology, we are not going to divide the suffix. The only thing we should watch out is not to divide the morpheme preceding the suffix. There is no need to have too many hyphenation possibilities in order to obtain good looking printed document. Therefore, since объект-ната (*obekt-nata* 'object (adjective)') is permitted both according to the morphology and to the syllables, then there is no need to use обек-тната (*obek-tnata*), especially considering that объект- (*obekt-*) at the end of line provides the reader with more useful information than обек- (*obek-*). Similarly, since the division агент-ка (*agent-ka* 'agent woman') is permitted both by the morphology and by the syllables, then there is no need to use аген-тка (*agen-tka*), especially considering that агент- (*agent-*) provides the reader with more useful information than аген- (*agen-*).

And third, there are suffixes starting with more than one consonant (-ски (*-ski*), -ство (*-stvo*)). The traditional Bulgarian hyphenation did not allow such suffixes to be divided.⁴ Nevertheless, I assert that in the *Perfect Bulgarian Hyphenation* it is permissible to divide these suffixes. In fact, it is not just permissible to do so, but it is also preferable to do so. When the eyes of the reader reach the end of line and he sees there, say, братс- (*brats-*), then he will know that there are very good chances that this is one of the words братски (*bratski* 'brotherly') or братство (*bratstvo* 'brotherhood') and the suffix is -ски (*-ski*) or -ство (*-stvo*). If, on the other hand, the reader sees at the end of line брат- (*brat-*) then he will know that брат (*brat*) is the root of the word, but there will be too many other possibilities for the word besides братски (*bratski*) and братство (*bratstvo*). While the hyphenation братс-ки (*brats-ki*) is not morphological, it does not deceive the expectations of the reader and it makes the reading easier because

²Even the illiterate people feel the prefixes intuitively. The current official hyphenation rules leave to the discretion of the writer whether to respect the prefix or not. I think this is the best possible decision about this issue.

³Remember that we are not permitted to insert a glottal stop before the suffix -ар (*-ar*).

⁴It seems before 1945 this was a mandatory rule and after 1945—only a recommendation.

it gives more clues to the reader about what follows on the next line.

With this I can conclude this section of the article. We saw that the case about the suffixes was clear and unambiguous, even if somewhat complex. Some things about the prefixes were more ambiguous and depended on the personal preferences. Fortunately, there is software using a smart line-breaking algorithm which is able to produce good results even when only few hyphenation possibilities are available. One such software is \TeX . With such software perfect results can be achieved when the hyphenation rules permit a word division only when it is compatible *both* with the prefix morphology and with the syllables. Therefore, when we use such software, both *paз-opa* (*raz-ora*) and *pa-зopa* (*ra-zora*) should be forbidden and the software will still be able to produce a good printed document.

5. The Bulgarian Syllables

Many things about how our brain processes the speech are still unknown. It seems that the audio signal is processed as a hierarchy of carefully arranged segments. For example the intonation helps us to divide the signal into sentences. The stress helps us to divide the sentence into words. That's why in so many languages the stress has a fixed position in the word (penultimate, ultimate, or word-initial). One interesting thing about the stress is that it does not exist as such in the audio. Instead, it is an illusion, created by our perception. Several different factors, such as the tone, the rhythm within the sentence, the loudness and the reduction of the vowels, are used together in a complex way in order to determine where the stress is. Even when the information provided by these factors is inconclusive, we may still perceive the stress in its most probable position.

On a lower level, just as the stress helps divide the sentence into words, the sonority within the audio signal helps us to divide the words into syllables. The real nature of the sonority is still unknown. When the sonority reaches a peak above a certain threshold (which depends on the language), then we perceive a syllable. The peak of the sonority is exactly at the *nucleus* of the syllable. The part of the syllable which is before the nucleus is called *onset* and the part after the nucleus is called *code*.

When the sonority reaches a peak which is below the threshold, then such a peak does not signal the existence of a syllable (Zec, 1995). Such peaks make the speech perception more difficult. That's why the languages try to eliminate such false peaks. This is known as the *sonority sequencing principle*. It says that the sounds within the onset have raising sonority while the sounds within the code have decreasing sonority.

It is said that the syllables are abstract phonological constituents without clear phonetic correlates (Ladefoged and Maddieson, 1996). Nevertheless, the syllables are not just fanciful artificial creations whose only purpose is to amuse the linguists. They correspond to the way our brain processes the speech. Because of this they can be the base of a very natural system for hyphenation. That's why they are used for this purpose in many languages, Bulgarian included.

Several Bulgarian grammar books agree that the following sonority scale is valid for Bulgarian:

voiceless obtrusive < voiced obtrusive < sonorant consonant < vowel

According to my investigations, it seems that Bulgarian respects the sonority sequencing principle more accurately than most other languages. The only exception to the above scale in the written language is due to the letter *в* (*v*) which is a voiced obtrusive but it can be used as if it were voiceless obtrusive. This exception is due to a spelling particularity of the Bulgarian language. Whenever the letter *в* (*v*) seemingly violates the sonority sequencing principle, in the spoken language the letter *в* (*v*) is read as *ф* (*f*) (which is a voiceless obtrusive). For example the word *отвсякъде* (*otvsyakǎde* 'all round') is read as *отфсякъде* (*otfsyakǎde*).⁵

I have found that the sonorant consonants in Bulgarian have their own sonority scale: $m < n < l < p < \check{y}$ ($m < n < l < r < y$). Only a few words such as *жанр* (*zhanr* 'genre') and *химн* (*himn* 'anthem') violate this scale. Such words are always loan-words and their pronunciation is somewhat problematic for the native Bulgarian speakers.

⁵Since no Primitive Slavonic word contained the phoneme *ф* (*f*), we can hypothesize that in the Primitive Slavonic language the consonants *ф* (*f*) and *в* (*v*) were two positional variants of a single phoneme.

From the sonority sequencing principle we can deduce the following two hyphenation rules. First, in a sequence MK where M is a consonant with higher sonority than K, we are not permitted to hyphenate before M (except when M is *v* and K is a voiceless consonant). And second, in a sequence KM where M is a consonant with higher sonority than K, we are not permitted to hyphenate after M.

In addition to the Sonority Sequencing Principle, the consonant clusters within the Bulgarian syllable adhere to the following principles:

1. Both in the onset and in the code, the labial and dorsal plosives precede the coronal plosives and affricates.
2. If the onset or the code contains two plosives or affricates, then there are no fricatives between them. Few words with the Latin root 'text' are exceptions: КОНТЕКСТ (*kontekst* 'context').
3. If the onset or the code contains two fricatives other than *v* (*v*), then there can be no plosives or affricates between them.
4. If the onset or the code contains two plosives or affricates, then they both have equal sonority (both are voiced, or both are voiceless).
5. If the onset or the code contains two fricatives other than *v* (*v*), then they both have equal sonority (both are voiced, or both are voiceless).
6. Neither the onset, nor the code may contain two labial plosives, or two coronal plosives or affricates or two dorsal plosives.
7. Neither the onset, nor the code may contain two equal consonants with the exception of *v* (*v*) (for example ВТВЪРДИ (*vtvardi* 'indurate')).⁶

From these seven properties we can deduce corresponding hyphenation rules. For example from the first property we deduce that in a consonant sequence where a coronal plosive or affricate T is followed by a labial or dorsal plosive K, we separate T from K. From the second property we deduce that in a sequence KBT where K and T are plosives or affricates and B is fricative, we separate K from T. Etc.

With so many prohibitive rules, a question arises: if we apply all these rules, aren't we going to eliminate too many hyphenation possibilities? The answer is no. All that these rules do is helping the software to determine more accurately the exact boundary between the syllables. It can be demonstrated that between any two consecutive syllables at least one separation point will be permitted.

6. Finding the Morphemes

How a computer can find the morphemes in a word? It turns out, there is no need to do this. At least not too often. We saw already one reason for this—there are cases when we have to ignore the morphology (remember хлеб-ар (*hleb-ar*)). And the second reason is the following: when the second morpheme starts with a consonant, then the morpheme boundary coincides with the syllable boundary. So we only have to discover the syllable boundary. In the previous section we saw how we can do this with sufficient precision.

The reason the morphology so often does not contradict the syllables is the following. First, every language has a tendency for simplification during its natural evolution. When a particular simplification concerns a single morpheme then it is easier for this simplification to actually happen in the language. However, when a simplification concerns the contact between two separate morphemes, then this simplification is more difficult and can actually happen only in the following two cases: 1. when it concerns unproductive and obscure morphemes or, 2. when it is a result of a regular phonological law in the

⁶Actually, the letter *v* (*v*) is not a real exception because in all such cases this letter denotes two different consonants—*v* (*v*) and *φ* (*f*). In the word ВТВЪРДИ (*vtvardi*) the first *v* (*v*) is pronounced as *φ* (*f*). Only in the Russian loan-word ВЗВОД (*vzvod* 'platoon') the two letters *v* (*v*) denote a repeating consonant *v* (*v*).

language. The case 1 should not concern us. As for case 2, it is rare because the Bulgarian orthography is largely morphological. This means that morphemes are written according to their pronunciation. However, Bulgarian orthography usually ignores the phonological changes that happen in the spoken language at the contact of two morphemes. In other words, case 2 of the above two happens rarely in the written Bulgarian language. Because of this, when the contact of two morphemes is at a consonant cluster, then the place of greatest complexity within the cluster is boundary both of the syllables and of the morphemes.

In order to discover the morphological hyphenation rules, first select an arbitrary morpheme. Then try to predict when it will have the potential to generate different hyphenation with respect to the hyphenation based solely on the syllables. For example when a prefix ends with a consonant, then we will be interested by cases when this prefix is followed by a vowel, like in the word *раз-ора* (*raz-ora*). This is so because in such cases the hyphen determined by the morphology will differ from the hyphen determined by the syllables. You will find that such cases are not numerous. It is possible without too much efforts to manually observe all potential cases and write hyphenation rules for each prefix.

A somewhat more complex is the situation with prefixes like *по-* (*po-*) and *под-* (*pod-*). It can be summarized by the following rules:

1. When a word starts with *по-* (*po-*) and the next letter is not *д* (*d*), then *по-* (*po-*) is likely a prefix.
2. When a word starts with *под-* (*pod-*) and the next letter is a consonant, then *под-* (*pod-*) is most likely a prefix.
3. When a word starts with *под-* (*pod-*) and the next letter is not a consonant, then *по-* (*po-*) is most likely a prefix.

We only need to describe the exceptions to the above rules and such exceptions are not numerous.

Let me give a complete example. For the prefixes *о-* (*o-*), *об-* (*ob-*) and *от-* (*ot-*) I have found the following rules:

prefix *о-* (*o-*) when the following letter is not *б* (*b*) nor *т* (*t*)⁷

Exceptions: *оазис* (*oazis*), *овц* (*ovc*), *овч* (*ovch*), *огн* (*ogn*), *окси* (*oksi*), *окт* (*okt*), *олтар* (*oltar*), *омлет* (*omlet*), *омни* (*omni*), *онбаш* (*onbash*), *ондул* (*ondul*), *онзи* (*onzi*), *онко* (*onko*), *онлайн* (*onlayn*), *онто* (*onto*), *опт* (*opt*), *опци* (*opci*), *опб* (*opb*), *орг* (*org*), *орд* (*ord*), *орк* (*ork*), *орл* (*orl*), *орн* (*orn*), *орт* (*ort*), *орф* (*orf*), *орх* (*orh*), *осман* (*osman*), *осмин* (*osmin*), *осмиц* (*osmic*), *осмич* (*osmich*), *осмо* (*osmo*), *осте* (*oste*), *остро* (*ostro*), *осци* (*osci*), *охва* (*ohva*), *охка* (*ohka*), *охна* (*ohna*).

prefix *об-* (*ob-*) when it is followed by a consonant

Exceptions when this is not *об-* (*ob-*) but *о-* (*o-*): *облаго* (*oblago*), *облаж* (*oblazh*), *обрем* (*obrem*), *обрул* (*obrul*), *обръс* (*obras*), *овдов* (*ovdov*), *овлад* (*ovlad*). Exception when this is neither *об-* (*ob-*), nor *о-* (*o-*): *общн* (*obshn*). Cases of *об-* (*ob-*) followed by a vowel: *обагн* (*obagn*), *обигр* (*obigr*), *обясн* (*obyasn*), *обобщ* (*obobsh*), *обозн* (*obozn*), *обозр* (*obozr*), *обосн* (*obosn*), *обособ* (*obosob*), *обузд* (*obuzd*), *обусл* (*obusl*).

prefix *от-* (*ot-*) when it is followed by a consonant

Cases of *от-* (*ot-*) followed by a vowel: *отив* (*otiv*), *отид* (*otid*), *отуч* (*otuch*).

7. Implementation

The author has implemented a shell script⁸ which generates Bulgarian hyphenation patterns in the form expected by \TeX . The output of this script is about to be used by \TeX , LibreOffice, OpenOffice and the browser Mozilla. The script is configurable—the user chooses whether or not to use the morphology,

⁷Despite that the Bulgarian hyphenation rules do not permit lone letters, we have to discover this prefix anyway. This is so because we have to ensure the root is not divided in the vicinity of the prefix.

⁸<https://sourceforge.net/p/bgoffice/code/HEAD/tree/trunk/hyph-bg/hyph-bg.sh>

whether or not to use only good syllable divisions and which version of the hyphenation rules published by the Institute of Bulgarian language to use (1945, 1983 or 2012).

Some statistics follow. When the script is used to generate patterns strictly adhering to the rules published by the Institute for Bulgarian Language (no detection of the syllables and no morphology), then it will output 1676 patterns. If we chose to detect the syllables, then we will have 5798 patterns. And if, in addition, we chose to use also the morphology, then we will have 6886 patterns.

Any computer implementation of a morphology based hyphenation will make mistakes. According to a test with 303 randomly chosen words we have the following figures (after the sign \pm is the expected deviation of the estimation): in $3.3\% \pm 1.0$ of the words there is a hyphenation which contradicts the prefix morphology, such as *несп-равям* (*nesp-ravyam* 'failure to do s.th.') and $2.9\% \pm 1.0$ of the words are badly hyphenated compound words, such as *самок-ритичен* (*samok-ritichen* 'critical to o.s.').

No words were found where the hyphenation contradicts both the morphology and the syllables. In order to find a more precise estimate of this worst case, an additional test with 122 words was run, using only words whose hyphenation with and without the morphology is not identical (there are about $10.2\% \pm 0.1$ such words). Again no such cases were found. Based on both tests, we can expect that only 0.06% of the words are hyphenated in a really bad way.

8. Debate

Now, some people might ask: is the hyphenation that important to justify all the big efforts to implement them in software? And to this I give the following response: but are the efforts really that big? We have to develop good hyphenation rules only once and then thousands can use them for years to come. The rules I have developed are part of the standard distributions of \TeX , LibreOffice and Firefox, so all users will benefit for free and with no efforts.

Others will ask: OK, maybe it is not difficult to implement the rules in software. But clearly, these rules are too complex to be used by people. Well, when people use computers, how often do they hyphenate the words by themselves? Seldom. We live in the 21st century! In these days hyphenation is done by computers, not by people. Only in handwriting people still hyphenate themselves. Are the people going to use simplified hyphenation in their handwriting? Of course. Is there a problem?

But the Institute for Bulgarian Language has published *the official* hyphenation rules. Shouldn't we follow these rules exactly instead of inventing our own? Well, we do follow the official rules. But do we have to hyphenate *изг-рев* (*izg-rev* 'sunrise') only because the official rules say this is OK? No, because the official rules say *из-грев* (*iz-grev*) is also OK and we like the second option more. Thanks be to God for after 2012 abominations like *селскос-топански* (*selskos-topanski* 'agric-ultural') are no longer compulsory.

9. Conclusion

Since I don't have a list of the Bulgarian compound words, the morphological hyphenation rules I have developed are not concerned with the morphology of such words. Because of this, the Bulgarian computer users are still coerced to accept crazy things like *селскос-топански* (*selskos-topanski*). Let us hope that in the future some good person with love for the Bulgarian language will make such a list. Then we all will benefit.

Fortunately, it wasn't that difficult to develop morphological rules about the prefixes. These rules make very few errors. I haven't started with the suffixes yet but I hope they won't be difficult either.

The development of the rules about the Bulgarian syllables has given me so much fun! I had to dive deep into the wonders of the Bulgarian phonology. So many questions and kindling curiosity! Why the coronal consonants follow the labial and the dorsal consonants? Does this happen only by accident or there is a more significant reason?

In the battle between the syllables and the morphology, each pushing its own principles, we found that the syllables were victorious. But they were a generous victor who leaved to the morphology quite a lot of governing rights. To save ourselves from problems we will have to reckon with both of them.

Gratitude is due to all who have worked before me in the area of hyphenation.

References

- Andreychin, L. (1945). *Pravopisen rechnik na bulgarskiya knizhoven ezik*. Sofia, Hemus.
- Belogay, E. (1988). Algoritam za avtomatichno prenasnyane na dumi. *Kompyutar za vas*, 3:12–14.
- Georgieva, E. et al. (1983). *Pravopisen rechnik na savremenniya bulgarski knizhoven ezik*. Sofia, BAN.
- Hadzhov, I. and Minkov, T. (1945). *Pravopisen i pravogovoren narachnik*. Sofia, Bulgarska kniga.
- Koeva, S. (1999). Pravila za prenasnyane na chasti na dumata na nov red. *Bulgarski ezik*, 1:84–86.
- Ladefoged, P. and Maddieson, I. (1996). *The sounds of the world's languages*. Oxford, Blackwell.
- Liang, F. M. (1983). *Word Hy-phen-a-tion by Com-put-er (Hyphenation, Computer)*. Ph.D. thesis, Stanford University, Stanford, CA, USA. AAI8329742.
- Murdarov, V. et al. (2012). *Oficialen pravopisen rechnik na bulgarskiya ezik*. Sofia, Prosveta.
- Noncheva, V. (1988). Algoritam za avtomatichno prenasnyane na dumi v bulgarskiya ezik. In *Matematika i matematischesko obrazovanie. Sb. dokladi na 17 PK na SMB*, pages 479–482. Sofia, Izd. na BAN.
- Pashov, P. (1989). *Prakticheska bulgarska gramatika*. Sofia, Narodna prosveta.
- Stoyanov, S. (1993). *Gramatika na bulgarskiya knizhoven ezik*. Sofia, Universitetsko izdatelstvo “Sv. Kliment Ohridski”.
- Topalov, A. (1995). *Algoritmi i programi v tekstoobrabotkata*. Master's thesis, Sofia University, Faculty of Mathematics and Informatics. <http://www.mind-print.com/diploma/>.
- Vasilev, V. (1997). *Ultimativniyat TeX. Udovolstvieto da pravim predpechatna podgotovka sami*. Sofia, Intela.
- Zec, D. (1995). Sonority constraints on syllable structure. *Phonology*, 12:85–129.

Russian Bridging Anaphora corpus

Anna Roitberg

IMPB RAS

Branch of KIAM RAS, Russia

HSE RSU, Russia

cvi@yandex.ru

Denis Khachko

IMPB RAS

Branch of KIAM RAS, Russia

mordol@lpm.org.ru

Abstract

In this paper, we present a bridging anaphora corpus for Russian, introduce a syntactic approach for bridging annotation and discuss the difference between the syntactic and semantic approaches. We also discuss some special aspects of bridging annotation for Russian and other languages where definite nominal groups are not marked so frequently as e.g. in Romance or Germanic languages. In the end we list the main cases of annotator disagreement.

1. Introduction

Anaphoric links are very important for text cohesion. In 1975, Clark (Clark, 1975) contrasted direct anaphora and indirect anaphora (bridging). The term *direct anaphora* (1) is used for cases where anaphorically linked entities are coreferent. In the case of an *indirect anaphora* (2), anaphorically linked entities are not coreferent, but associated reflecting more complicated semantic relations.

(1) *I looked at his car yesterday. A really old **vehicle**.*

(2) *I looked at his car yesterday. **The door** was rusty.*

In researches of the direct anaphora, the terms *anaphoric element* (for vehicle in (1)) and *antecedent* (for car in (2)) are usually used. In the bridging anaphora researches, the terms *bridging element* (instead of anaphoric element) and *anchor* (instead of antecedent) are more common.

Bridging anaphora involves a very wide spectrum of semantic relations from the part-whole relations to relations between different arguments accompanying a single predicate. So all studies on bridging, as we know them, limit the number of bridging relations which they work with. The most common way to constrain the amount of bridging types is to consider several types of semantic relations, the most popular of which are part-whole and set-subset relations. This approach is used in inspiring the Poesio's projects: GNOME corpus (Poesio, 2000; Poesio et al., 2004) and ARRAU for English (Poesio and Artstein, 2008); in the second edition of the ARRAU corpus, the set of bridging relations became wider but it is still limited on the semantic ground. The same approach can be found in the CESS-ECCE corpus for Spanish (Recasens et al., 2007) and AnCora for (Recasens, 2008) Catalan, and PAROLE for French (Gardent et al., 2003). A wide spectrum of semantic types of bridging relations is annotated in the Prague Dependency Treebank for Czech (Mikulová et al., 2017; Nedoluzhko and Mírovský, 2011). The semantic constraints are also used in recent researches: multilingual corpus for English, German and Russian (Grishina, 2016) and GUM corpus for English (Zeldes, 2017)

We call this approach *semantically oriented*.

The second approach appeared through development of computational methods in linguistics. No semantic constraints are used here. This approach is less popular, but it is used in (Hou et al., 2013) where very impressive results are shown. The state of the art system for bridging resolution (Hou et al., 2016) is based on this corpus.

The paper is structured as follows: Section 2 provides our syntactic oriented approach for bridging anaphora and introduces the term *genitive bridging*, Section 3 presents RuGenBridge corpus and annota-

tion scheme, Section 4 describes inter-annotator agreement, in Section 5 we discuss main cases of typical disagreement.

2. Bridging anaphora annotation approach for Russian

Bridging anaphora is a very complicated high-level phenomenon and the research is in its infancy. Due to this, there are no standard annotation schemes up to the present. The annotation scheme used usually corresponds appropriately to the study purposes.

Our main goal is to develop an automatic bridging recognition system (set of classifiers) based on machine learning techniques. First of all, we considered bridging relations between noun phrases (NP) and marked just heads of noun phrases. Recall that there are no articles in Russian, we could not focus solely on definite NPs as in studies for Romano-Germanic languages. Afterwards we decided to annotate only those features which could be useful for ongoing classifiers. That led to a decision not to use the semantic-oriented approach, because we cannot utilize and implement this knowledge. Russian WordNet and similar resources are not as developed as English analogues.

2.1. Genitive bridging

On that basis we concentrated on a new formal-oriented, syntactic approach. We decided to restrict the amount of bridging cases to one syntactic construction, more specifically, the genitive construction. The genitive construction $N+N_{gen}$ is very common in Russian, it typically marks possessive relations (in a broad sense). So it is associated with (but not limited to) such semantic relations as *item – possessor*, *part – whole* etc.

Therefore, we annotated the cases of bridging anaphora where the bridging element and anchor could form a grammatical genitive construction as in the following:

- (3) *Ja kupil telefon, no knopki okazalis' slishkom malen'kimi.*
'I bought the phone, but the buttons turned out to be too small'

On the one hand, in the example above, the words that mean “phone” and “buttons” are anaphorically linked: there are not just some buttons but specifically the buttons of the previously mentioned phone. On the other hand, in Russian the anchor “phone” and the bridging element “buttons” can form a grammatical genitive construction bridging *element + anchor.Gen*: “*knopki telefona.Gen*” ‘*the buttons of the phone*’. We called this kind of bridging relations “*genitive bridging*”.

In our corpus we annotated only cases of genitive bridging.

3. Corpus RuGenBridge

Our corpus materials were short news texts from online news agencies. Short means 100 – 200 words. We chose such short texts due to the complexity of this phenomenon: annotators make more mistakes in long texts, because of the difficulty of keeping in mind discourse relations over a large distance.

At the time of writing, we annotated 339 texts or 61076 tokens, and tagged 609 genitive bridging pairs.

All bridging cases were manually annotated using the BRAT tool¹. Parts of speech and syntactic links were annotated automatically, using FreeLing² and MaltParcer³ (Nivre et al., 2006), correspondingly.

On the engineering side, our corpus is a SQL database, which consists of 3 main tables: 1) Table of texts; 2) Tables of lemmas and 3) Tables of relations

3.1. Boundary markables

We postulate a principle of minimum possible size for markables. Where possible, we mark single nouns – the heads of the corresponding noun phrase. In “the smallest house in the lane” only a “house” will be

¹<http://brat.nlplab.org>

²<http://nlp.lsi.upc.edu/freeling/>

³www.maltparser.org

marked. In the case of having an anchor (or a bridging element) as a named entity, we mark all the entity, so in “Cherry Tree Lane” we mark “Cherry Tree Lane”; the same for names of persons, organizations, geographic names etc.

3.2. Semantic labels

Despite not using the semantic approach to corpus annotation, we can still use some semantic labels for anchors and bridging elements to mark the most popular semantic types which could be relevant for future work.

The set of labels is given below.

1. Geo – for proper names of geographic objects (Brazil, Indian Ocean, Grand Canyon). Compare (4) and (5):

(4) *The government of Moscow.Geo is continuing to discuss transportation.*

(5) *The government of the city is continuing to discuss transportation.*

2. ORG – for proper and common names, refers to official organisations, public institutions etc.: government, Russian Orthodox Church, BBC. We take into account the contextual meaning of a noun phrase. Compare (6) and (7):

(6) *BBC World Service.ORG has announced the extension of the agreement...*

(7) *She used to listen to the BBC especially news programs...*

3. POST – job titles: president, coach, cardinal, priest, dean

(8) *FC Barcelona.ORG has announced that the coach.POST was dismissed.*

The total number of semantic labels in RuGenBridge corpus is shown in Table 1.

Semantic label	Anchor	Bridging-element	Total
GEO	148	9	157
ORG	11	24	35
POST	22	-	22

Table 1: Semantic labels in RuGenBridge

These types of semantic labels were chosen in view of the fact that the lists of such lexical groups can be extracted from dictionaries and ontologies. This information can be used to construct a bridging anaphora recognition and resolution system, which was the main purpose of the project.

3.3. Bridging relations in RuGenBridge

Considering that we are using a new approach to bridging, we tried to analyze what types of bridging pairs (on semantic point of view) were annotated. We compared what kind of bridging relations become annotated by using each of the approaches. As a reference, we choose semantically oriented annotation scheme, using in Prague Dependency Treebank (PDT) – (Nedoluzhko and Mírovský, 2011). There are two advantages of this scheme for our project: 1) it is the one of the most developed semantic oriented scheme, 2) it was constructed for Czech – Russian’s relative language. There are 6 types of bridging relations are emphasized in PDT: (1) PART-WHOLE and WHOLE-PART, as e.g. in face – eyes), (2) SUBSET- SET and SET-SUBSET, as in a group of students – some students – a student), (3) the relation between an entity and a singular function on this entity (subtypes P-FUNCT and FUNCT-P, as in company – director) (4) the relation between coherence-relevant discourse opposites (type CONTRAST,

as in black flags – white flags), (5) non-coreferential explicit anaphoric relation (type ANAPH, as in first world war – at that time) and (6) further underspecified group REST consisting of six other bridging subtypes (e.g. relations between family members, event – argument, locality – inhabitant, etc.). We provided two experiments, fully described in (Roitberg and Nedoluzhko, 2016). In the first experiment, eight texts of the corpora were annotated with both schemes: genitive bridging and PDT. We annotated 69 bridging pairs using the PDT scheme, 22 pairs using the RuGenBridge scheme, but there were only 7 coincidence cases. During the second experiment, we added PDT annotation marks (for semantic type of bridging relations) for all genitive bridging pairs in 200 texts (more than a half of texts). All bridging relations using in PDT are listed in (Nedoluzhko et al., 2009). We analyzed what semantic types are most frequent among the genitive bridging pairs. The most frequent were: PART-WHOLE (WHOLE-PART), SET-SUB (SUB-SET), FUNCT-P (P-FUNCT); the last type is often used for government positions (parliament – speaker). Besides bridging relations we annotate coreference chains, but only for entities that were previously annotated as anchors or bridging elements.

We also analyzed which types of bridging relations were annotated with the genitive bridging approach are usually missed when semantic approach to annotation is used. We found out that just a half of genitive bridging pairs can be marked with any of semantic PDT labels. There are two main groups of cases, which cannot be classified as any of of PDT types of bridging: 1) the pairs that reflect text cohesion more than semantic relations. For example geographic names – something located there, like ‘Moscow – hospitals’; and 2) bridging relations between non-referential nouns, like ‘oil – barrel’; non-referential nouns were not marked in PDT on formal ground. Such syntactic oriented approach can be useful for those researches who study these types of bridging anaphora.

4. Evaluating the quality

4.1. Inter-annotator agreement

High-level annotation is a challenge. The higher-level phenomenon is less strictly described in theoretical models, so there are a lot of borderline cases which are difficult to annotate. Moreover, the discourse annotation requires close attention because an annotator has to keep in mind the text as a whole, not just a solitary word. This said, the inter-annotator agreement in high-level annotation is usually not as high as, for example, in part of speech tagging.

Corpus RuGenBridge was annotated by three annotators and a supervisor. The statistics for all annotations are shown in Table 2.

	Annotator 1	Annotator 2	Annotator 3
Anchors	167	419	663
Bridging elements	273	620	846
Bridging links	273	620	846

Table 2: Labels statistics for different annotations

The first annotator was inclined to miss some genitive bridging cases, whereas in contrast other annotators (especially Annotator 3) marked several false pairs.

In spite of visible differences, the level of agreement (F-measure) between Annotator 1 and Annotator 2 was sufficient in more detail see (Table 3).

An 1. Total links	An 2. Total links	True positive	False positive
273	620	147	473

Table 3: Inter-annotator agreement between Annotator 1 and Annotator 2.

While computing the Inter-annotator agreement, we considered one annotation as a gold standard and computed F-measure regarding this annotation.

An 1. Total links	An 3. Total links	True positive	False positive
273	846	105	741

Table 4: Inter-annotator agreement between Annotator 1 and Annotator 3.

We used F-measure for inter-annotator agreement in line with (Nedoluzhko and Mírovskỳ, 2013). The more widespread Cohen’s kappa can not be applied to such rare phenomenon as bridging anaphora. For rare phenomenon the number of no-no cases (close to *true negatives*) in confusion matrix is incomparably higher than the number of yes-yes cases (close to *true positives*) and yes-no cases, so Cohen’s kappa would always be in the neighborhood of 1.

As presented in Table 3, we considered Annotation 2 regarding Annotation 1. Notice that *True positive* – is the set of bridging pairs that are matched between Annotator 1 and Annotator 2; *true negative* – is the set of pairs which were labeled as bridging by Annotator 2 and in contrast were not labeled as bridging by Annotator 1.

On account of the data represented in Table 3, the inter-annotator agreement between Annotator 1 and Annotator 2 is at F-measure = 0.71

Unfortunately the level of agreement between Annotator 3 and other annotators was unacceptably low as shown in Table 4.

The F-measure for this pair of annotations is just F=0.37. Since this annotation contained multiple errors, we did not use this annotation in our results.

In the final release of the RuGenBridge Corpus the supervisor combined the annotations of Annotator 1 and Annotator 2 and removed all false pairs, which in truth were not the cases of genitive bridging.

4.2. Cases of typical disagreement

Bridging annotation requires both solid annotator’s experience and well-thought-out guidelines, but the main problem for annotators is to keep in mind the text and to concentrate on deciding if the noun in question has a bridging link to some anchor.

We summarized up the main types of inter-annotator agreement errors. In obvious way, there are three main groups of errors: 1) to omit a bridging pair, 2) to add a false pair, 3) to choose the wrong anchor for some bridging element. Beside errors, there are also some cases of insignificant differences between annotations.

We provide examples of each case in what follows.

4.2.1. Omitting of bridging-pairs

Omission of bridging-pairs is obviously the most common type of annotation errors, but happens more frequently where a bridging element and an anchor are linearly close to each other. The anaphoric link seems to be trivial in such cases, but it should be annotated on formal ground.

- (9) *Prezident v obrash’enií zayavil...*
 ‘The President announced in the **address**...’

It is worth mentioning that to miss bridging pairs, to miss bridging pairs at a long-distance (those with an anchor in the very beginning of the text and bridging element at the end of the text) bridging relations was the second most frequent type of errors of this sort.

4.2.2. Adding false pairs

Genitive bridging criteria is formal and “machine-friendly”, but in some situations it was difficult to follow this criteria, because there were some cases semantically close to genitive bridging relations. Sometimes such pairs were annotated by mistake. One of the most frequent cases was bridging relations between two geographic objects, where one is a part of another. In Russian, two geographic names can usually form grammatical genitive construction, when the head is a name of a country and the dependence is a name of some region of the country. In Russian, the dependence usually contains such general words

as *oblast'*, *kraj* means 'region'. Several expressions of that type can be used in genitive constructions as follows:

(Part_Geo) + (Whole_Geo).Gen

“Moskovskaja oblast” and “Rossijskaja Federacija” can form a grammatical genitive construction, see Example (10).

(10) *Moskovskaja oblast' Rossijskoj Federacii.Gen* ‘Moscow region of Russia federation’

However, even more expressions can not form grammatical genitive construction on formal ground even though they are semantically very close to previous ones. Example (11) below is ungrammatical.

(11) **Sibir' Rossijskoj Federacii.Gen*

4.2.3. Mismatches in coreference chains

In the RuGenBridge corpus annotation guideline it was mentioned that the annotator should choose the linearly closest preceding anchor. In several cases, the annotators missed the closest anchor and made a link to some other coreferential NP. We consider cases of this sort as insignificant, so such errors was ignored while computing inter-annotator agreement.

(12) (...) Tol'jatziazota, v sluchae esli ne soglashus' na ih uslovija po prodazhe predpriyatija (...) ih tsel' rejderskij zahvat predpriyatija (...) ne lehche li bylo by vykupit' dolyu **minoritarijev**.
‘Of’Tol'jatziazot’, If I do not accept their conditions for a business transfer (...) their goal is a asset-grabbing (...) was not it easy to buy out the (...) **minority interest**’

One of Annotators drew an arrow from “minority interest” to “business” and the second annotator connected the “minority interest” to “Tol'jatziazot” (the name of the company). “Tol'jatziazot” and “business” are coreferential expressions.

4.2.4. Comprehension disagreement errors

A minor proportion of errors was caused by different comprehension of texts as in Example bellow.

(13) *V Instagrame Papy Rimskogo pojavilas' fotografija Papy, obnimajush'ego dvuh devochek s sindromom Dauna s zhelto-goluboj lentoj v rukah* .
‘In Papa’s Instagram a photo appeared of Papa, hugging two girls with Down syndrome, holding yellow and blue ribbons in the **hands**.’

One annotator linked the bridging element “hands” with anchor “Papa”, while the other annotator connected “hands” with “girls”.

Importantly, in Russian a possessive pronoun before “hands” is not required, so there is a case of ambiguity.

It is interesting to note, that our automatic bridging recognition system marked highly likely both mentioned variants as bridging relations.

5. Conclusion

We have described a syntax-oriented annotation scheme used in the RuGenBridge corpus. The RuGenBridge corpus represents an inventory of bridging anaphora relations which are not limited to common semantic relations such as part-whole, set-subset etc. We have also shared an experience in bridging anaphora annotation. In line with our expectations, the development of the corpus reveals the complexity of discourse-level annotation, which leads to a lower level of inter-annotator agreement. To increase inter-annotator agreement, we consider training future annotators in discourse theory in general and especially in anaphora theory.

The RuGenBridge corpus can be used as a training and test data set for bridging anaphora recognition; see our pilot results in (Roitberg and Khachko, 2017). The corpus is available on request. The

supplementary materials on the project are available on GitHub repository ⁴.

References

- Clark, H. H. (1975). Bridging. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, pages 169–174. Association for Computational Linguistics.
- Gardent, C., Manuélian, H., and Kow, E. (2003). Which bridges for bridging definite descriptions. In *Proceedings of the EACL 2003 Workshop on Linguistically Interpreted Corpora*, pages 69–76.
- Grishina, Y. (2016). Experiments on bridging across languages and genres. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 7–15.
- Hou, Y., Markert, K., and Strube, M. (2013). Global Inference for Bridging Anaphora Resolution. In *HLT-NAACL*, pages 907–917.
- Hou, Y., Markert, K., and Strube, M. (2016). Unrestricted Bridging Resolution. *Computational Linguistics*, (Just Accepted):1–68.
- Mikulová, M., Mírovský, J., Nedoluzhko, A., Pajas, P., Štěpánek, J., and Hajič, J. (2017). PDTSC 2.0-Spoken Corpus with Rich Multi-layer Structural Annotation. In *International Conference on Text, Speech, and Dialogue*, pages 129–137. Springer.
- Nedoluzhko, A. and Mírovský, J. (2011). Annotating extended textual coreference and bridging relations in the Prague Dependency Treebank. *Annotation manual. Technical report*, (44).
- Nedoluzhko, A. and Mírovský, J. (2013). Annotators’ Certainty and Disagreements in Coreference and Bridging Annotation in Prague Dependency Treebank. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 236–243.
- Nedoluzhko, A., Mírovský, J., Ocelák, R., and Pergler, J. (2009). Extended coreferential relations and bridging anaphora in the Prague Dependency Treebank. In *Proceedings of the 7th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2009), Goa, India*, pages 1–16.
- Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219.
- Poesio, M. and Artstein, R. (2008). Anaphoric Annotation in the ARRAU Corpus. In *LREC*.
- Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. (2004). Learning to resolve bridging references. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 143. Association for Computational Linguistics.
- Poesio, M. (2000). Annotating a Corpus to Develop and Evaluate Discourse Entity Realization Algorithms: Issues and Preliminary Results. In *LREC*.
- Recasens, M., Martí, M. A., and Taulé, M. (2007). Text as scene: Discourse deixis and bridging relations. *Procesamiento del lenguaje natural*, 39:205–212.
- Recasens, M. (2008). Discourse deixis and coreference: Evidence from AnCora.
- Roitberg, A. and Khachko, D. (2017). Bridging Anaphora Resolution for the Russian Language. In *Proceeding of 23rd Conference on Computational Linguistics and Intellectual Technologies Dialogue-2017*.
- Roitberg, A. and Nedoluzhko, A. (2016). Bridging Corpus for Russian in comparison with Czech. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 59–66.
- Zeldes, A. (2017). The GUM corpus: creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

⁴<https://github.com/Anna-Roitberg/RuGenBridge>

Aspectual and temporal characteristics of the past active participles in Bulgarian – a corpus-based study

Ekaterina Tarpomanova
University of Sofia St. Kliment Ohridski
tochitsaaa@gmail.com

Abstract

The paper presents a corpus-based study of the past active participles in Bulgarian with respect of their aspectual and temporal characteristics. As this type of participles combine two morphological markers, a special attention is paid on their interaction in different tenses, moods and evidentials. The source of language material used for the study is the Bulgarian National Corpus. The paper is organized in terms of morphological oppositions, aspectual and temporal, analyzing the functions of the participles in compound verbal forms.

1. Introduction

In the modern Bulgarian there are five types of participles: present active, aorist active, imperfect active, past passive, and gerund. Being a verbal form, participles can be marked for tense, aspect and voice, but they also may share some of the categories of the adjective as gender, number and definiteness. However, their hybrid nature does not allow for the complete manifestation of the grammatical categories and especially with regard to the verbal categories participles are only partially marked with the respective grammatical meanings (GSBKE: 374).

The paper aims at studying the functions of the aorist and imperfect active participles by using the empirical data of the Bulgarian National Corpus. The aorist active participle is an old form that is found in all the Slavic languages. It is considered a formally, semantically and functionally stable form both in dialectal and standard varieties of Bulgarian. The imperfect active participle is an innovation in Bulgarian and a number of researchers share the opinion that its formation is connected to the grammaticalisation of the category of evidentiality. The study is organised in terms of morphological oppositions – aspectual and temporal, based on the respective characteristics of the participles. In such situations the speaker has to make a choice between morphologically marked forms according to his/her communicative intentions and the context that can enhance or restrict the usage of a certain form.

2. Research method

The Bulgarian National Corpus (BulNC) is used as a source of empiric language material being the largest electronic resource for Bulgarian (its monolingual part contains over 1,2 billion words). BulNC has been designed mainly for computational linguistic tasks focusing on volume and structure. Although representativeness and balance are not considered key features, the corpus covers the language production since 1945 up to now and the language varieties of different text types. The online search system and the linguistic annotation make it suitable for linguistic research too. For detailed description of BulNC, see Koeva et al. 2012.

Participles in BulNC are annotated as deverbal forms with several grammatical characteristics, for example *правел* {V PE T s q}: V = verb, PE = perfective, T = transitive, s = singular, q = imperfect participle. Theoretically the combination of two features – aspect (imperfective or perfective) and type of participle (past aorist or past imperfect) should provide all the grammatical information through the respective morphological markers for a correct annotation, but in fact there are many instances of

incorrect interpretation, especially concerning specific forms such as imperfect participles derived from perfective verbs. For that reason I chose three imperfective verbs representative for the three conjugations in Bulgarian, and their perfective counterparts¹: *пиша, напиша* ‘write’; *правя, направя* ‘do; make’; *казвам, кажа* ‘say’. The verbs are frequent and semantically neutral, so that the results of a search by word form allow for general conclusions about the types of participles under study.

3. Past active participles in Bulgarian: formation, meaning and usage

The aorist past participle is formed by adding the suffix *-l* to the aorist stem of an imperfective or a perfective verb:

пиша – *писал* ‘write, impf.’
напиша – *написал* ‘write, pf.’
правя – *правил* ‘do; make, impf.’
направя – *направил* ‘do; make, pf.’
казвам – *казвал* ‘say, impf.’
кажа – *казал* ‘say, pf.’

The aorist past participle denotes property of an action that is performed before a given interval of reference (GSBKE: 379; Nitsolova 2008: 434). It is used in the resultative tenses of indicative (perfectum, plusquamperfectum, futurum exactum, futurum exactum praeteriti), in the structure of indirect evidentials (renarrative, inferential and dubitative) and admirative, and in the Slavic type of the conditional mood.

The imperfect past participle is formed from the imperfect stem of an imperfective or a perfective verb and the suffix *-l*:

пиша – *пишел* ‘write, impf.’
напиша – *напишел* ‘write, pf.’
правя – *правел* ‘do; make, impf.’
направя – *направел* ‘do; make, pf.’
казвам – *казвал* ‘say, impf.’
кажа – *кажел* ‘say, pf.’

According to Nitsolova, the imperfect active participle denotes an action whose interval is larger than a present or a past interval of reference (Nitsolova 2008: 436). It can be used only in indirect evidential forms (renarrative, inferential or dubitative) and in admirative. Unlike the aorist participle, it cannot function as an adjective.

The 3rd conjugation verbs have only one stem for all the tenses, i.e. the present stem, and for that reason the aorist and the imperfect participles are homonymous.

Past active participles in Bulgarian are organised in a complicated system with two morphological markers: for aspect (imperfective vs. perfective) and for tense (aorist vs. imperfect). Their functioning can be analysed in terms of two oppositions: aspectual and temporal (as they are all active, the opposition by voice is not relevant).

4. Aspectual oppositions

4.1. Indicative

Perfect and pluperfect

писал vs. *написал*, *правил* vs. *направил*, *казвал* vs. *казал*

Participles display the common characteristics of the respective aspect, i.e. participles of imperfective verbs present the event as atelic, more often iterative, non-concrete (general) or processual², while the participles derived from perfective verbs view the event as telic, usually single and/or concrete. The examples of BulNC show that there are several typical contexts of each type of participle.

¹ The prevalent opinion for the aspectual oppositions in Bulgarian is that a basic imperfective verb (*пиша*) and a prefixed perfective verb (*напиша*) do not form an aspectual pair, but for the purpose of this study verbs are selected for their frequency and variety of forms.

² For the concrete aspectual meanings I use the classification of Valentin Stankov (Stankov 1980).

Imperfective: iterativity

The imperfective verbs and participles respectively are typically used for unbounded iterativity and habituality, while the perfective is associated with bounded iterativity. Iterativity is often enhanced by adverbials of the type ‘many times’, and habituality by adverbs and adverbials with the ‘always’.

(1) *Всичко това той го е казвал и преди безброй пъти.*

‘He has said that before, thousands of times.’

(2) *Да, тъкмо това беше правил винаги – носеше се по пързалката на течението.*

‘Yes, he had always done this – he was drifting on the stream.’

Perfective: bounded iterativity

In Bulgarian bounded iterativity is regularly expressed by perfective verbs, usually in a lexical context specifying the number of times the event is repeated. With respect to the system of participles, bounded iterations are connected with the aorist participle. Still, in the structure of the perfect tense this is not a central meaning of this type of participle. A possible explanation is that the bounded iterativity combines better with aorist than with perfect because the event is presented as localized in a past moment, which contradicts to the main meaning of the perfect. Another restricting factor is the extension of the scope of the inferential in the field of the perfect, especially in 2nd and 3rd person.

The language data in BulNC illustrate the clear preference for aorist instead of perfect with aorist participle: 16 instances of aorist vs. 1 instance of perfect for the verb *кажа* ‘say, pf.’ in 1st p. sg. In the examples below the usage of perfect in (4) should be interpreted as emphatic.

(3) *Три пъти казах “добър вечер”.*

‘I said “good evening” three times.’

(4) *Хиляди пъти съм казал, че ненавиждам боя...*

‘I have said thousands of times that I hate fight.’

Imperfective: general factuality

The imperfective participle is used when the event is viewed as a general fact, without any specifications of its properties (Stankov 1980). This is one of the typical meanings of the imperfective aspect, but it is also strongly connected with the perfect tense. A very frequent lexical context in interrogative sentences are adverbials with the meaning ‘ever’.

(5) *Писмото звучеше сякаш го бе писал той.*

‘The letter sounded as if he had written it.’

(6) *Да съм казвал някога, че планът е свършен?*

‘Have I ever said that the plan was perfect?’

General factuality is often expressed in negative context, and in such cases it can be enhanced by adverbials ‘never’, ‘at all’, etc.

(7) *Аз например никога не съм писал нещо криминално.*

‘As for me, I have never written detective stories.’

(8) *Никога по-рано не съм правил това!*

‘I have never done this before!’

Perfective: concrete factuality

According to Stankov (1980), the concrete factual meaning is the central particular meaning of the perfective that expresses a single complete event stated as a fact in the concrete circumstances of its realization. Among the past tenses it is connected mostly with the aorist, denoting a concrete and a completed event in the past, but it is compatible with the perfect too. As compared to the aorist, the perfect meaning can be more expressive or to put an emphasis on the event. In a sentence with a perfective verb its arguments describe explicitly the situation of the event realization.

(9) *Егон, не си го измислям. – Не съм казал това.*

‘Egon, I’m not making it up. – I didn’t say that.’

(10) *Направил съм това проследяващо устройство.*

‘I made this tracking device.’

Imperfective: process

Processuality is a central meaning of the imperfective aspect. To express a process, participles of imperfective verbs are more often used in pluperfect with a taxis function.

(11) *Разбира се, знаеше всичко това, докато беше писал текста.*

‘Of course, he knew all that while he was writing the text.’

Imperfective instead of perfective

A perfective reading of basic imperfective verbs (non-prefixed and non-suffixed) is inherited by the respective participles.

(12) *Казвай де, какво ти е писал?*

‘Come on, tell me what he wrote you.’

Futurum exactum and futurum exactum praeteriti

Due to their meaning both tenses more often comprise in their structure perfective participles. FE and FEP refer to an event whose result is situated before the completion of another event (FEP is mostly used in conditional sentences). The fact that the second event is completed generally implies the completion of the first event too, that is why those two tenses usually choose participles of perfective verbs.

(13) ... *тридесет минути след приемането на химикала, ще е казал на трътеите всичко, което Елиът иска да знае.*

‘... thirty minutes after consuming the substance, he will have told the drones everything Eliot wanted to know.’

(14) *Станеше ли то, за няколко месеца щеше да е направил кариера и то каква!*

‘If this happened, for a few months he would have made a career, and a great one!’

The combination *ще е* + participle of imperfective verb has usually a presumptive reading. In fact, all the examples of such combination found in BulNC are presumptives (130 results):

(15) *Някога, на младини, тя ще е била стройна и хубава.*

‘When she was young, she must have been slender and beautiful.’

(16) *Когато най-сетне се приготви, беше невероятен – така ще е блестял Харун ал Рашид на първата си сватба.*

‘When he finally got ready, he looked amazing – Harun al-Rashid must have shined like this at his first wedding.’

(17) *От това може да се съди, че в Букурещ тя ще е преболедувала доста сериозно.*

‘One may conclude that in Bucharest she must have been very sick.’

A few exceptions are found, for biaspectual verbs with perfective interpretation:

(18) *След около два часа вашата нервна система ще е асимилирала вече Зеко.*

‘In a couple of hours your nervous system will have assimilated Zeke.’

(19) *На практика, населението ще е гласувало за умерено, традиционно или поне реформистко правителство, а ще се установи режим на твърдата левица...*

‘In actual fact, the people will have voted for a moderate, traditionalist or at least reformist government, but a regime of the hard left will establish itself.’

However, out of the corpus examples are found in which FE form with an imperfective participle is used to express a general fact in the future situated before a future event. Therefore, despite of the corpus data, the usage of imperfective participle in the structure of FE and FEP is possible, although limited in terms of frequency.

(20) *Просто ще е правил секс, а стеснителността му ще си остане...*

‘He will just have had sex, but his shyness will remain the same.’

4.2. Conditional

писал vs. написал, правил vs. направил, казвал vs. казал

Aspectual opposition between past active participles is present in the Slavic type of the conditional mood formed by the auxiliary *бих* and the aorist participle of an imperfective or perfective verb. As the conditional forms are unambiguous, some statistical data may be obtained by a searching by word forms (2 and 3 p. sg. of the selected verbs).

imperfective	<i>би писал</i>	31	<i>би правил</i>	78	<i>би казвал</i>	12
perfective	<i>би написал</i>	59	<i>би направил</i>	141	<i>би казал</i>	1290
				2		

Table 1: Instances of the imperfective and perfective aorist participles in conditional.

The results presented in Table 1 show a clear predominance of the perfective verbs in conditional, except for the verbs *пиша* / *напиша* ‘write’ with a ratio of only 1:2 between imperfective and perfective. The conclusion is that conditionals combine better with telic events, while atelic are peripheral. This observation is apparent for the aspectual pair *казвам* / *кажа* ‘say’ where the imperfective verb is suffixed and cannot be used with a perfective meaning, unlike *пиша* ‘write’ and *правя* ‘do; make’.

Taking into account the usage of the selected verbs, the most frequent concrete meaning of the imperfective participles in conditional forms is general factual (21), while iterative (22), habitual (23) and processual (24) meanings are occasional. A very frequent context for the imperfective participles is a *what*-question – 41 instances of the 78 occurrences of the form *би правил* ‘would do’.

(21) *Съвсем други скокове би правил този тигър на свобода.*

‘This tiger would make quite different jumps if he was free.’

(22) *Ако не чакаха някаква облага, и говорещият истината би лъгал колкото лъжеца, и лъжецът би казвал истината, колкото нелъжещия.*

‘If they did not expect some benefit, the truth teller would lie as much as the liar and the liar would tell the truth as often as the truth teller.’

(23) *Какво би правил обикновено? – Нищо особено.*

‘What would he usually do? – Nothing special.’

(24) *Днес някой спяха от Аинтаб би казвал: ...*

‘Today, some spahi from Aintab would say: ...’

Perfective conditional forms refer to a concrete event.

(25) *Така би направил един обикновен гражданин.*

‘That’s what a common citizen would do.’

4.3. Evidentials

Evidential present and imperfect

пишел vs. *напишел*, *правел* vs. *направел*, *казвал* vs. *кажел*

The evidential present and imperfect formed with imperfective participle display the characteristics of the respective tenses of indicative. The most frequent aspectual meanings associated with these tenses are the following: processual (26), iterative (27), habitual (28) and general factual (29). The examples below illustrate the usage of the participles in renarrative.

(26) *Но Зайо Байо не **правел** нищо особено.*

‘But the Rabbit wasn’t doing anything special.’

(27) *Кажете им, че ей сега тръгвам – **казвал** той на пратениците, а това "ей сега" нямаше край.*

‘Tell them that I’m leaving right away – he used to say every time to the messengers, and this right away was endless.’

(28) *Ахав, великият миротворец на Вискос, често **казвал**: ...*

‘Ahav, the great peacemaker of Viskos, used to say: ...’

(29) *Не **правела** така.*

‘She never does that, she said.’

The imperfect participle of perfective verbs can be used in dependent clauses only, or in imperative and optative clauses, which corresponds to the usage of the perfective verbs in indicative. The dependent clauses are more often introduced by the conjunction *да* ‘to’, other subordinating conjunctions (*за да* ‘in order to’), relative pronouns and adverbs.

(30) *Който **кажел** една нова истина, вдигали му паметник.*

‘Whoever told a new truth, they raised him a monument.’

(31) *... триста пъти да **кажел** „Отче наш“ и триста пъти „Аве Мария“.*

‘He had to say three hundred times the Lord’s Prayer and three hundred times Ave Maria.’

(32) *Бащата запази Анри при себе си, за да го **откъснел** от влиянието на майката и да го **направел** добър католик.*

‘The father kept Henry for himself, so as to bar him from his mother’s influence and to raise him as a good catholic.’

Evidential perfect and pluperfect

писал vs. написал, правил vs. направил, казвал vs. казал

The evidential perfect and pluperfect expressed by a single form are formed with the past active participle of the auxiliary *съм* 'be', i.e. *бил*, and aorist participle of the lexical verb. The usage of the participles in the renarrative tenses is identical with their functioning in indicative and they are found in a similar lexical context.

The imperfective participles refer to repetitive events (33), general facts (34), in many cases in negative context.

(33) *Колко пъти ѝ бил казвал на тази патка, че краката му са вечно студени и ако не бъдат добре затоплени, той изобщо не може да заспи!*

'How many times he told this idiot that his feet are always cold and if they don't get heated up well, he can't fall asleep at all.'

(34) *От разменените приказки разбрах, че не се е мил от Пролетния празник. Никой не му бил казвал да го направи след смъртта на майка му.*

'I understood from what he said that he hadn't washed since the spring holiday. Nobody told him to do that after his mother's death.'

The perfective participles denote a concrete fact (35), a few examples are found with the particular meaning of bounded iterativity (36). In the majority of cases the combination of the evidential auxiliary *бил* + aorist participle with iterative meaning is a dubitative aorist.

(35) *Престояло цяла седмица в храма, защото никой не им бил казал къде си отседнал.*

'It remained a whole week in the temple, because nobody had told them where you had put up.'

(36) *100 пъти му бил казал...*

'He told him 100 times.'

5. Temporal opposition

The temporal opposition holds between the aorist and the imperfect participles, which is only possible within the evidentiality system, where the two types of participles are used to form the temporal structure of the category. Tenses are organized by pairs expressed by a single form: present and imperfect; perfect and pluperfect; future and *futurum praeteriti*; *futurum exactum* and *futurum exactum praeteriti*; aorist. Thus the temporal opposition imperfect vs. aorist is expressed by the imperfect and aorist participles, respectively. Due to the two participial paradigms the evidential temporal system can express all types of events and their relations as the indicative tenses.

5.1. Renarrative/inferential imperfect vs. aorist

пишел, напишел vs. писал, написал

правел, направел vs. правил, направил

казвал, кажел vs. казвал, казал

The imperfect denotes an event that is simultaneous to a past moment, while the aorist refers to a completed event in the past. The temporal relations are illustrated with two text excerpts in renarrative (37) and inferential (38), which are the evidentials that may be used in longer texts.

(37) *Щом свършил първият танц, Петер се наредил с дамата си горе на площадката до Краля на танца и щом онзи рипнел три стъпки над земята, Петер скачал четири. Направел ли онзи чудни, изящни стъпки, Петер започвал да усуква и върти краката си така, че хората, които го гледали, се захласвали от удоволствие и възторг.*

'When the first dance finished, Peter lined up himself and his lady on the stage next to the King of the Dance and when he jumped three feet from the floor, Peter jumped four. If he did those wondrous, elegant steps, Peter started to fling and twist his feet in such a way that people who looked at him were struck with delight and amazement.'

(38) *"Който и да е бил, трябва първо бавно и безшумно да е убил дежурния, след това – Ту Май, като е запушил устата на младия хан с ръка, докато го е събарял надолу." Бен се обърна. "Да, и е трябвало вратата да остане затворена, докато го е правел, иначе е щял да бъде видян от мъжете около масата." Затвори очи, видял всичко ясно. "Офицерът се е*

оттеглял, когато се е обърнал с лице към Брок, извадил е оръжието си, без да даде на Брок време да стане от стола.”

“Whoever he was, first he must have killed the guard slowly and quietly, and then Tu Mai by gagging the young khan with his hand while wrestling him down.” Ben turned around. “Yes, and the door must have remained closed while he was doing it, otherwise the men around the table would have seen him.” He closed his eyes and saw everything clearly. “The officer must have been withdrawing when he turned to face Brock and drew his weapon without giving Brock time to get up from his chair.”

Example (31) describes a competition in dancing between two characters, Peter and the King of the Dance. The story begins by two single completed actions expressed by perfective aorist participles denoting the renarrative aorist tense: *свършил* ‘finished’, *се наредил* ‘lined up’. The following sentences comprise repetitive events expressed by imperfective imperfect participles in the main clauses that refer to imperfect tense (*скачал* ‘jumped’, *започвал* ‘started’), and in the dependent time clauses two specific verbal forms occur – *рипнел* ‘jumped’ and *направел* ‘did’, which are imperfect participles derived from perfective verbs and correspond to a peculiar meaning of the perfective aspect when combined with imperfect tense to denote repetitive events through a single example (Maslov 1959: 232). The excerpt ends with two continuous actions (*гледали* ‘looked at’, *се захласвали* ‘were struck’) expressed by imperfective imperfect participles.

Similarly, in (32) imperfect and aorist participles are used in inferential forms to express temporal relations in a murder scene inferred by a character in the novel. In that excerpt the typical contrast between aorist and imperfect can be seen, the aorist referring to single and completed events in the past (*е запушил* ‘gagged’, *се е обърнал* ‘turned’, *извадил е* ‘drew’), and the imperfect denoting continuous and incompleted acts that serve as a background for the completed ones (*е събарял* ‘wrestling’, *е правел* ‘was doing’, *се е оттеглял* ‘withdrawing’).

6. Aspect, tense and adjectives

The aorist participles may have adjectival usage and in these cases the perfective stem is preferred. Nevertheless in particular contexts both perfective and imperfective participles may be used as adjectives inheriting the aspectual and the temporal characteristics of the respective participle.

(39) *Четящият впоследствие ще почувства душата на писалия.*

‘The reader will afterwards feel the soul of the writer.’

(40) „Часът на зеления прилив“ очевидно е някакво предварително определено време между *написалия* документа и онзи, който трябва да го прочете.

“The hour of the green flow” is obviously some time period between the writer of the document and the one who has to read it.’

7. Distribution

The general distribution of the past active participles of the verbs *правя* ‘do; make impf.’, *направя* ‘do; make pf.’, and *пиша* ‘write impf.’, *напиша* ‘write pf.’ without specification of the compound verb form is shown in Table 2. The verbs are chosen to illustrate the forms distribution with respect to their frequency and the possibility to compare all four participles.

	aorist		imperfect	
imperfective	<i>правил</i>	6 023	<i>правел</i>	1 121
perfective	<i>направил</i>	24 847	<i>направел</i>	43
imperfective	<i>писал</i>	3 163	<i>пишел</i>	385
perfective	<i>написал</i>	4 529	<i>напишел</i>	3

Table 2: Distribution of the past active participles

The number of occurrences may be analyzed in several viewpoints. Aorist participles are more frequent than imperfect participles as they may be used in perfect tenses of indicative, in conditional and in evidential tenses. The usage of the imperfect participles is limited to few tenses of the indirect evidentiality. With respect to the compatibility of the grammatical features aorist participles derive more often from perfective verbs and imperfect participles – from imperfective verbs. Taking into

account those two trends, the highest frequency of the perfective aorist participle is not unexpected, as well as the smallest number of occurrences of the perfective imperfect participle.

8. Conclusion

Past active participles in Bulgarian form a complex system combining aspectual and temporal characteristics. Their usage in different tenses, moods and evidentials depends on the compatibility of the respective grammatical meanings. Corpus-based studies outline the general tendencies of their usage, the specific contexts that require a given type of participle and the restrictions due to incompatible aspectual and temporal meaning. In general, participles in compound temporal forms cover all the central particular meanings of verb aspects in Bulgarian, thus creating a possibility to express aspectual opposition within perfect tenses, conditional mood and evidentiality. Matching the general trends in aspectual functions, perfective participles have homogenous meaning and usage, and, on the contrary, imperfective ones display much more diversity in their functions and none of their particular meanings can be pointed out as predominant. In terms of frequency, aorist participles prevail considerably above imperfect, the latter being restricted within the evidential system.

Acknowledgement

This research was supported by the project *The Balkan languages as an emanation of the ethnical and cultural community of the Balkans (verb typology)*, financed by the Scientific Research Fund at the Ministry of Education and Science, contract ДН 20/9/11.12.2017.

References

- GSBKE: *Gramatika na savremenniya balgarski knizhoven ezik. Tom II. Morfologiya*. Sofia: Bulgarian Academy of Sciences, 1983.
- Koeva, S., Stoyanova, I., Leseva, S., Dimitrova, T., Dekova, R., and Tarpomanova, E. (2012). The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, 0(1): 65–110.
- Maslov, Y. S. (1959). Glagolnii vid v sovremennom bolgarskom literaturnom iazyke (znachenie i upotreblenie). – *Voprosy grammatiki bolgarskogo literaturnogo iazyka*. Moscow, pp. 157 – 312.
- Nitsolova, R. (2008). *Balgarska gramatika. Morfologiya*. Sofia: University Press „Sv. Kliment Ohridski“.
- Stankov, V. (1980). *Glagolniiyat vid v balgarskiya knizhoven ezik*. Sofia: Nauka i izkustvo.

Unmatched Femininitives in a Corpus of Bulgarian and Ukrainian Parallel Texts

Olena Siruk

Institute of Mathematics
and Informatics
Bulgarian Academy of Sciences
olebosi@gmail.com

Ivan Derzhanski

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

Abstract

Femininitives are formed and used in all Slavic languages, but the productivity of their formation and the intensity of their use are not the same everywhere. They are often subject to various intralinguistic and extralinguistic restrictions. In this paper we present a study of femininitives based on a parallel Bulgarian–Ukrainian corpus, with a focus on those occasions on which a femininitive in one language corresponds to a masculine (rarely neuter) noun in the other. The experiment shows that Bulgarian uses femininitives with considerably greater regularity than Ukrainian does, and we discuss the semantic classes of nouns that fail to form femininitives most often and the effect of the source language in translated text and of the author’s and translator’s individual preferences.

1. Introduction

The late 19th and the early 20th centuries saw a great increase of the representation of women in various social roles outside the family as a result of diverse objective causes of historical development, among them the industrialisation of production in the leading countries of Europe and North America and female emancipation. In the course of the 20th century women gained access to education at all levels and to a broad spectrum of professions, the opportunity to participate in elections and to be economically independent. Such profound changes in public life and culture could not but find their mark in many languages. In both Bulgarian and Ukrainian, one such process has been the intensification of the use of the mechanism of formation of **femininitives**, especially terms for denoting women by trade, social rank and role and political views, following the older models for deriving feminine correlates of masculine nouns expressing nationality, place of residence or individual characteristics. This process is ongoing, because the question of gender equality has not yet found its definitive social and linguistic resolution, and particularly dynamic in periods of intense social shake-up, as in the early 21st century in Ukraine.

This paper presents a comparative bilingual study of Bulgarian and Ukrainian femininitives based on a parallel corpus, with a focus on the cases where only one of the matching sentences contains a femininitive. We are not aware of other similar corpus-based cross-linguistic studies on femininitives.

2. The Corpus

The working Bulgarian–Ukrainian parallel corpus is composed entirely of fiction (mostly novels, but also short stories), including both original Bulgarian and Ukrainian texts and translations from other languages. The material has been obtained from electronic libraries or from paper editions through scanning, optical character recognition and error correction by *ad hoc* software tools and by hand. (See Siruk and Derzhanski (2013) for more details on the general make-up of the corpus, and Siruk (2017) on

its use in several earlier projects.) The current version is made of ten sectors, each composed of texts with the same original language and counting approximately 800,000 words on the Bulgarian and 700,000 words on the Ukrainian side. This amounts to a total size close to 15 million words. Two sectors contain translations from Russian and two from English (because of the larger amount of text available); the remaining original languages are Bulgarian, French, German, Italian, Polish and Ukrainian. All texts are aligned at sentence level.

The parallel analysis of translations into closely related languages reveals challenges for translations that these languages both share and differ in. Among these challenges are culturally marked signs, by which scholars of translation studies (Nekryach and Chala, 2013: 8–9) understand such lexical units that may have equivalents in the target language and be understandable to readers of the translation but evoke a different train of association (a combination of socio-cultural and historical associations that a certain concept comes with for representatives of a particular culture at a particular historical stage) than the readers of the original get. Femininitives may fall under this definition as items having socio-cultural peculiarities, which gives rise to divergence between the original and the translation, as well as between parallel translations: the translator sacrifices the formal and factual precision of the translation in order to recreate the associations of the original text.

3. Femininitives in Bulgarian and Ukrainian

Both Bulgarian and Ukrainian have several productive mechanisms for deriving femininitives, some shared (and going back to their common ancestor), some not: *student-k-a* ‘female student’ is both Bulgarian and Ukrainian; Bulgarian *glez-l-a* and Ukrainian *kapryz-ulj-a* ‘fickle woman’ each contain morphology that is not found in the other language. However, in both languages masculine terms are also often used for women, both because of lexical gaps and because of certain stylistic limitations on the use of femininitives, in part due to foreign influence (mostly of Russian and, more recently, English).

This affects the two languages to different degrees. Whereas the Bulgarian femininitives are declared to have the same stylistic characteristics, as well as the same lexical meanings, as the masculine nouns from which they are derived (Stoyanov, 1983: 55) and the avoidance of their use to be at variance with normative grammar (ibid.: 103), traditional Ukrainian grammar places the emphasis on the fact that ‘it is far from every noun for a person of male sex that a term for a person of female sex can be formed from’ (Moisiyenko, 2013: 176). Which is true in general, but is too categorical: the nouns *faxivec* ‘specialist’, *naukovec* ‘scientist’, *službovec* ‘employee’ are given as examples but form femininitives in fact; *faxivčynja* and *naukovka* or *naukovycja* are used in colloquial and journalistic speech, *službovka* is registered by lexicography (SUM) and is present in our corpus as well. Attempts to classify femininitives as potential but unrealised risk meeting the opposition of usage: the examples *spikerka* ‘(female) speaker’ and *medijnycja* ‘woman from the mass media’ (Moisiyenko, 2013: 178) have already been ‘realised’, are used in the press and thrive on the Net. But examples given by Bulgarian grammarians have similar problems: some of (Stoyanov, 1983: 113)’s examples of masculine nouns that form no femininitives (*profesor* ‘professor’, *docent* ‘associate professor’, *ministăr* ‘minister’, and especially *šofjor* ‘driver’) can no longer be called that (if they could at the time of writing).

4. Unmatched femininitives

The corpus was searched for occurrences of nouns with one of the feminine suffixes *-an(a)*, *-ic(a)*, *-inj(a)*, *-k(a)*, *-l(a)*, *-uš(a)* on the Bulgarian side and of nouns marked in SUM as *žin. do* ‘fem. to’ on the Ukrainian side. The results were proofread and sentences with false femininitives (i. e., their homographs, such as Bulgarian *špionka* ‘spyhole’ not ‘female spy’, Ukrainian *cukernycja* ‘sugar bowl’ not ‘woman sweets maker/seller’, or *zemljanka* ‘dugout, earth house’ not ‘Terran woman’ in both languages; also individual word forms, as Ukrainian *korolevi* dat. sg. of *korol* ‘king’ or nom. pl. of *korolevyj* ‘royal’ not dat. sg. of *koroleva* ‘queen’) were filtered out.

In the Bulgarian texts 292 femininitives were counted that had Ukrainian masculine nouns as their counterparts. In the Ukrainian texts 57 femininitives were found to which masculine nouns correspond on the Bulgarian side. Table 1 presents their distribution by original language.

language of the original	Bg	De	En	Fr	It	Pl	Ru	Uk	total
Bg m : Uk f	12	4	18	4	8	3	5	3	57
Bg f : Uk m	20	16	92	26	12	3	112	11	292
total	32	20	110	30	20	6	117	14	349

Table 1: Number of unmatched femininives by original language

Bearing in mind that there was twice as much text translated from English as written originally in Bulgarian, one has to conclude that the texts by Bulgarian authors were startlingly rich in masculine nouns applied to women, for which the Ukrainian translators chose to substitute femininives. On the other hand, the occasions on which the Bulgarian translators were more eager to use femininives were mostly in the texts by English and Russian authors. Note that these are the two most ‘femininive-hostile’ languages of the eight, in spite of the great differences in their grammatical structure. Contrariwise, the translations from Polish, a ‘femininive-friendly’ Slavic language, show the least discrepancy.

The numbers given above do not include cases where the translator who used the masculine noun appears not to have had a woman in mind. For example, in (1) there is a chambermaid in the Bulgarian translation but a stableman in the Ukrainian one:

- (1) Bg: *Kogato se vǎrnax na slednata sutrin, mene veče me čakaše kamerierkata na mistǎr Kandi i tutaksi me otvede v stajata na svoja gospodar.*
 Uk: *Koly ja povernuvsjax na druhyj ranok, mene vže čekav bilja dverej pereljakanyj konjux mistera Kendi j vidrazu ž poviv mene v kimnatu svoho xazjajina.*
 En: *When I got back the next morning, I found Mr. Candy’s groom waiting in great alarm to take me to his master’s room.*
- (Wilkie Collins, *The Moonstone*)

In (2) the Bulgarian translator has understood the fellow-passenger to be a man and the Ukrainian has imagined a woman:

- (2) Bg: *Po pǎtjxa ot London do Hampšǎr misis Klemǎnts razbrala, če edin ot spǎtnicite ì dobre poznava okolnostite na Blakuotǎr [...]*
 Uk: *Dorohoju vid Londona do Hempšǎru z”yasuvalos’, ščo odna jixnja susidka po kupe čudovo znaje Blekvoter ta joho okolyci [...]*
 En: *On the journey from London to Hampshire Mrs. Clements discovered that one of their fellow-passengers was well acquainted with the neighbourhood of Blackwater [...]*
- (Wilkie Collins, *Woman in White*)

In (3) the Ukrainian translator has altered the gist of Tutmosis’ words from ‘if you knew Jewish girls, you wouldn’t try to ingratiate yourself with one of them by talking nonsense about Jews before her’ to ‘if you knew Jews, you’d realise that what you’re saying about them isn’t true’:

- (3) Bg: *Vižda se, če ti nikak ne poznavax evrejkite!...*
 Uk: *Odrazu vydno, ščo ty zovsim ne znaješ jevrejiv.*
 Pl: *Jak to widać, że nie znasz Żydówek!...*
 “How evident it is that thou knowest not Jewesses!”
- (Bolesław Prus, *The Priest and the Pharaoh* [English tr. by Alexander Glovatski])

Also excluded are pairs of sentences in which the different genders are caused by linguistic reasons, as when Baloo, who is a she-bear in the Bulgarian translation of *The Jungle Book* by Rudyard Kipling because the default word for ‘bear’ – *mečka* – is feminine, is accordingly referred to as *učitelka* ‘(female) teacher’ of the Law and chastises himself (that is, herself) as a *glupačka* ‘(female) fool’ for having let

Mowgli off with the Bandar-log; in Ukrainian *vedmid'* 'bear' is masculine, making Baloo a *včytel'* '(male) teacher' and a *duren'* '(male) fool'. But we include examples where no such reasons are in sight, as in (4), where the words for 'teacher' are the same (feminine in Bulgarian as in the German original, masculine in Ukrainian by default), although the word for 'passion for power' is neuter in both languages:

- (4) Bg: *Vlastoljubie: strašnata učitelka na velikoto prezrenie [...]*
 Uk: *Vlastoljubstvo — hriznyj učytel' velykoji znevahy [...]*
 De: *Herrschaft: die furchtbare Lehrerin der grossen Verachtung [...]*
 "Passion for power: the terrible teacher of great contempt [...]"
 (Friedrich Nietzsche, *Thus Spake Zarathustra* [English tr. by Thomas Common])

4.1. Cross-unmatched terms

The numbers given above confirm the fact that Ukrainian eschews femininitives more often than Bulgarian does. In those semantic fields where both do, however, the patterns can be complex.

Table 2 demonstrates several lexical items of the field 'friend, comrade', with the Bulgarian ones labelling the rows and the Ukrainian ones the columns, and each field showing the number of occasions in the corpus on which they match. The masculine and the feminine words are separated by double lines. (The following tables are organised in the same way.) Because of the way the experiment was set up, we did not count how many times a Bulgarian masculine noun corresponds to a Ukrainian masculine one.

	<i>tovaryš</i>	<i>pryjatel'</i>	<i>druh</i>	<i>tovaryška</i>	<i>pryjatel'ka</i>	<i>podruha</i>	<i>podružka</i>
<i>drugar</i>				4	1		
<i>prijatel</i>				1	1		
<i>družka</i>				3	1	14	6
<i>drugarka</i>	1		1	18	3	28	1
<i>prijatelka</i>	1	3	71	8	155	156	6

Table 2: Words for 'friend, comrade'

We see that, although the most frequent Ukrainian correspondences of Bulgarian *prijatelka* are the feminine nouns *pryjatel'ka* and *podruha*, the masculine *druh* also has a significant presence. At the same time Bulgarian *drugar* and *prijatel* can also be found to refer to women. The reason for this complexity is to be sought in the many associations that the concept of friendship comes with, including its numerous varieties (comradeship, friendship between women or across sexes, etc.).

The field 'enemy' is less ramified, but still not simple, in part because in Bulgarian no feminitive is formed from *vrag* 'enemy, foe', but as Table 3 shows, Ukrainian *voroh*, which has no such limitation, can also denote a woman.

	<i>voroh</i>	<i>vorohynja</i>	<i>neprijatel'ka</i>	<i>suprotyvnycja</i>	<i>nenavysnycja</i>
<i>vrag</i>		5	1	1	
<i>neprijatelka</i>	2			3	1
<i>protivnička</i>			1		

Table 3: Words for 'enemy'

For ‘witness’ both languages can be seen to use a masculine as well as a feminine word, though with different frequency. The middle column in Table 4 corresponds to the instrumental plural, which the two Ukrainian words share.

	<i>svidok</i>	<i>svidkamy</i>	<i>svidka</i>
<i>svidetel</i>	1	2	1
<i>svidetelka</i>	16	2	2
<i>očevidka</i>	1		

Table 4: Words for ‘witness’

We see that Bulgarian usually uses the feminitive whilst Ukrainian usually does not. The 1 in the top left cell of Table 4 reflects (5).

(5) Bg: *Tja čaka samo edna дума ot men, za da dojde v Jorkšir i prisăstvuvu v kačestvoto si na svidetel [...]*

Uk: *Vona žde lyše vidpovidi vid mene, ščob pojixaty v Jorkšir i buty prysutn’oju jak svidok [...]*

En: *She only waits a word of reply from me to make the journey to Yorkshire, and to be present as one of the witnesses [...]*

(Wilkie Collins, *The Moonstone*)

This is one of two registered occasions on which neither text uses a feminitive, although both might; the other is (6):

(6) Bg: *Stanala li e veče voljata svoja sobstvena izbavitelka i blagovestitel?*

Uk: *Xiba volja vže stala sobi spasytelem i visnykom radosti?*

De: *Wurde der Wille sich selber schon Erlöser und Freudebringer?*

“Has the Will become its own deliverer and joy-bringer?”

(Friedrich Nietzsche, *Thus Spake Zarathustra* [English tr. by Thomas Common])

What is remarkable about the last example is that, whereas ‘will’ is feminine in both target languages (though the German word is masculine), the Bulgarian translator has chosen to make ‘deliverer’ feminine and ‘joy-bringer’ masculine and in the Ukrainian text both are masculine.

Finally, the word for ‘teacher’ is 2 times masculine in Bulgarian and feminine in Ukrainian and 3 times the other way around, and the word for ‘disciple, pupil’ is 4 times masculine in Bulgarian and feminine in Ukrainian and 2 times the other way around.

4.2. Bulgarian masculine, Ukrainian feminitive

With ‘enemy’ having been mentioned already, the remaining words in this subsection show no room for generalisation, nor can any be called frequent, unless we count the 5 times on which Bulgarian *pomošnik* ‘helper, assistant’ corresponds to Ukrainian *pomičnyca*, the 3 times on which *beglec* ‘fugitive’ is used where the other side has *vtikačka*, and the 3 times when *maneken* ‘mannequin’ in a Bulgarian original (by Bogomil Rainov) is translated into Ukrainian as *manekennycja*. On all these occasions a feminitive could have been used in Bulgarian as well. The only two exceptions – actually one, occurring twice – are *doktor* used before a female doctor’s name, as is usual in Bulgarian, where the Ukrainian has *likarka* in the same position:

(7) Bg: *Doktor Anna Georgievna săšto e mogla da pronikne.*

Uk: *Likarka Hanna Heorhijivna tež mohla probratysja.*

Ru: *Doktor Anna Georgiyevna tože mogla probrat’sja.*

‘Doctor Anna Georgievna might also have got in.’

(Alexander Mirer, *Chief Noon*)

Note that the use of a generic masculine form of the title *doktor* in Bulgarian but a feminine in Ukrainian correlates with the fact that the vast majority of the Bulgarian surnames have an ending which indicates gender, whereas a significant portion of the Ukrainian surnames do not, so the frequent formula ‘*doktor* [initials] <surname>’ (ditto with other similar titles) tends to be more informative in Bulgarian. Absence of information often leads to failed stereotypical expectations and thence to misunderstandings, so one might see in this a strong stimulus for the use of feminines in Ukrainian, but it interacts with the conflicting requirements of the official and the colloquial style.

The other lexical items appear no more than twice and seem arbitrary.

4.3. Bulgarian feminine, Ukrainian masculine

Most words in this subsection denote professions. Some of the most frequent ones are in translations from Russian and reflect a general avoidance of feminines in that language (especially in the scientific, and by extension the science fictional, genre) which has been copied in the Ukrainian translations. Thus the Bulgarian feminine nouns *lekarka* ‘physician’ (20 times) and *astronavigatorka* ‘astronaut’ (7) correspond to masculine nouns in the Ukrainian text. So do *istorička* ‘historian’ (17), *bioložka* ‘biologist’ (7), *geoložka* ‘geologist’ (6) and *sekretarka* ‘secretary’ (5), whose feminine counterparts in Ukrainian (resp. *istorikynja*, *biolohynja*, *heolohynja*, *sekretarka*) are not accepted by all speakers. The same is true of Bulgarian *členka* ‘member’ (6), which (unlike the ones listed hereto) occurs mostly in texts outside the Russian sector. The Ukrainian correspondences *členka* and *členkynja* are actively used by the diaspora. Close to this semantic field is *poznavačka* ‘connoisseur’, Ukrainian *znavec* with no feminine in common use (*znavčinja* is rare at present).

An intriguing example which does not fit this paradigm is (7):

(8) Bg: *Bezdelnički, lůžkiny... kljukarki... shte kaža na majka-igumenka...*

Uk: *Darmoždky, brexunky... jazykodzvonny... nexaj on matušci-ihumeni skažu...*

‘Spongers, liars ... twaddlers ... just let me tell Mother Abbess’

(Mykhailo Kotsiubynsky, *Into the Sinful World*)

The word *jazykodzvin* seems to be the author’s neologism, from the set expression *dzvonyty jazykom* lit. ‘to ring with one’s tongue’, i. e., ‘to wag one’s tongue [as a bell’s clapper]’, so the presumed meaning is ‘idle talker’. Words with this pattern are technically harder to form a feminine from, but anyway the use of a masculine noun (after the hesitation pause marked by dots) makes the statement more abstract.

Terms denoting women by nationality or place of residence are conspicuously absent from both the preceding subsection’s material and this one’s. So are kinship terms and other words from the oldest layer of feminines.

4.4. Feminines with non-masculine counterparts

It happens that Bulgarian uses a neuter noun where Ukrainian has a feminine, especially due to a lexical gap. For example, there is no word for ‘female dwarf’ in Bulgarian, only *džudže* ‘dwarf’, which is neuter.

(9) Bg: *Izvikaše li majmunata, vseki păt izkreštjavaše i džudžeto, i negovijat glas beše mnogo po-životinski.*

Uk: *Koly mavpa vereščala, skrykuvala ščorazu j karlycja, i holos jiji buv šče menše sxožyj na lyuds’kyj.*

De: *Schrie der Affe, schrie jedesmal die Zwergerin mit, und ihre Stimme war tierischer.*

‘When the monkey screamed, the dwarf screamed too, and her voice was far more beast-like.’

(Heinrich Mann, *Young Henry of Navarre* [English tr. by Eric Sutton])

Or there are no single words in standard Bulgarian corresponding to Ukrainian *odynak* ‘single son’, *odynačka* ‘single daughter’, so two-word expressions have to be used, often based on *dete* (n.) ‘child’.

(10) Bg: *Edinstveno dete li e mis Havišam?*

Uk: *Mis Hevišem bula odynačkoju?*

En: *Miss Havisham was an only child?*

(Charles Dickens, *Great Expectations*)

Lastly, Bulgarian diminutives (especially from masculine nouns) are often neuter and gender-neutral.

- (11) Bg: *Eto ja, malkoto drugarče* [n], *čieto štastie trjabvaše da osiguri, dokolkoto možeše* [...]
 Uk: *Os' vona, joho malen'ka tovaryška* [f], *i vin doklade vsix zusyl', ščob vona bula jakomoha ščaslyviša* [...]
 En: *There she was, his little companion, to be made as happy as ever he could make her* [...]
 (John Galsworthy, *The Forsyte Saga*)

4.5. Gender-unmarked forms

In both Bulgarian and Ukrainian it is common (pun intended) for nouns or their forms to be able to belong to both the masculine and the feminine gender and refer to men or women. This may be because the lexical item is of the so-called common gender, as Bulgarian *rodnina* ‘relative’ or Ukrainian *susida* ‘neighbour’, or because a masculine and a feminine lexeme overlap in part of their paradigms, as was said about Ukrainian *svidok* (m.) and *svidka* (f.) ‘witness’ above. The overlap may be restricted to the written form, as in Ukrainian *hostěj* gen./acc. pl. of *hist* ‘(male) guest’ ~ *hóstej* ditto of *hostja* ‘(female) guest’; since it is written text we are dealing with, such syncretism is as good as complete.

We did not count such forms as part of this experiment, because their interpretation as feminine is a possibility at least, but we mention them here because they are significant as a potential factor of change. In Bulgarian, for example, masculine nouns with the suffix *-nik* have female correlates in *-nica* or *-nička*, and whether one or both are formed and used depends, largely idiosyncratically, on the noun: ‘deceased, late (woman)’ is always *pokojnica*, ‘woman worker’ always *rabotnička*, and ‘(female) fellow traveller’ can be *spātnica* or *spātnička*. But the plural forms of the feminines in *-nica* coincide with the plurals of the masculine nouns (*spātnici* is plural of *spātnik* as well as *spātnica*), which may have one (or both) of two effects: make speakers prefer the derivatives in *-nička* (first in the plural and then in the singular as well) or enhance the acceptability of the use of the same terms for men and women. Time, as well as separate studies, will show if this is the case.

4.6. Feminines referring to men

On very rare occasions a feminine noun may have a male referent. Two such involve strong censure:

- (12) Bg: *Ti si naj-lošijat meždu ricarite, a ne naj-dobrijat. Ti, vaša milost, si prosto razvratnik* [m], *kojto tǎrguva s devstvenostta si!*
 Uk: *Ty najhirsšyj sered rycariv, a ne najlipšyj, poljubovnycja* [f], *ščo prodaje cnotu.*
 Pl: *Najgorszyś między rycerstwem nie najlepszy, po prostu gamratka z waszmości, która cnota handluje!*
 “You are the worst among knights, not the best, — simply a *drab*, trading in virtue.”
 (Henryk Senkiewicz, *With Fire and Sword* [English tr. by Jeremiah Curtin])

—and a similar example (with a feminine noun in the Bulgarian translation and a masculine one in the Ukrainian) in Thomas Mann’s *Doctor Faustus*. And on two occasions the somewhat disdainful feminine noun *pehotinka* ‘infantryman, foot soldier’ appears in the Bulgarian texts of Erich Maria Remark’s novels, corresponding to *rjadovyj* ‘private’ and *soldat* ‘soldier’ in the Ukrainian. Either way one sees that the feminine gender is associated with lesser worth. The widespread feeling that feminines are best avoided scores another point here.

5. Conclusions

Being a phenomenon characteristic of all Slavic languages, feminines are present in Bulgarian as well as Ukrainian. Both languages have centuries-old but still active models for forming feminines, and they are very much alive in the colloquial style. Historical circumstances at the end of the 19th and at the start of the 20th century (industrialisation, female and social emancipation) have increased the demand for them, a process which continues, with varying intensity, to this day.

One would think that there is no obstacle to their functioning and development. What we see instead is a conflict between intralinguistic and extralinguistic factors. On one hand, the wealth of derivational mechanisms offers all possibilities for creating and using feminines (particularly in Ukrainian, with its

greater variety of feminine suffixes in common nouns), and the demand for them is undeniable (again, especially in Ukrainian, whose frequent gender-neutral surnames increase the need for alternative ways of expressing gender). On the other hand, in practice their derivation and employment is thwarted by the expansive influence of the geographically close Russian (constant upon Ukrainian and episodic upon Bulgarian) and the globally pervasive English (especially in the last two decades), in which feminines are severely restricted, be it by social opposition (in Russian to the point of banning the use of suffixal models analogous to the closely related Ukrainian ones) or structural traits (the levelling of the distinction between masculine and feminine being an unwavering tendency in English).

This contradiction is unambiguously reflected by the material of the parallel corpus: the pair 'Bulgarian feminine ~ Ukrainian masculine' is substantially more frequent than the pair 'Bulgarian masculine ~ Ukrainian feminine'. This despite the fact that suffixation as a typical derivational model for feminines has a larger number of formal manifestations in Ukrainian: Bulgarian has fewer feminine suffixes but applies them with greater regularity.

In translated texts the frequency of the use of feminines appears to depend on the source language. Translations from Russian to Ukrainian are considerably poorer in feminines than translations from Polish. Similarly, translations from German to Bulgarian are richer in feminines than translations from English (although the correlation is predictably weaker). This is a typical situation when there is a choice of translation variants but no conscious choice of translation strategies.

The employment of feminines may also be a marker of the author's or the translator's style. Characteristically, whilst in Ukrainian it is the enhanced use of feminines (as by authors P. Zahrebelny and V. Drozd and translator M. Lukash) that is marked, in Bulgarian it is their avoidance in typical contexts (e. g., by B. Rainov).

A question which remains open, due to the peculiarities of the parallel corpus, is the correlation between the use of feminines of various semantic classes and the genre and time of writing of the text. A comparison of the results of our investigations with observations made on large monolingual corpora of Bulgarian and Ukrainian may shed light on this matter.

References

- Moisiyenko, A., Ed. (2013). *Sučasna ukrajins'ka mova. Morfolohija*. Kyjiv: Znannja.
- Nekryach, T. and Chala, Yu. (2013). *Viktors'ka doba v ukrajins'komu xudožn'omu perekladi*. Kyjiv: Kondor.
- Siruk, O. B. (2017). Kul'turno markovani realiji v ukrajins'kyx ta bolhars'kyx paralel'nyx perekladax. In *Multikulturalizăm i mnogoezičie. Sbornik s dokladi ot Trinadesetite meždunarodni slavistični četenija – Sofija, 21–23 april 2016 g. Tom I. Lingvistika*. Veliko Tărnovo: Faber, pp. 649–659.
- Siruk, O. and Derzhanski, I. (2013). Linguistic Corpora as International Cultural Heritage: The Corpus of Bulgarian and Ukrainian Parallel Texts. In *Digital Presentation and Preservation of Cultural and Scientific Heritage (III/2013)*, pp. 91–98.
- Stoyanov, S., Ed.-in-Chief (1983). *Gramatika na săvremennija bălgarski knižoven ezik. Tom 2: Morfologija*. Sofia: Bulgarian Academy of Sciences.
- SUM (1970–1980). *Slovník ukrajins'koji movy v 11 tomach*. Kyjiv: Naukova dumka, 1970–1980, <http://sum.in.ua> (consulted on 4 February 2018).

The Bulgarian Summaries Corpus

Viktoriya Petrova

Bulgarian Academy of Sciences

v.k.petrova@abv.bg

Abstract

This article aims to present the Bulgarian Summaries Corpus, its advantages, its purpose and why it is necessary. It explains the selection of texts and process of summarization and the tool used, in addition of a quick overview of the current situation in Bulgaria. The paper also presents a general outline of the market needs, the use of this kind of tools and a short list of examples of a variety of corpora around the world both in language and field.

1. Introduction

Web content¹ has become a science with a list of new jobs² because of its growing importance. This has increased the volume of information available online and with it the need of its quick processing in a rapid and effective way. The necessity of extracting the most valuable parts of documents has also grown slightly, although papers and studies in this direction have been written for more than twenty years. Since more information is becoming available, more tools are needed to handle it (Mani, I. and Maybury, M.T. 1999). Some summarization-related technologies attracted substantial investment companies.

Universally, a “summary” is to be understood as a text that is produced from another bigger text, and that conveys the most important information from the original text. It should be no longer than half of the original.

Nowadays summarization is applied in multiple areas: from scientific articles to web pages content, to the creation of large and especially designed corpora. They all adopt different methods and techniques, such as deleting textual units that are considered unimportant for the main message (it often happens by using a discursive structure of text) or the structure trees that compute different segments of the text or sentence compression (consists in removing lexical units that are not important enough in the sentence to change or distort its main meaning).

2. Corpora around the world

A large variety of summarization corpora has been developed. Each of them stresses on a particular point of what the texts can be used for: length of the document, interpretation of the text (especially in

¹ Even though it is generally divided into textual, visual and aural, the focus in this paper is only on the first one and will be understood as such in the entire paper.

² Web content writer, Web content Manager, Web content Editor etc.

the area of politics), whether they are multi or monolingual, or are related to a particular area. Various examples are:

- The Japanese Text Summarization Corpus, especially developed to be able to judge the credibility of the information collected on the web. Another purpose is also the preparation of gold standard data to evaluate smaller sub-processes within the extraction and summary generation process, and the investigation of the summaries made by human summarizers (Nakano M., Shibuki H and al., 2010).
- The composed entirely on French Human Reference Corpus for Multi-Document Summarization and Sentence Compression whose purpose is the development of automatic methods for multi-document summarization including text, audio and video.
- The multi-document multilingual summarization corpus made for Arabic, English, Greek, Chinese, Romanian and others, whose aim was to evaluate a series of language-independent algorithms and the problem of summarizing news topics.
- TweetMotif, a tool created specifically for the search and topic summarization of Twitter messages.
- The Polish Summaries Corpus, created for the support of the tools for automated single-document summarization of texts in the Polish language.

3. Origins and purpose of the corpus

The Bulgarian Summaries Corpus was created under the guidance of the Institute for Bulgarian Language of the Bulgarian Academy of Science. It is the first corpus of its kind in the country and it is a part of the Bulgarian National Corpus³. The aim is not to be left behind the rest of the world, and to help in the application of the different linguistic areas and other research purposes. It is also expected to become a resource of the Bulgarian language on the internet, especially since it represents a peripheral language⁴.

4. Text selection

When choosing texts, the type of the corpus that has to be taken into consideration. In some cases randomly selected documents are acceptable, but in others they are not. Due to its particular nature, a variety of articles was selected for the Bulgarian Summary Corpus. They cover different journalistic domains and a large variety of styles. The main subjects vary between political analysis and newspaper articles, followed by health issues, diseases and their possible cures. The documents were subjected to an additional filter, where interviews and files with more than one text inside were deleted. In this way, every text was put on a different file.

The texts are divided in two main groups:

- Texts containing 1000 to 1999 words;
- Texts containing 2000 to 2999 words.

After the process of summarization is completed, two more files are created – one made up by sentences and another containing only the main closes from the file with the sentences. In this way the total number of files is 3. When there are both computer and human summarizations of texts, it is possible to compare the results from the machine and the people – it is a process similar in some extent

³ It was developed between 2001 and 2009, with over 240 000 text samples. Access to the corpus: <http://dcl.bas.bg/bulnc/en/>

⁴ According to the The Global Language System of de Swaan and its hierarchy of four levels (the peripheral, central, supercentral and hypercentral languages) the linguistic dimension of the world goes hand in hand with the political and economic aspects. The present global situation of languages is the product of prior conquest and domination and of ongoing relations of power and exchange. “Peripheral languages” are 98% of the world's languages and spoken by less than 10% of the world's population.

to the machine translation evaluation. The only difference here is that there are no translations, but purposely omitted parts of a text.

5. Summarization process

Normally, the summarization of a text may be language-dependent (when an algorithm is specifically designed for a certain language) or independent (mostly based on algorithms for which it is not important). Over the years many scientific papers describe different processes and techniques for the summarization of information and the different purposes of its use. Some examples are:

- The PageRank algorithm used by Google Search to rank websites in their search engine results, with its graph-based ranking algorithms;
- Classifier4J⁵ with its micro service “Summarizer”⁶. Its purpose is to “extract sentences from a text document, determine which are most important, and return them in a readable and structured way.”
- A linear-time algorithm for lexical chain computation. It makes lexical chains as an intermediate representation for automatic text summarization. By lexical chains is understood the cohesion among an arbitrary number of words that can be computed in a source document by grouping sets of words that are semantically related and have a sense flow.

Although it is undeniable that the current technological advancements are remarkable, much is still to be done in this area. For this reason is not recommended to give full credibility to algorithms. The same applies if they are language-dependent or independent. No matter the improvement, they are still not able to match human judgment, especially about the nuances that each word contains within itself. This is one of the reasons why automatic text summarization is a very difficult task.

This said, here is also another point of view to consider: when a human summarizes a piece of text, he or she usually reads it with the purpose of developing his or her understanding. Then, when he or she writes the summary, there is a tendency to highlight points that are related to the person’s own background. This implies rating as “important” information that other individuals might consider superfluous. In order to avoid this risk, when engaged in the process of summarizing, it is strongly recommended to give the task to more than one person.⁷

For the Bulgarian Summarization Corpus was chosen human summarization.

For the process of summarization was necessary the following:

- Texts containing 1000 to 1999 words had to be reduced respectively reduced by 40%, 20% and 10% of their initial volume when using entire sentences and by 32%, 16% and 8% when reducing by leaving only simple closes.
- Texts containing 2000 to 2999 words to be reduced by up to 24%, 12% and 6% of their initial volume when using entire sentences and by 20%, 10% and 5% when using simple closes.

⁵ A Java library designed to do text classification. <http://classifier4j.sourceforge.net/>

⁶ The tool may be found on the following website: <https://algorithmia.com/algorithms/nlp/Summarizer>

⁷ An example is the Polish Summaries Corpus, where manual summarization was conducted by 11 annotators. Texts were randomly assigned.

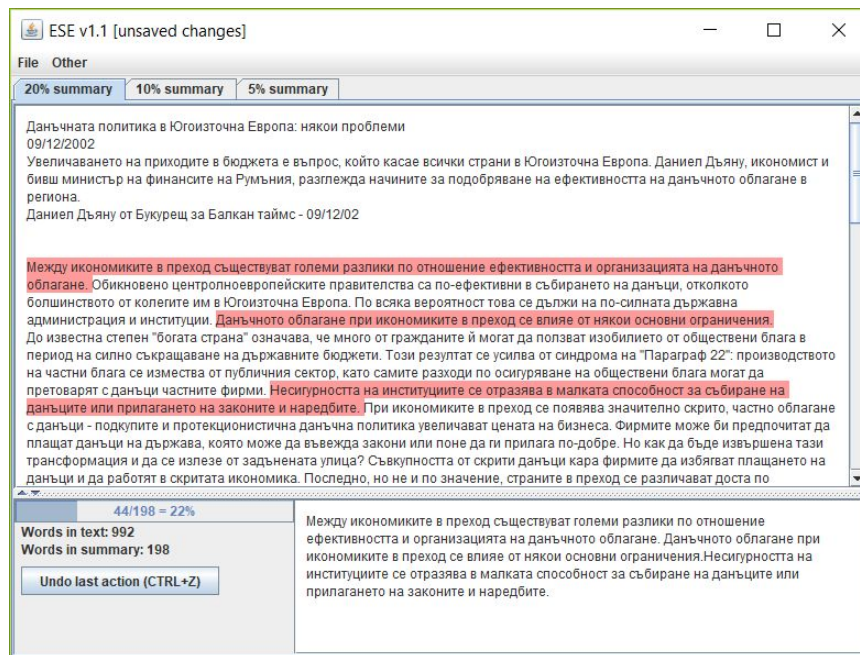


Figure 1: The interface of the ExtraSumAnnotator

For both groups, the reduction of simple clauses was done using the file containing the sentences selected in the first stage. All superfluous information in the texts such as author's name, dates or résumés as well as titles and abstracts had to be ignored in the process of summarizing the information in order to eliminate any possible interference with the results.

The tool ExtrSumAnnotator from the toolbox of the SummaryAnnotationTools⁸ was used. It was originally created for the Polish Corpus mentioned in paragraph 2. Figure 1 shows its user friendly interface with the different stages of the summarization process. The selection of percentages is performed manually by the user, while the word count is automatic. All selected information is highlighted in red and shown in the little window under the main text. This information is then transferred to the next tab, where the process is repeated again until the end.

5.1 Selection of sentences

The selection of sentences was done as follows for every text:

In the first stage of the work the whole text was read;

Then started the selection of entire sentences respecting the selected percentages. The changes were saved as a new file;

This new file was then opened again and from it were selected only the simple clauses. This second file was then saved, again separately.

As mentioned above, the total number of files at the end was 3.

It followed a specific pattern during the entire process – mostly the first and last sentences were selected for each paragraph of the text. In specific cases the middle part was preferred, but this procedure is to be considered an exception.

It is noteworthy that for the majority of articles were chosen the first and last sentences of every paragraph. Also, in certain cases the summarization process was extremely difficult due to the impossibility of reduction of the volume to the percentage required. This happened mainly because of the risk of loss of information.

⁸ They can be found on link of the Explore GitLab:
<http://git.nlp.ipipan.waw.pl/summarization/SummaryAnnotationTools>

The selection of simple clauses was a much easier process, since the message was almost always in the main sentence.

A practical example of the summarization process can be described as follows:

The main text contains 992 word (since it is in the first group, it has to be reduced respectively by 40%, 20% and 10% for entire sentences and by 32%, 16% and 8% for simple closes). In numbers this means that the reduction is:

For entire sentences:

992 words – 396 words (40%)

992 words – 198 words (20%)

992 words – 99 words (10%)

For simple clauses:

992 words – 317 words (32%)

992 words – 158 words (16%)

992 words – 79 words (8%)

Below parts of the text have been copy-pasted⁹.

Step 1 is the reduction by 40%:

“Между икономиките в преход съществуват големи разлики по отношение ефективността и организацията на данъчното облагане. Обикновено централноевропейските правителства са по-ефективни в събирането на данъци, отколкото болшинството от колегите им в Югоизточна Европа. По всяка вероятност това се дължи на по-силната държавна администрация и институции. Данъчното облагане при икономиките в преход се влияе от някои основни ограничения.

До известна степен "богата страна" означава, че много от гражданите ѝ могат да ползват изобилието от обществени блага в период на силно съкращаване на държавните бюджети. Този резултат се усилва от синдрома на "Параграф 22": производството на частни блага се измества от публичния сектор, като самите разходи по осигуряване на обществени блага могат да претоварят с данъци частните фирми. Несигурността на институциите се отразява в малката способност за събиране на данъците или прилагането на законите и наредбите. При икономиките в преход се появява значително скрито, частно облагане с данъци – подкупите и протекционистична данъчна политика увеличават цената на бизнеса. Фирмите може би предпочитат да плащат данъци на държава, която може да въвежда закони или поне да ги прилага по-добре. Но как да бъде извършена тази трансформация и да се излезе от задънената улица? Съвкупността от скрити данъци кара фирмите да избягват плащането на данъци и да работят в скритата икономика. Последно, но не и по значение, страните в преход се различават доста по способността си да печелят пари на чуждите капиталови пазари. [...]

Ниските данъци и опростени наредби са съществени за развитието на малките и средни предприятия (МСП). Но добрите идеи и предприемаческият дух може да не са достатъчни, когато има нужда от банково финансиране, а банките изискват трудни за получаване гаранции. Новите наредби на Банката за международни разплащания за банковото осигуряване на заеми може да удари сериозно малките и средно големи фирми, освен ако банките не намерят творчески начини за финансиране. Капиталовите пазари функционират лошо в Югоизточна Европа и често не могат да се използват като вариант за самофинансиране. Един начин за намаляване на трудностите за МСП е да се учредят

⁹ Since the working language is English and the text are in Bulgarian, this is to be considered just as an example and the will focus only on the percentages and how the selection works visually. Because of this, there are only the first and the last paragraphs. The underlined sentences represents the red marks in the ExtraSumAnnotator.

специални финансови институции, които да задоволяват техните нужди. Положителен знак е, че ЕБВР е измежду спонсорите на банките, занимаващи се с МСП, създадени в региона.”

Step 2 is a reduction of 20%:

“Данъчното облагане при икономиките в преход се влияе от някои основни ограничения.

Несигурността на институциите се отразява в малката способност за събиране на данъците или прилагането на законите и наредбите. При икономиките в преход се появява значително скрито, частно облагане с данъци – подкупите и протекционистична данъчна политика увеличават цената на бизнеса. Фирмите може би предпочитат да плащат данъци на държава, която може да въвежда закони или поне да ги прилага по-добре. Съвкупността от скрити данъци кара фирмите да избягват плащането на данъци и да работят в скритата икономика. Последно, но не и по значение, страните в преход се различават доста по способността си да печелят пари на чуждите капиталови пазари. [...]

Ниските данъци и опростени наредби са съществени за развитието на малките и средни предприятия (МСП). Капиталовите пазари функционират лошо в Югоизточна Европа и често не могат да се използват като вариант за самофинансиране. Един начин за намаляване на трудностите за МСП е да се учредят специални финансови институции, които да задоволяват техните нужди.”

Step 3 is a reduction by 10%:

“Несигурността на институциите се отразява в малката способност за събиране на данъците или прилагането на законите и наредбите. [...]

Ниските данъци и опростени наредби са съществени за развитието на малките и средни предприятия (МСП). Капиталовите пазари функционират лошо в Югоизточна Европа и често не могат да се използват като вариант за самофинансиране. Един начин за намаляване на трудностите за МСП е да се учредят специални финансови институции, които да задоволяват техните нужди.”

As mentioned above, the selected sentences are saved in a separate file that is used for the simple clauses. Here are also visible the aforementioned difficulties and the risk of loss of information, since many of the sentences cannot be reduced to simple clauses.

Step 1 is a reduction of 32%

“Между икономиките в преход съществуват големи разлики по отношение ефективността и организацията на данъчното облагане. Обикновено централноевропейските правителства са по-ефективни в събирането на данъци, отколкото болшинството от колегите им в Югоизточна Европа. По всяка вероятност това се дължи на по-силната държавна администрация и институции. Данъчното облагане при икономиките в преход се влияе от някои основни ограничения.

До известна степен "богата страна" означава, че много от гражданите ѝ могат да ползват изобилието от обществени блага в период на силно съкращаване на държавните бюджети. Този резултат се усилва от синдрома на "Параграф 22": производството на частни блага се измества от публичния сектор, като самите разходи по осигуряване на обществени блага могат да претоварят с данъци частните фирми. Несигурността на институциите се отразява в малката способност за събиране на данъците или прилагането на законите и наредбите. При икономиките в преход се появява значително скрито, частно облагане с данъци – подкупите и протекционистична данъчна политика увеличават цената на бизнеса. Фирмите може би предпочитат да плащат данъци на държава, която може да въвежда закони или поне да ги прилага по-добре. Но как да бъде извършена тази трансформация и да се излезе от задънената улица? Съвкупността от скрити данъци кара фирмите да избягват плащането на данъци и да работят в скритата икономика. Последно,

но не и по значение, страните в преход се различават доста по способността си да печелят пари на чуждите капиталови пазари. [...]

Ниските данъци и опростени наредби са съществени за развитието на малките и средни предприятия (МСП). Но добрите идеи и предприемаческият дух може да не са достатъчни, когато има нужда от банково финансиране, а банките изискват трудни за получаване гаранции. Новите наредби на Банката за международни разплащания за банковото осигуряване на заеми може да удари сериозно малките и средно големи фирми, освен ако банките не намерят творчески начини за финансиране. Капиталовите пазари функционират лошо в Югоизточна Европа и често не могат да се използват като вариант за самофинансиране. Един начин за намаляване на трудностите за МСП е да се учредят специални финансови институции, които да задоволяват техните нужди. Положителен знак е, че ЕБВР е измежду спонсорите на банките, занимаващи се с МСП, създадени в региона.”

Step 2 is a reduction of 16%.

“Данъчното облагане при икономиките в преход се влияе от някои основни ограничения.

Несигурността на институциите се отразява в малката способност за събиране на данъците. При икономиките в преход се появява значително скрито, частно облагане с данъци – подкупите и протекционистична данъчна политика увеличават цената на бизнеса. Фирмите може би предпочитат да плащат данъци на държава, която може да въвежда закони или поне да ги прилага по-добре. Съвкупността от скрити данъци кара фирмите да избягват плащането на данъци и да работят в скритата икономика. Последно, но не и по значение, страните в преход се различават доста по способността си да печелят пари на чуждите капиталови пазари. [...]

Ниските данъци и опростени наредби са съществени за развитието на малките и средни предприятия (МСП). Капиталовите пазари функционират лошо в Югоизточна Европа. Един начин за намаляване на трудностите за МСП е да се учредят специални финансови институции, които да задоволяват техните нужди.”

Step 3 is a reduction of 8%

“Несигурността на институциите се отразява в малката способност за събиране на данъците. [...]

Ниските данъци и опростени наредби са съществени за развитието на малките и средни предприятия (МСП). Капиталовите пазари функционират лошо в Югоизточна Европа и често не могат да се използват като вариант за самофинансиране. Един начин за намаляване на трудностите за МСП е да се учредят специални финансови институции, които да задоволяват техните нужди.”

6. Conclusion

Summarization procedures and techniques will increase and improve in the following years due to the constant rise of information available on the internet. Considering the current situation and the growing market needs, it is to be expected that instruments such as corpora will be very useful and appreciated by professionals and common users, especially when they belong to peripheral languages like Bulgarian.

As for the Bulgarian Summaries Corpus, hopefully it will grow with more texts, which will cover additional fields.

Hopefully, this work will contribute to the future development of the Bulgarian Summaries Corpus and will increase the popularization of this sort of instruments.

References

- Hristov, D. (2017). Automatic Text Summarization for the Bulgarian Language. Faculty of Mathematics and Informatics, Sofia University.
- Mani, I. and Maybury, M.T. (1999). *Advances in Automatic Text Summarization*. (MITRE Corporation) Cambridge, MA: The MIT Press
- Nakano, M., Shibuki, H., Miyazaki, R., Ishioroshi, M., Kaneko, K., Mori, T. (2010). *Construction of Text Summarization Corpus for the credibility of Information on the Web*. Graduate School of Environment and Information Sciences Yokohama National University
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.675.2386&rep=rep1&type=pdf>
- De Loupy, C., Guégan, M., Ayache, C., Seng, S., Torres Moreno, J-M. (2010). *A French Human Reference Corpus for Multi-Document Summarization and Sentence Compression*. Syllabs, Laboratoire Informatique d'Avignon (UAPV), Ecole Polytechnique de Montréal
http://www.lrec-conf.org/proceedings/lrec2010/pdf/919_Paper.pdf
- Li, L., Forascu, C., El-Haj, M., Giannakopoulos, G. (2013). *Multi-document multilingual summarization corpus preparation, Part 1: Arabic, English, Greek, Chinese, Romanian*. BUPT, China, UAIC, Romania, Lancaster Univ., UK, NCSR Demokritos, Greece.
<https://pdfs.semanticscholar.org/c4ba/4a4bc313a93850091b0b17ac27e2fbae569e.pdf>
- Ogrodniczuk, M., Kopéc, M.,(2014). *The Polish Summaries Corpus*. Institute of Computer Science, Polish Academy of Sciences
https://www.researchgate.net/publication/263087349_The_Polish_Summaries_Corpus
- O'Connor, B., Krieger, M., Ahn, D. (2010). *TweetMotif: Exploratory Search and Topic Summarization for Twitter*. Carnegie Mellon University, Meebo, Inc. Microsoft, Inc
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.478.5512&rep=rep1&type=pdf>
- Mihalcea, R. (2004). *Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization*. Department of Computer Science, University of North Texas
<http://www.aclweb.org/anthology/P04-3020>
- Silber, H. G., McCoy, K. F. (2002). *Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization*. University of Delaware
<http://www.aclweb.org/anthology/W00-1438>
- Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., Kochut, K., (2017). *Text Summarization Techniques: A Brief Survey*. Computer Science Department, University of Georgia, Institute of Bioinformatics, Department of Mathematics, Institute of Bioinformatics
http://cobweb.cs.uga.edu/~pouriyeh/Text_Summarization_Techniques_a_Brief_Survey.pdf
- “Web content” – https://en.wikipedia.org/wiki/Web_content
- “PageRank” – <https://en.wikipedia.org/wiki/PageRank>
- “Global language system” – https://en.wikipedia.org/wiki/Global_language_system

Ontologies for Natural Language Processing: Case of Russian

Natalia Loukachevitch

Research Computing Center of
Lomonosov Moscow State University
louk_nat@mail.ru

Boris Dobrov

Research Computing Center of
Lomonosov Moscow State University
dobrov_bv@mail.ru

Abstract

The paper describes the RuThes family of Russian thesauri intended for natural language processing and information retrieval applications. RuThes-like thesauri include, besides RuThes, Sociopolitical thesaurus, Security Thesaurus, and Ontology on Natural Sciences and Technologies. The RuThes format is based on three approaches for developing computer resources: Princeton WordNet, information-retrieval thesauri, and formal ontologies. The published version of RuThes thesaurus (RuThes-lite 2.0) became a basis for semi-automatic generation of RuWordNet, WordNet-like thesaurus for Russian. Currently researchers can use both RuThes-lite or RuWordNet and compare them in applications. Other RuThes-like resources are being prepared to publication.

1. Introduction

Term "ontology" is used in broader and narrower senses. In a broader sense, ontology is considered as an umbrella concept for various resources such as glossaries, thesauri, taxonomies, subject headings, or formal axiomatic systems. This understanding corresponds to the well-known definition of an ontology as "a formal specification of a shared conceptualization" (Guarino et al., 2009), (Gruber, 1995), (Studer et al., 1998), because all these resources represent some conceptualization of the external world.

In a narrower sense, ontology is a formal representation system of concepts defined using logical formalism to explain the meanings in a computable way (Pease, 2011). Such ontologies should be independent of any specific natural language (Nirenburg and Raskin, 2004), (Nirenburg and Wilks, 2001). The main author of well-known CYC ontology Doug Lenat wrote that taking the meanings of words into account can only confuse, the meanings of words divide the world ambiguously, and the division lines come from a variety of reasons: historical, physiological, etc. (Lenat et al., 1995).

In contrast to the above-mentioned approaches, the WordNet thesaurus is often mentioned as a linguistic, or lexical ontology, that is an ontology, whose concepts are mainly based on senses of existing lexical units, the terms of the subject field (Magnini and Speranza, 2002), (Veale and Hao, 2008). Linguistic ontologies cover most of the words of the language or a subject field, and at the same time they have an ontological structure represented in relations between the concepts. Synsets of WordNet are often considered as lexicalized concepts. Later, the WordNet structure was reproduced in various WordNet-like resources (wordnets) created for many languages (Azarowa, 2008), (Derwojedowa et al., 2008), (Koeva, 2010), (Kunze and Lemnitzer, 2010).

However, WordNet has been created as a lexical rather than ontological resource (Fellbaum, 1998)) in the framework of relational semantics (Miller et al., 1990). WordNet is mainly intended to describe lexical relations, which is quite different from the primary aim of ontologies to describe knowledge about the world, not about language. The WordNet structure was criticized from the ontological point of view (Guarino, 1998). Guarino and Welty (Guarino and Welty, 2009), developed the OntoClean approach for stricter description of relations and applied it to WordNet. Other authors (Wilks, 2009) suppose that NLP resources such as WordNet should not be subjects of such strict procedures, because of vagueness of their units, word senses.

Conventional information retrieval thesauri can also be considered as linguistic ontologies because they are based on real terms of a subject field (Will, 2012), (Clarke and Zeng, 2012). A term is defined as one or more words referring to a concept; a concept is considered as a unit of thought, regardless of the terms that express it (NISO, 2005). Contemporary standards for developing of information-retrieval thesauri stress that thesaurus relations are established between concepts, not between terms (Clarke and Zeng, 2012). However, information-retrieval thesauri are not intended for use in automatic processing of texts: they should be used in manual indexing by human experts for improvement of information retrieval in physical or digital libraries.

Thus, there exist different approaches to representing models of linguistic ontologies for natural language processing on the scale from more lexical to more conceptual resources. In this paper, we consider the approach to developing Russian ontological resources having the format of the RuThes thesaurus (Loukachevitch and Dobrov, 2014) and created for automatic processing of documents in information-analytical systems and natural language processing. These resources are linguistic ontologies uniting some principles of their organization from WordNet, information-retrieval thesauri and formal ontologies. They were utilized in various information-retrieval and NLP applications (Loukachevitch and Dobrov, 2014). RuThes was successfully evaluated in text summarization (Mani et al., 2002), text clustering (Loukachevitch et al., 2017), text categorization (Loukachevitch and Dobrov, 2014), detecting Russian paraphrases (Loukachevitch et al., 2017), etc.

We compare the RuThes model with the WordNet-like model of knowledge representation and describe some applications of RuThes-like resources for text analytics. We also consider the structure and current state of RuWordNet, WordNet-like Russian thesaurus, semi-automatically generated from RuThes data. The specificity of RuWordNet generation allows better understanding of the differences between representation models of the thesauri.

2. Existing Russian Thesauri

For the Russian language, there were at least four known projects for creating a wordnet. In the RussNet project (Azarova, 2008), the authors planned to create a Russian wordnet from scratch, guided by the principles of Princeton WordNet. In two different projects (Gelfenbeyn et al., 2003), (Balkova et al., 2008), attempts were made to automatically translate WordNet into Russian, with all the original thesaurus structure preserved. The results of (Gelfenbeyn et al., 2003) have been published¹, but the analysis of the thesaurus generated in this way shows that it requires considerable editing efforts.

The last Russian wordnet project YARN (Yet Another Russian wordNet) (Braslavski et al., 2016) was initiated in 2012 and initially was created on the basis of crowdsourcing, i.e. involvement of a large number of non-professional native speakers. Currently, YARN contains a significant number of synsets with a small number of relationships between them. The published version of the YARN² thesaurus contains too many similar or partially similar synsets introduced by different participants.

In (Azarova et al., 2016), the authors describe a current project on the integration of the RussNet thesaurus (Azarova, 2008) and the YARN thesaurus YARN (Braslavski et al., 2016) into a single linguistic resource, where the expert approach and the crowdsourcing will be combined.

For Russian, traditional information-retrieval thesauri in social sciences and the humanities have been developed and are supported in the Institute of Scientific Information of Russian Academy of Sciences (INION RAN). This institution publishes separate issues of thesauri on economics, sociology, linguistics etc., which were developed according to the guidelines of international and national standards. These thesauri cannot be used for automatic processing of document and news flows because they are, in fact, lists of selected keywords, denoting the most significant concepts of the domain, with low coverage of real texts (Mdivani, 2013). There are also several Russian versions of international information-retrieval thesauri or controlled vocabularies (Lipscomb, 2000), (Kupriyanov et al., 2016).

¹wordnet.ru

²<https://russianword.net/>

3. RuThes Family of Resources

The structure of the RuThes thesaurus of the Russian language is based on three approaches for developing computer resources: information-retrieval thesauri, WordNet-like thesauri, and formal ontologies.

The RuThes thesaurus is created in form of a linguistic ontology, which concepts are based on senses of really existing words and phrases. RuThes is a concept-oriented resource as much as possible in describing senses of Russian words and expressions. Each concept has a unique, unambiguous name. In this, RuThes is similar to information-retrieval thesauri and formal ontologies. Rules for inclusion of phrases in the thesaurus are more similar to information-retrieval thesauri guidelines (NISO, 2005).

Each concept is linked with words and phrases conveying the concept in texts (text entries). Detailed description of lexical units (words in specific senses), representation of senses of ambiguous words are closer to wordnets. Types of relations between concepts originate from information-retrieval thesauri, but some explications are made on the basis of ontological studies. There exist several large Russian thesauri presented in the same format:

- RuThes thesaurus comprising words and phrases of literary Russian together with terms of so-called sociopolitical domain (see below) (Loukachevitch and Dobrov, 2014);
- RuThes-lite³, a published version of RuThes, can be obtained for non-commercial purposes (Loukachevitch et al., 2014);
- Sociopolitical Thesaurus comprising lexical items and terms from the sociopolitical domain. The sociopolitical domain is a broad domain describing everyday life of modern society and uniting many professional domains, such as politics, law, economy, international relations, finances, military affairs, arts, and others. Terms of this domain are usually known to not only professionals, but also to ordinary people (Loukachevitch and Dobrov, 2015). Thus, this thesaurus contains important knowledge for processing news flow, legal documents, and developing new domain-specific resources. The Sociopolitical thesaurus can exist and be used separately. At the same time it is included as a part into three larger thesauri: RuThes, OENT ontology, and the Security Thesaurus;
- Ontology on Natural Sciences and Technologies (OENT) includes terms of mathematics, physics, chemistry, geology, astronomy etc., terms of technological domains (oil and gas, power stations, cosmic technologies, aircrafts, etc.). It also contains the Sociopolitical thesaurus as a part because scientific and technological problems can be discussed together with political, economical, industrial, and other issues (Dobrov and Loukachevitch, 2006);
- Security Thesaurus is an extension of the RuThes thesaurus and includes terminology related to social, national and religious conflicts, extremism and terrorism, information security.

The Table 1 contains quantitative characteristics of the above-mentioned resources.

Table 1: RuThes-like Thesauri

Thesaurus	Number of concepts	Number of Text Entries	Number of Conceptual Relations
RuThes	55,275	170,130	226,743
RuThes-lite	31,540	111,559	128,866
Sociopolitical Thesaurus	41,426	121,292	161,523
OENT	94,103	262,955	376,223
Security Thesaurus	66,810	236,321	271,297

³www.labinform.ru/pub/ruthes/index.htm

4. Specific Features of RuThes Structure

The main unit of RuThes is a concept as a unit of thought regardless of words expressing it. A concept has a unique, unambiguous name. Concept names are similar to descriptors in information-retrieval thesauri, that is precisely formulated terms referring to implied concepts. If an unambiguous and clear name in form of an existing word or a phrase cannot be found, than an ambiguous word can be used for naming and supplied with a “relator” (a brief note in parentheses).

The RuThes concepts are not divided into part-of-speech-oriented nets as in wordnets. This approach is closer to formal ontologies. Therefore, text entries of a specific concept can comprise single words of different parts of speech, including ambiguous ones, and phrases that can be either idiomatic or compositional groups. Large rows of synonyms and term variants are collected to provide better recognition of concepts in texts. The concept-based approach seems to be more convenient for text analytics and information-analytical systems in specific domains.

Fig. 1-2 show the interface of thesaurus developing. The upper left form contains a list of concepts in alphabetical order. Fig. 1 shows concepts from the Sociopolitical thesaurus: *Import of weapons*, *Import of information*, *Importer*, *Import dependence*, *Import quota*, *Import license*, *Import tax*. The lower left form shows text entries for the highlighted concept (*Import dependence*), which include: *to depend on import*, *import dependence*, *import-dependent*, *dependence on imported goods*, etc.

The right upper form presents the relations of the highlighted concept. Fig. 1 shows the relation of *Import dependence* concept with such concepts as *Economy dependence*, *Energy dependence*, *Import substitution*, *Imported goods*, and *Import*. The lower right form shows text entries for a related concept. The low right form of Fig. 1 describes text entries of *Import substitution* concept.

Fig. 2 shows a fragment from the Security thesaurus. The visible list of concepts includes: *Attack on payment system*, *Attack on search engine*, *Attack on embassy*, *Attack on vulnerability*, *Attack on torrent*, *Zero-day attack*, *DOS-attack*. Text entries of *Attack on vulnerability* concept comprise in particular such a phrase as *exploiting vulnerability*. It should be noted that this is a true, but non-evident synonym of *attack on vulnerability*, found in real texts.

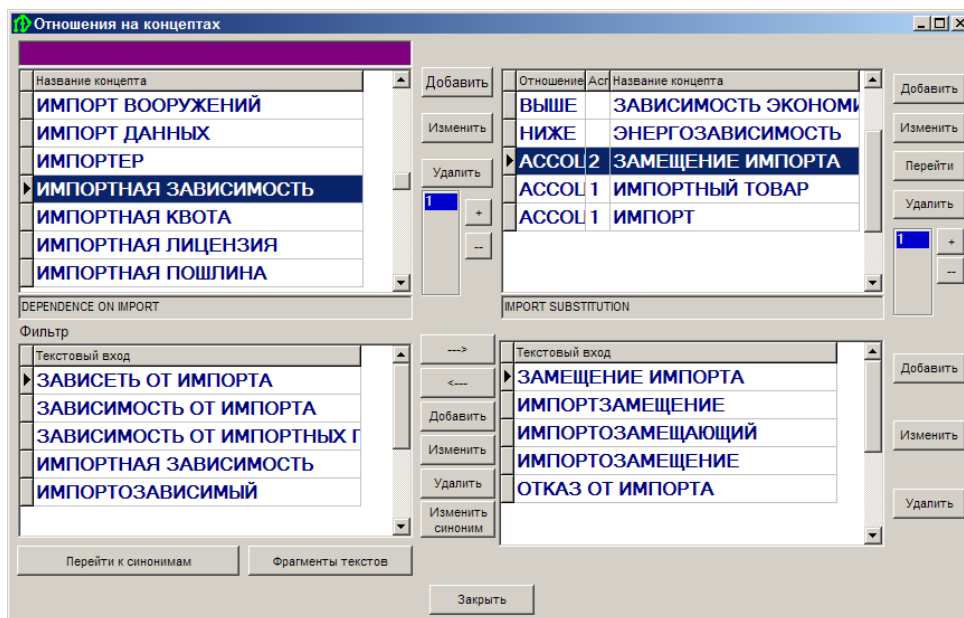


Figure 1: Relations and text entries for *import dependence* concept from Sociopolitical thesaurus

In RuThes-lite, thesauri there are four basic types of relationships between concepts. The first type of the relations is the class-subclass relationship, it has the properties of transitivity and inheritance.

The second type of the RuThes relations is the part-whole relation. An important condition for establishing this relationship in RuThes is that the concept-parts must be rigidly connected with their whole,

that is, each example of the concept-part must, throughout its entire existence, require the existence of the concept-whole. This corresponds to the guidelines of information-retrieval thesaurus standards recommending that part-whole relations should be established when the part-concept "inherently included" in the whole-concept, regardless of context (NISO, 2005). This idea can be explicated in ontological terms of inseparable parts or mandatory wholes (Guizzardi, 2011).

Under these conditions, it is possible to rely on the transitivity property of the part-whole relation, which is very important for automatic logical inference in the process of automatic text processing (Loukachevitch and Dobrov, 2015). In RuThes, the part-whole relation is used not only to describe physical parts, but also to other internal attributes, such as properties or roles for situations (Guarino et al., 2009).

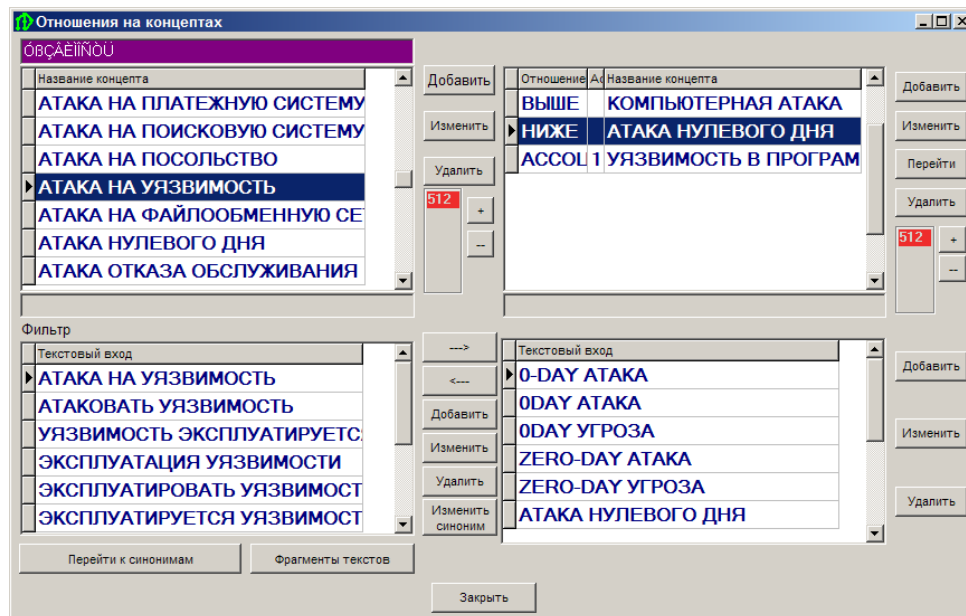


Figure 2: Relations and text entries for *zero-day vulnerability* concept from Sociopolitical thesaurus

Two other relations asymmetric and symmetric associations originate from the related term relation in information retrieval thesauri (NISO, 2005).

The asymmetric association $asc_1 - asc_2$, connects two concepts that cannot be linked by the class-subclass or part-whole relationships, but when one of which does not exist without the existence of another, for example, the *import dependence* concept can exist if only the *import* concept exists, or the *vulnerability attack* concept can appear if only the *computer vulnerability* concept exists. This relation is close to the external ontological dependence relation in ontological terms (Guarino et al., 2009).

The last type of relationships is the symmetrical association, it links concepts that are very similar in meaning, but which seems difficult to represent as one concept.

Thus, the system of the RuThes thesaurus relations describes the most significant relationships of concepts. It originates from relations used in conventional information-retrieval thesauri and explicated in existing ontological terms.

5. RuThes Ontologies in Information-Analytical Systems

RuThes ontologies are used in information-analytical systems as a tool of conceptual indexing and search, various forms of query expansion can be carried out. One of the main applications of RuThes-like ontologies is text categorization and other tasks of text analytics.

The mainstream technology of automatic document categorization is the machine-learning approach. This approach assumes that there is a sufficient training collection for learning the algorithms. However, many organizations have a need in automatic text categorization, when even a category system (system

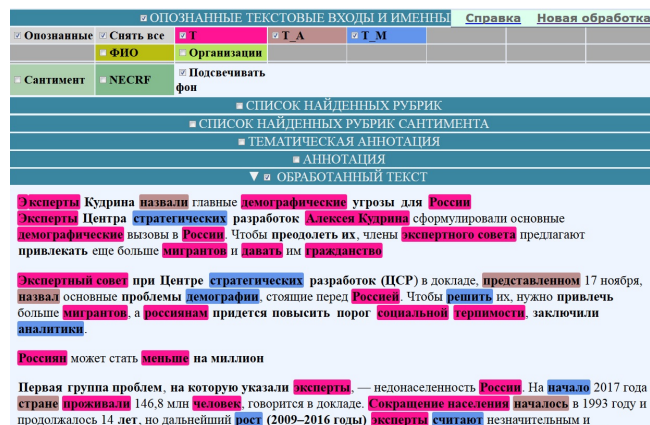


Figure 3: Security thesaurus terms found in a text. Brown and blue boxes show ambiguous terms, which should be disambiguated

of subject headings) may be absent and should be created from scratch or with the use of existing similar categorial systems.

In such conditions, machine-learning approaches cannot be applied, and knowledge-based methods of text categorization, i.e. exploiting manual rules for describing categories, are more acceptable. When one creates a hierarchical system of categories and rules for the text categorization in a broad subject domain, it is convenient to use the thesaurus support, because the thesaurus allows working not with separate words and expressions, but with concepts and substructures of the thesaurus (Loukachevitch and Dobrov, 2015).

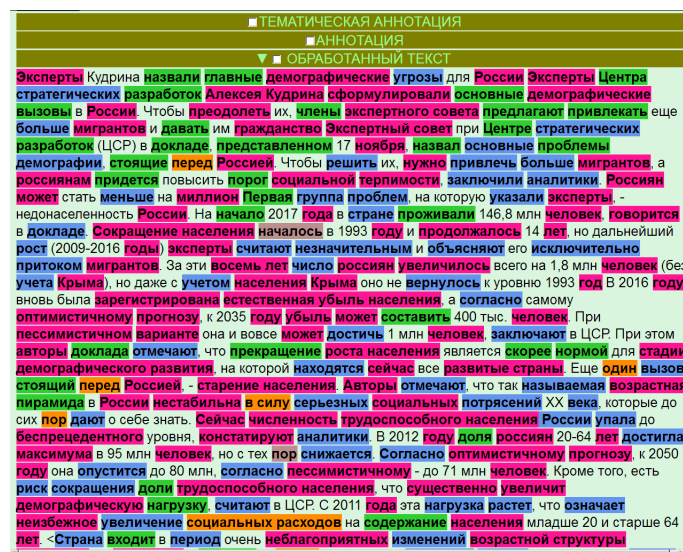


Figure 4: RuThes coverage of the same text

In RuThes-based text categorization, each category is represented by the disjunction of alternatives, each clause is a conjunction. The conjuncts, in turn, are described by experts with the help of so-called "support" concepts. For each support concept, the rule of its extension $f(\bullet)$ is defined, which determines what subordinate concepts should be included in the category profile: without an extension (denoted by the symbol "N"), the full extension in the hierarchy tree ("Y"), extension only by subclass relations ("L"), etc. In every step, automatically derived concepts in the category description can be edited.

For example, the music category can be described with the single concept *Musical art_Y*, where Y means full expansion to lower levels of the hierarchy, including hyponyms, parts, and dependent concepts. The full Boolean expression for this category looks like a disjunction of more than 400 concepts,

including musical styles, musical instruments, musicians, musical compositions, musical performances, musical groups and organizations: *Adagio or Accordion or ... Jazz music or ... Musician ... or Opera or ... or Orchestra or ... (etc.)*.

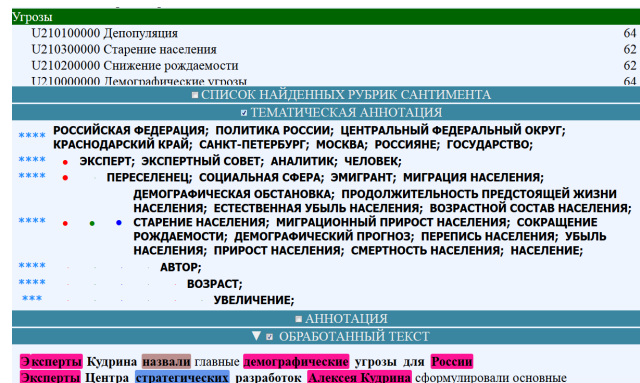


Figure 5: The upper part shows the threat categories found for a text. The central part presents topic nodes of related concepts such as "demographic situation" node.

The automatic text categorization is based on the automatically constructed thematic representation of documents that models the main topic and sub-topics of the document in sets (thematic nodes) of similar concepts mentioned in the document (Loukachevitch and Dobrov, 2015). Such a basis for the text categorization makes it possible to process texts of different types and sizes: normative acts, newspaper articles, news reports, scientific publications, or sociological surveys.

5.1. Automatic Document Processing for Text Analytics

The main stages of thesaurus-based document processing include:

- Tokenization and lemmatization, that is, the transfer of word forms to dictionary forms (lemmas);
- Matching with the thesaurus based on the lemma representation of the document. Multiword terms from a thesaurus are matched with the text using lemma sequences. Fig. 3 shows the term coverage of news text "Kudrin's experts named the main demographic threats for Russia"⁴, according to the Security thesaurus. Fig. 4 shows the coverage of matching the same text with RuThes text entries;
- Disambiguation of ambiguous text entries. Brown and blue boxes on Fig. 3 highlight ambiguous terms, which were automatically resolved. For example, Russian word *demografiya* (*demography*) can mean *demographic situation* or *demographic science*. The quality of the disambiguation procedure was previously evaluated as 75% (F-measure) for domain-specific thesauri (Security thesaurus and Sociopolitical thesaurus). Green and orange boxes on Fig. 4 show ambiguous words that were not disambiguated in the current processing. The quality of disambiguation for RuThes is much lower than for domain-specific thesauri because of the presence of ambiguous general words;
- Grouping semantically related concepts into so-called thematic nodes. This provides better determination of concept weights, which are calculated on the basis of the concept frequency in the given document and the significance of the corresponding thematic node. Fig. 5 (in the center) demonstrate such thematic nodes for the above-mentioned document about the demographic threats. The important thematic node about the demographic situation includes the following concepts: *Demographic situation*, *Life expectancy*, *Natural population decline*, *Age structure of population*, *Population aging*, *Net migration rate*, *Decline in birth rate*, *Demographic prognosis*, etc.;
- Forming the conceptual index of the document. Conceptual index of a document consist of concepts found in the document and their assigned weights. The weight of a concept accounts for the

⁴(<https://www.rbc.ru/economics/17/11/2017/5a0eb1d39a79470f724250b4>)

significance of the corresponding thematic node and the frequency of the concept in the document. In the example text, the important threat "population aging" was explicitly mentioned only once in the text, and it could obtain a too low frequency-based weight, but with the support of the main topic node "demographic situation", its weight is considerably higher;

- Calculation of category weights in dependence of concepts included into the rules of the inference for this category. Fig. 5 (upper part) shows the categories found in the mentioned document, including "Depopulation", "Population aging", "Fertility decline";
- The results of document processing, including the word index, the conceptual index, the calculated categories, etc. are loaded into an information-analytical system.

6. Generating of Russian WordNet from RuThes-lite

As it was described in Section 2, there did not exist large and qualitative Russian wordnet, but there is a demand from researchers to have a thesaurus in the WordNet format for Russian. Therefore such a thesaurus called RuWordNet was semi-automatically generated from the published version of RuThes (RuThes-lite 2.0) (Loukachevitch et al., 2018).

Table 2: Quantitative characteristics of synsets and entries in RuWordNet

Part of Speech	Number of Synsets	Number of Unique Entries	Number of Senses
Nouns	29,296	68,695	77,153
Verbs	7,634	26,356	35,067
Adjectives	12,864	15,191	18,195

The main above-mentioned differences of RuThes from WordNet-like thesauri are as follows: representation of word senses without division to parts of speech in a single hierarchical net, and conceptual relations. Thus, the first step to construct RuWordNet was to divide the source resource into three nets of nouns, verbs, and adjectives. This subdivision was based on the morpho-syntactic representation of RuThes-lite 2.0 text entries, that is part-of-speech labels for single words and syntactic classes (noun group, verb group, and adjective group) for phrases. The divided synsets were linked to each other with the relation of part-of-speech synonymy (cross-categorical synonymy). The Table 2 contains quantitative characteristics of the RuWordNet synsets, senses, and unique entries.

The hyponym-hypernym relations were established between synsets of the same part of speech. These relations include direct hyponym-hypernym relations from RuThes-lite 2.0. In addition, the transitivity property of hyponym-hypernym relations was employed in cases when a specific synset did not contain a specific part of speech but its parent and child had text entries of this part of speech. In such cases, the hypernymy-hyponymy relation was established between the child and the parent of this synset.

Similar to the current version of Princeton WordNet, in RuWordNet class-instance relations are established. By now, they had been generated semi-automatically for geographical synsets for indication of their types. The part-whole relations from RuThes were semi-automatically transferred and corrected according to traditions of WordNet-like resources.

Adjectives in RuWordNet similarly to German or Polish wordnets (Gross and Miller, 1990; Maziarz et al., 2012; Kunze and Lemnitzer, 2010) are connected with hyponym-hypernym relations. Antonyms relations were transformed from association relations and currently are established between synsets, not between lexical units. Verb synsets additionally have cause and entailment relations.

Besides, to overcome so-called "tennis problem" (Miller et al., 1990), the domain system was introduced in RuWordNet. The tennis problem is that synsets from the same domain (*tennis player*, *racket*, *court*) are very far from each other in the WordNet hierarchy. The WordNet domain system proposed in (Magnini and Cavaglia, 2000) was adapted for the RuWordNet synsets. Then domain labels were semi-

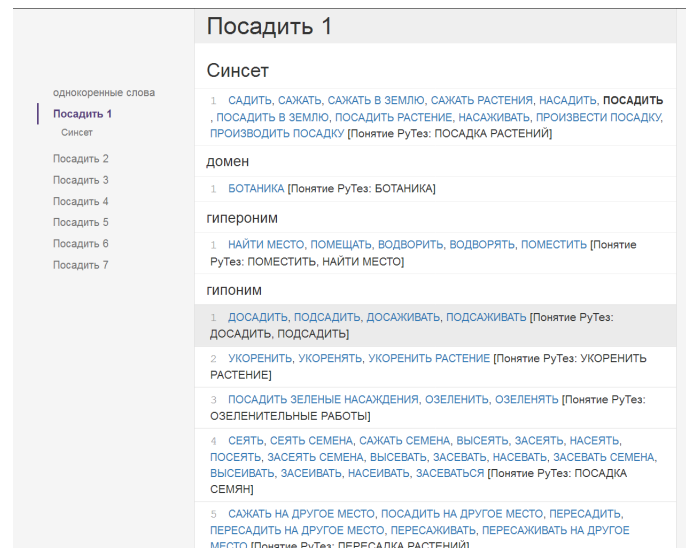


Figure 6: Screen of RuWordNet online version

automatically assigned to the RuWordNet synsets. The domains labels were represented as categories in the RuThes interface for creating knowledge-based categorization systems (see Section 5).

Fig. 6 presents description of Russian verb "posadit" in RuWordNet. It has seven senses, the first sense corresponds to English *to plant seeds, seedlings, or plants*. The description includes synonyms and variants in the synset, the reference to the source RuThes concept *Plant setting*, the link to the domain concept (Botany), the hypernym synset (to put, locate), and hyponym synsets.

7. Conclusion

In this paper we presented the RuThes family of Russian thesauri intended for natural language processing and information retrieval applications. RuThes-like thesauri include, besides RuThes, Sociopolitical thesaurus, Security Thesaurus, and Ontology on Natural Sciences and Technology. The RuThes format is based on three approaches for developing computer resources: Princeton WordNet, information-retrieval thesauri, and formal ontologies.

The published version of RuThes thesaurus (RuThes-lite 2.0) became a basis for semi-automatic generation of RuWordNet, WordNet-like thesaurus for Russian. Currently researchers can use both RuThes-lite or RuWordNet and compare them in their applications. Other RuThes-like resources are being prepared to publication.

Acknowledgements

This work is partially supported by the Russian Foundation for Basic Research (project 16-29-09606) and the Ministry of Education and Science of the Russian Federation (project N 14.601.21.0018).

References

- Azarova, I., Braslavsky, P., Zakharov, V., Kiselev, Y., Ustalov, D., and Khohlova, M. (2016). Integration of thesauri russnet and yarn. In *Proceedings of Conference "Internet and Modern Society"*, pages 7–13.
- Azarowa, I. (2008). Russnet as a computer lexicon for russian. *Proceedings of the Intelligent Information systems IIS-2008*, pages 341–350.
- Balkova, V., Suhonogov, A., and Yablonsky, S. (2008). Some issues in the construction of a russian wordnet grid. In *Proceedings of the Forth International WordNet Conference, Szeged, Hungary*, volume 44.
- Braslavski, P., Ustalov, D., Mukhin, M., and Kiselev, Y. (2016). Yarn: Spinning-in-progress. In *Proceedings of the Eight Global Wordnet Conference*, pages 58–65.

- Clarke, D. and Zeng, M. L. (2012). From iso 2788 to iso 25964: The evolution of thesaurus standards towards interoperability and data modelling. *Information Standards Quarterly (ISQ)*, 24(1).
- Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawistawska, M., and Broda, B. (2008). Words, concepts and relations in the construction of polish wordnet. In *Proceedings of the Global WordNet Conference, Seged, Hungary*, pages 162–177.
- Dobrov, B. and Loukachevitch, N. (2006). Development of linguistic ontology on natural sciences and technology. In *Proceedings of Linguistic Resources and Evaluation Conference*, pages 1077–1082.
- Fellbaum, C., Ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Gelfenbeyn, I., Goncharuk, A., Lehelt, V., Lipatov, A., and Shilo, V. (2003). Automatic translation of wordnet semantic network to russian language. In *International Dialog 2003 Workshop*, pages 148–154.
- Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5-6):907–928.
- Guarino, N. and Welty, C. A. (2009). An overview of ontoclean. In *Handbook on ontologies*, pages 201–220. Springer.
- Guarino, N., Oberle, D., and Staab, S. (2009). What is an ontology? In *Handbook on ontologies*, pages 1–17. Springer.
- Guarino, N. (1998). Some ontological principles for designing upper level lexical resources. In *Proceedings of First International Conference on Language Resources and Evaluation LREC-1998*, pages 28–30.
- Guizzardi, G. (2011). Ontological foundations for conceptual part-whole relations: the case of collectives and their parts. In *Advanced Information Systems Engineering*, pages 138–153. Springer.
- Koeva, S. (2010). Bulgarian wordnet—current state, applications and prospects. *Bulgarian-American Dialogues*, pages 120–132.
- Kunze, C. and Lemnitzer, L. (2010). Lexical-semantic and conceptual relations in germanet. *Lexical-semantic relations: Theoretical and practical perspectives*, (28):163–183.
- Kupriyanov, V., Kossilov, A., Maximov, N., and Kupriyanova, I. (2016). *A Semantic-Based Approach for Preserving Operational Experience of Nuclear Installations*. Technical report.
- Lenat, D., Miller, G., and Yokoi, T. (1995). Cyc, wordnet, and edr: critiques and responses. *Communications of the ACM*, 38(11):45–48.
- Lipscomb, C. E. (2000). Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.
- Loukachevitch, N. and Dobrov, B. (2014). Ruthes linguistic ontology vs. russian wordnets. In *Proceedings of Global WordNet Conference GWC-2014*.
- Loukachevitch, N. and Dobrov, B. (2015). The sociopolitical thesaurus as a resource for automatic document processing in russian. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 21(2):237–262.
- Loukachevitch, N., Dobrov, B., and Chetviorkin, I. (2014). Ruthes-lite, a publicly available version of thesaurus of russian language ruthes. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue, Bekasovo, Russia*, pages 340–349.
- Loukachevitch, N., Shevelev, A., Mozharova, V., Dobrov, B., and Pavlov, A. (2017). Ruthes thesaurus in detecting russian paraphrases. In *Conference on Artificial Intelligence and Natural Language*, pages 242–256. Springer.
- Loukachevitch, N., Lashevich, G., and Dobrov, B. (2018). Comparing two thesaurus representations for russian. In *Proceedings of Global WordNet Conference GWC-2018*.
- Magnini, B. and Cavaglia, G. (2000). Integrating subject field codes into wordnet. In *LREC*, pages 1413–1418.
- Magnini, B. and Speranza, M. (2002). Merging global and specialized linguistic ontologies. *Proceedings of Ontolex 2002*, pages 43–48.
- Mdivani, R. (2013). Thesauri of the isiss ras for social sciences and humanities. *Scientific and Technical Information Processing*, 40(3):137–141.

- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.
- Nirenburg, S. and Raskin, V. (2004). *Ontological semantics*. Mit Press.
- Nirenburg, S. and Wilks, Y. (2001). What's in a symbol: ontology, representation and language. *Journal of Experimental & Theoretical Artificial Intelligence*, 13(1):9–23.
- NISO. (2005). *Z39.19 - Guidelines for the Construction, Format and Management of Monolingual Thesauri*. NISO.
- Pease, A. (2011). *Ontology: A practical guide*. Articulate Software Press.
- Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1-2):161–197.
- Veale, T. and Hao, Y. (2008). A context-sensitive framework for lexical ontologies. *The Knowledge Engineering Review*, 23(1):101–115.
- Wilks, Y. (2009). Ontotherapy, or how to stop worrying about what there is. *Recent advances in natural language processing V*, pages 1–20.
- Will, L. (2012). The iso 25964 data model for the structure of an information retrieval thesaurus. *Bulletin of the Association for Information Science and Technology*, 38(4):48–51.

Resource based WordNet augmentation and enrichment

**Ranka Stanković and
Ivan Obradović**
Faculty of Mining and Geology
University of Belgrade, Serbia
ranka@rgf.bg.ac.rs
ivano@rgf.bg.ac.rs

Miljana Mladenović
College for
Preschool Teachers
Bujanovac, Serbia
ml.miljana@gmail.com

**Cvetana Krstev and
Marko Vitas**
Faculty of Philology
University of Belgrade, Serbia
cvetana@matf.bg.ac.rs
vitas.marko@gmail.com

Abstract

In this paper we present an approach to support production of synsets for Serbian WordNet (SerWN) by adjusting Princeton WordNet (PWN) synsets using several bilingual English-Serbian resources. PWN synset definitions were automatically translated and post-edited, if needed, while candidate literals for Serbian synsets were obtained automatically from a list of translational equivalents compiled from bilingual resources. Preliminary results obtained from a set of 1248 selected PWN synsets show that the produced Serbian synsets contain 4024 literals, out of which 2278 were offered by the system we present in this paper, whereas experts added the remaining 1746. Approximately one half of synset definitions obtained automatically were accepted with no or minor corrections. These first results are encouraging, since the efficiency of synset production for SerWN was increased. There is also space for further improvement of this approach to wordnet enrichment.

1. Introduction

Semantic networks, such as wordnets, are among the most important resources in Human Language Technologies. Thus, for example, the Princeton WordNet - PWN (Fellbaum, 1998), has been in use for more than two decades as the standard lexical database for English. Several projects inspired by PWN for the development of wordnets for clusters of other languages have subsequently emerged, such as EuroWordNet - EWN (Vossen, 1998) for a cluster of European languages, and BalkaNet -BWN (Tufis et al., 2004) for a cluster of languages from the Balkans.

The EuroWordnet and BalkaNet projects extended the PWN approach by introducing an Inter-Lingual Index (ILI), used to align synsets in different languages within each cluster on basis of PWN structure. In BWN, the structure of PWN was extended to accommodate concepts specific to Balkan languages that were not found in PWN.

Manual development and maintenance of wordnets has proved to be a demanding task, requiring a considerable amount of human resources. As a consequence, some wordnets developed within the EWN and BWN projects, as well as some others wordnets, developed independently, have been stalled or frozen. In order to overcome this issue, methods for semi-automatic and automatic development and enhancement of wordnets have been proposed.

Even PWN, despite being one of the most comprehensive wordnets, is facing the development problem, as many concepts, especially domain specific, or newly emerging concepts generated by technological developments, need to be added. Targeting specifically PWN WordNet 3.0, SemEval-2016 proposed Task 14 for development of systems for automatic enrichment of taxonomies with new word senses, drawn from other lexicographic resources. The task provided prospective systems with a set of word senses that were not in PWN, where each word sense comprised a lemma, a part of speech, and a definition. The systems were to identify the most plausible point for placing each of the word senses from this set in PWN, either by merging it into an existing synset, or adding it as a new hyponym synset.

Keywords: WordNet, bilingual resources, term alignment, parallel lists

Five teams submitted 13 systems, with all teams performing better than chance, but only one team surpassing a simple baseline, thus pointing towards a need for further efforts aimed at improving taxonomy enrichment (Jurgens and Pilehvar, 2016).

When non-English wordnets are concerned, one of the common approaches to wordnet enhancement is the use of a developed wordnet (most commonly PWN) as the basic source for enrichment, and a method for aligning the synsets of the two wordnets, such as the ILI. Automatic alignment of synsets belonging to different languages is closely related to the task of pairing their word senses. This approach was followed by Matuschek and Gurevych (2013) who solved the word sense alignment (WSA) task by pairing senses with the same meaning from different lexical-semantic resources.

Besides alignment with a developed wordnet, the use of other available resources for development and enrichment of wordnets have also been proposed. Thus, Oliver and Climent (2014) used parallel corpora for five European languages to produce aligned wordnets. The English part of each corpus was semantically tagged, after which the process of wordnet creation was transformed into a word alignment problem, where wordnet synsets in the English part of the corpus were aligned with in the target language part of the corpus. The obtained precision was satisfactory, but the overall number of extracted synset pairs was too low, resulting in poor recall.

In methods used for automatically enriching wordnets using other available lexical resources, the successfulness of the method is strongly correlated with the comprehensiveness of the resource used in the alignment process (Hrstea, 2007). Different methods and resources can be used for alignment. One of the common approaches is to take PWN as the source for alignment, and a bilingual dictionary of English and the target language.

There are, however, several other approaches. In (Chugur et al., 2001) a monolingual and a bilingual Spanish-English dictionary were used for the enrichment of the Spanish WordNet with adjective synsets, where the bilingual dictionary was used for alignment with EWN. Bentivogli and Pianta (2003) proposed a method for extending MultiWordNet¹ with phrases, by extracting them from bilingual dictionaries and corpora with techniques similar to those used for collocation extraction. In (Bhingardive et al., 2014) a method for Sanskrit WordNet extension using a Sanskrit English dictionary is presented. The dictionary entries were automatically extracted and linked to PWN. In (Simões et al., 2016) an approach is given to enrichment of Portuguese WordNet with synonyms available in a standard monolingual Portuguese dictionary. Dictionary word senses were automatically identified, annotated and extracted, independently of definitions, and synsets were created. These synsets were then aligned with Portuguese WordNet synsets. The process resulted in both the addition of new synonyms to existing synsets, and in the creation of new synsets.

There were also some interesting approaches related to Slavic languages. Vintar and Fišer (2017) proposed a method to enrich the Slovenian WordNet (sloWNet) with domain-specific single and multi-word expressions. They used a large monolingual Slovene corpus of texts to extract terminology from the domain of informatics, and a parallel English-Slovene corpus and an online dictionary as bilingual resources to facilitate the addition of new terms to sloWNet, based on the semantic structure of the PWN. The Croatian Wordnet (CroWN) was enlarged with the WN-Toolkit and CroDeriV, a derivational database for Croatian (Oliver et al., 2015), from 10,026 synsets and 31,367 synset-lemma pairs to 23,137 synsets and 47,931 synset-lemma pairs. The WN-Toolkit, a set of Python programs for wordnet creation and expansion uses a dictionary, Babelnet and parallel-corpora based strategies.

In this paper we present a method to speed up the enrichment of the Serbian WordNet (SerWN) using PWN, several bilingual resources and the Google Translate API. The paper is organized as follows: in Section 2. we present wordnets and bilingual resources that are the base of our solution described in Section 3., while in Section 4. we discuss results of preliminary automatic and manual evaluation. Finally, in Section 5. we give some concluding remarks and plans for future research and implementation of results.

¹<http://multiwordnet.fbk.eu/english/home.php>

2. Resources used

2.1. WordNets

Serbian WordNet was initially developed within the scope of the BalkaNet project and subsequently manually enhanced and upgraded. At present it contains 22,530 synsets, out of which 18,248 are nouns, 2,249 verbs, 1,907 adjectives and 126 adverbs. There are 12,248 synsets with only one literal, 7,204 with 2 literals, 2,130 with 3 literals, 612 with 4 literals, 215 with five literals, and 121 synsets with more than five literals. The majority of SerWN synsets are aligned with corresponding PWN 3.0 synsets via the Interlingual Index, with the exception of a little over 1,000 Serbian specific synsets that do not exist in PWN.

In our research we used 20,221 aligned synsets (from a previous version of SerWN), coupled with a number of bilingual resources, to speed up the development of SerWN, by selecting and adjusting PWN synsets from a set of approximately 95,000 PWN synsets without an equivalent in SerWN (further on referred to as non-adjusted PWN synsets).

We analyzed at the outset the number of literals in the 20,221 aligned synsets of PWN and SerWN, and found out that 11,091 synsets (55%) had the same number of literals, that the number of literals in 6,107 synsets (30%) differed by 1, in 1,834 synsets (9%) by 2, in 682 for (3%) by 3 literals, whereas an even greater difference in the number of literals was found in the remaining 507 (3%) synsets. For the 11,091 synsets with same number of literals in English and Serbian the distribution according to the number of literals was as follows: there were 7063 (64%) synsets with one literal, 3288 (30%) with two, 561 (5%) with three, and 179 (1%) with more than three literals. As for the remaining synsets with different number of literals, the smaller number of literals was mainly found in SerWN. Namely, the average number of literals per synset in PWN is 2.01, whereas the average number of literals per synset in SerWN is 1.66. After the initial analysis, we used the 20,221 aligned synsets to produce a parallel list of 72,262 literals for the purpose of this research. For example, we used the aligned synset pair:

```
building, edifice -- zgrada, kuća
```

to produce the following list:

```
building -- zgrada
building -- kuća
edifice -- zgrada
edifice -- kuća
```

We have produced another list from non-adjusted PWN synsets, where each item in the list consisted of an ID, a definition and list of literals. The items were then translated using Google Translation API. We used the *LanguageApp* service in Google Apps Script² to create our own version of Language Translation API, which, unlike the official Google Language Translation API, produces text translated into Serbian in Latin script, instead of Cyrillic, and serializes it into a plain text file.³

An example of a list item is:

```
ENG30-08331011-n | a court with jurisdiction in equity | chancery; court of chancery
which is converted, after translation by our Translation API based on Google Language Translation API,
into:
```

```
ENG30-08331011-n | sud koji je nadležan za pravičnost | kancelarija ; sudski ured
```

If a Serbian translational equivalent for a PWN literal obtained by Google Translation is not offered by any other bilingual resource, it is taken into account only if the term is found in the Serbian morphological e-dictionary and if the POS of the term matched the synset POS.

2.2. Bilingual lists

In addition to the list of English-Serbian (en-sr) term pairs extracted from aligned PWN and SerWN synsets, we have used thirteen other bilingual resources to produce additional lists of term pairs, as

²Google Apps Script is a scripting language based on JavaScript that provides easy ways to automate tasks across Google products and third party services and build web applications – <https://developers.google.com/apps-script/overview>

³<https://cloud.google.com/translate/docs/translating-text>

a means to speed up the process of adjusting PWN synsets that still do not have their counterpart in SerWN. A brief description of these resources follows.

Parallel list is a simple bilingual parallel list, developed gradually from various resources and used as an auxiliary resource in WS4LR (later upgraded and dubbed LeXimir), a workstation for lexical resources we have developed (Krstev et al., 2006), to produce SerWN synsets on basis of PWN and to support multilingual queries in English and Serbian. The structure of bilingual parallel lists has in general the following form:

```
TermI[,TermI]*; TermII[,TermII]*
```

where TermI represents a word in one language, and TermII a corresponding word in another. A few examples from the en-sr parallel list are:

```
grandmother;baka,baba
soft drink;bezalkoholno piće
safe,secure;bezbedan,siguran
```

From the en-sr parallel list we produced a simple list of translation pairs to be used in our approach. For example, the four items from the en-sr parallel list listed above generated 9 items in the simple list of translation pairs:

```
grandmother;baka
grandmother;baba
soft drink;bezalkoholno piće
secure;bezbedan
secure;siguran
safe;bezbedan
safe;siguran
```

The Dictionary of Library and Information Sciences⁴ is a terminology dictionary that covers the library and information sciences and related disciplines in Serbian, English and German, developed at the National Library of Serbia (Kovačević et al., 2004) (further referred to as Biblio). It contains 12,060 aligned lexical entries, from which a list of 17,761 aligned term pairs was produced.

GeolISSTerm (Stanković et al., 2011) and RudOnto (Kolonja et al., 2016) are bilingual resources developed at the Faculty of Mining and Geology, University of Belgrade (FMG). GeolISSTerm⁵ is a thesaurus of geological terms with entries in Serbian and English, containing 4,788 concepts in Serbian and English. From this resource, a list of 6,213 aligned term pairs was produced, both from preferred terms and their synonyms. RudOnto⁶ is another complex bilingual terminological resource developed at FMG with the aim of becoming the reference resource in Serbian for mining terminology. It contains 1,041 concepts in Serbian and English, from which a list of 1,273 aligned term pairs was produced, both from preferred terms and their synonyms.

Termi application⁷ was also developed at the FMG, with the support of Tempus BAEKTEL project⁸. It has been aimed to support the development of terminological dictionaries in various domains within BAEKTEL. It contains 870 concepts in Serbian and English, which were used to produce 1,290 aligned term pairs.

The structure of lexical entries in GeolISSTerm, Rudonto and Termi is similar. Each term comes with a name, definition, an optional list of synonyms, abbreviations and a bibliographic source. Each term, except the top term in the dictionary tree, has only one hyperonym term, but it can have an arbitrary number of hyponym terms. However, in our approach we have used only the relation between aligned terms and their synonyms to produce aligned term pairs.

Eurovoc⁹ is a multilingual, multidisciplinary thesaurus covering the activities of the EU. Serbian is one of the 23 languages managed within this thesaurus. EuroVoc is managed by the Publications

⁴<http://rbi.nb.rs/en/dict.html>

⁵<http://geoliss.mre.gov.rs/recnik/>

⁶<http://rudonto.rgf.bg.ac.rs/>

⁷<http://termi.rgf.bg.ac.rs/>

⁸<http://baektel.eu/>

⁹<http://eurovoc.europa.eu/>

Office, which moved forward to ontology-based thesaurus management and semantic web technologies compliant to W3C recommendations, as well as latest trends in thesaurus standards. For this research we used the bilingual en-sr version 4.7 in xls format, with 6,939 term entries, and 6,971 aligned pairs of terms.

Microsoft language portal¹⁰ has published Microsoft Terminology Collection data in the form of a .tbx (ISO 30042:2008) file containing: Concept ID, Definition, Source term, Source language identifier, Target term, Target language identifier. The number of terms differ from language to language, due to varying levels of localization. The Microsoft Terminology Collection is a set of standard technology terms used across Microsoft products, and comprises 13,147 aligned terms for Serbian.

Additional bilingual lexicons from domains of ecology, psychology, sport and mathematics were compiled from various on-line resources and textbooks and used for generating lists of translational equivalents. Finally, we compiled a list of aligned en-sr terms from the web site of the Serbian Institute for Standardization.

3. Production

The final parallel list of translation equivalents compiled from all of the abovementioned resources had 151,641 different en-sr term pairs, as summarized in Table 1. The first two columns show the bilingual resource the term pairs originated from, and its abbreviation. The numbers of en-sr pairs generated from each bilingual resource are shown in column 3. The total number of different English terms in the final parallel list was 90,563, and column 4 shows the number of pairs, per each resource, where the English word within the pair was found among English literals from non-adjusted PWN synsets, while the last column shows the corresponding percentage. It comes as no surprise that the Parallel list, as a general-domain lexicon has a considerably higher percentage than domain-specific resources. However, when it comes to particular domains, aligned terms from those resources tend to be more relevant. There were 124,887 literals within the non-adjusted PWN synsets, out of which 22,645 had at least one corresponding Serbian literal in the compiled parallel list.

Src	Source	No of pairs	Mapped with PWN	Percentage
L	Parallel list	10542	9071	86
B	Biblio	17761	4296	24
P	Psih	11898	3945	33
K	Eko	4900	2850	58
I	Iss	16423	2560	15
E	Eurovoc	6971	1324	18
M	Microsoft	13147	2398	18
G	Geoliss	6213	701	11
R	Rudonto	1273	376	29
T	Termi	1290	346	26
S	Sport	634	194	30
H	Math	178	90	50
W	WordNet	72262	27600	38
O	Google	290639	290593	99

Table 1: Overview of term-pairs in the final parallel list and their mapping with PWN literals

Each term pair in the final parallel list is accompanied by information about POS (part of speech), number of occurrences (frequency) within all resources as well as information on the resources where it was found, as illustrated by examples in table 2. For example, the term pair `access;pristup` appeared a total of 9 times in six different resources: BIMPKO, namely B–Biblio, I–Iss, M–Microsoft, P–Psih, K–Eko, O –Google list.

¹⁰<https://www.microsoft.com/en-us/language/Terminology>

POS	English	Serbian	Frequency	Source
N	access	pristup	9	BIMPKO
N	base	baza	7	BILWO
N	base	osnova	6	BILWO
N	contact	kontakt	7	MPRWO
N	content	sadržaj	7	BMPWO
V	fall	padati	4	LWO
V	fall	pasti	4	LWO
V	enrich	obogatiti	3	MLO

Table 2: Examples of en-sr term pairs in the final parallel list

3.1. Feasibility assessment of the approach

In order to assess the feasibility of our approach, we have performed an experiment with the aligned PWN and SerWN synsets. Namely, we have used the compiled bilingual lists, excluding the list generated by the aligned synsets, to find the number of English literals for which the expected corresponding Serbian literal would be offered by the compiled bilingual list. For each synset, the number of literals for which the expected term was offered was recorded. In general, the expected corresponding literal was found for more than 25% of literals from all synsets. A correct corresponding literal for all literals was offered for 3,925 synsets, while 13,724 did not have any offered equivalent. In Figure 1 numbers of synsets with expected corresponding equivalents are shown in comparison to those with no equivalent offered, for synsets with 1, 2, 3, 4 and more than 4 literals, respectively. For example, 11,024 (54%) of all synsets had only one literal, and for 3390 (31%) of these synsets the expected Serbian equivalent was offered. Similarly, 6540 (32%) of synsets had two literals, and 1341 (21%) of them had the expected Serbian equivalent offered, and so on.

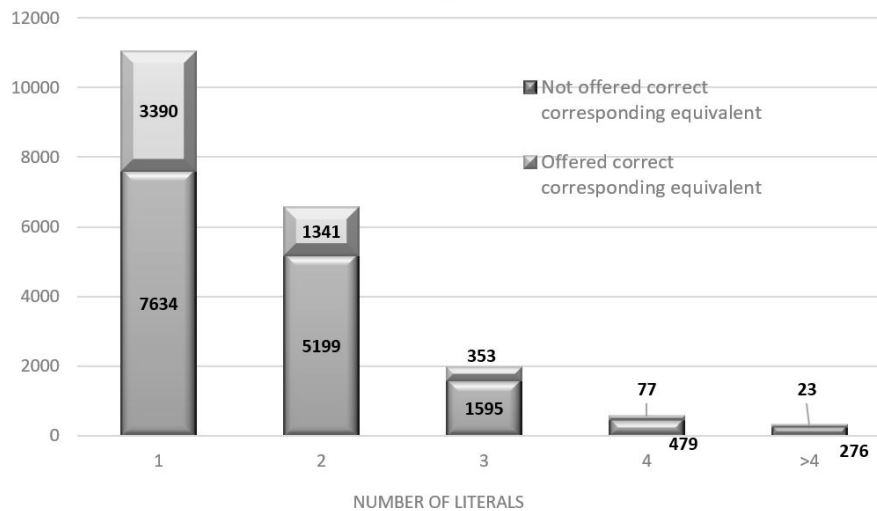


Figure 1: Number of synsets with and without expected corresponding equivalents

We found the results of this experiment encouraging, which motivated us to proceed with the procedure of semi-automatic adaptation of non-adapted PWN synsets, on basis of the final list of parallel terms. Given that the final parallel list was larger than the one we used in our experiment (as it included the list of parallel terms extracted from aligned synsets), we expected to get even better results than the ones in the experiment.

3.2. The semi-automatic adjustment procedure

In order to evaluate our approach for expanding SerWN the final list of parallel terms and a subset of non-adjusted PWN synsets were presented to experts, specialist for specific domains. Their task was to adjust the 1418 PWN synsets (categories) from the so called Base Concepts Set (BCS) that had no equivalents in SerWN. In addition to that, smaller sets of PWN synsets, for which corresponding terms in the parallel list of terms were available, were selected from the following specific domains: computer_science, earth, ecology, engineering, geology, a total of 803 synsets. The experts were expected to perform the following three tasks:

- post-edit Serbian version of synset definition produced by our Translation API;
- select appropriate literals for each Serbian synset from a list of translational equivalents offered by the system;
- add new literals, if adequate candidate(s) could not be found on the list of translational equivalents, but also in other cases if deemed appropriate.

Figure 2 presents an example of the document generated to support experts in their adjustment task. Column ID contains the interlingual index, used to bind definition and literals of a specific synset. Column domain shows the synset domain from PWN and is used to check whether the synset has been assigned to the appropriate domain expert. In order to reduce false alignments due to homographs, we introduced a constraint to check whether the POS mark for the PWN synset and the POS of the Serbian equivalent offered for a term in that synset are equal, where Serbian POS was available. Column POS shows the PWN synset POS, while the last column SrpPOS shows the POS for the Serbian equivalent, obtained from Serbian morphological dictionaries (Krstev et al., 2010), if this term was found in the dictionary.

Definition in English DefEn is intended to help the expert in correcting the definition in Serbian DefSr if needed, as the latter is generated using the Translation API, and is in general subject to post-editing. Pairs of literals from the original English synset and the final list of parallel terms are given in LiteralsEn, LiteralsSr. In the column Mark the expert marks if the offered literal is acceptable. The next column Freq shows the frequency of the candidate pair in the final list of parallel terms, whereas the resources in which the term pair was found are displayed in column Sources. In general, the higher the frequency and the number of resources, the greater the chance that the offered Serbian term is an appropriate candidate. Each resource, as mentioned earlier, is represented by one letter, e.g. ML next to the term pair update - ažurirati means that this pair was found in the list obtained from Microsoft Terminology Collection data (M) and the parallel list (L).

If the expert wants to add a new literal, he/she must insert a row, and copy the appropriate ID in the first column of the inserted row, in order to enable further automatic mapping of literals.

ID	domain	POS	DefEn	DefSr	LiteralsEn	LiteralsSr	Mark	Freq	Sources	Srp PC
ENG30-00170857-v	compute_r_science	V	bring to the latest state of technology	dovesti do najnovijeg tehnološkog stanja	update;	ažuriranje		0		
ENG30-00170857-v								0		
ENG30-00170857-v					update	ažurirati	da	3	ML	V
ENG30-00170857-v					update	osavremeniti	da	2	LK	V
ENG30-00170857-v					update	ažurirana verzija		1	B	
ENG30-00170857-v					update	modernizovati		1	K	V
ENG30-00200242-v	compute_r_science	V	locate and correct errors in a computer	pronaći i ispraviti greške u kodu računarskog programa	debug;	debug		0		
ENG30-00200242-v								0		
ENG30-00200242-v					debug	otклонiti grešku	da	1	M	
ENG30-00200242-v					debug	trebiti		1	P	V

Figure 2: An excerpt from the document to support the adjustment task

3.3. Integration into SerWN

At the beginning of its development, SerWN was maintained using VisDic (Horák and Smrž, 2004), a free tool for editing wordnets in XML format. A more powerful new web tool SWNE¹¹ was subsequently developed, inheriting all useful characteristics of VisDic and enriched by full XML support, distributed work, and advanced search, as presented in (Mladenović et al., 2014). Further improvements of this tool in developing and maintaining SerWN are presented in (Mladenović and Mitrović, 2014), the main improvement being the replacement of XML by a relational database. The tool provides monitoring of relevant statistics, such as the total number of synsets, the number of semantic relations, quality and speed of enriching SerWN (statistics listed by days and by authors), etc.

The current version of SerWN, thanks to the SWNE tool, integrated external resources – SUMO (Niles and Pease, 2003), SentiWordNet (Baccianella et al., 2010), WordNet Domains (Bentivogli et al., 2004) and PWN v3.0 (Fellbaum, 1998). The official mapping files¹² are used for mapping each synset with these resources. One of the goals of our approach is to further develop the semi-automatic procedure to enable automatic evaluation, periodical analysis and automatic extraction of XML synsets structure for further processing and integration into SWN with sumo, sentiments, relations etc.

In our approach, we used the procedure described in (Mladenović and Mitrović, 2014) for conversion from VisDic XML into MS SQL Server database format, and adjusted it for updating and inserting new synsets into SerWN. While inserting, the procedure checks if there are mapping entries for SUMO, SentiWordNet and Domains resources and generates the appropriate links.

4. Initial results

As the application of the procedure we have developed is a work in progress so far we can report the initial results for 1248 synsets. Results show that in the set of 1248 produced Serbian synsets having a total of 4024 literals, experts accepted 2278 literals offered by the system, and added the remaining 1746.

The average number of English literals in candidate en-sr pairs was 2.1, with the average number of 10.2 literals offered as candidates in Serbian. The average number of Serbian literals in produced Serbian synsets (both selected and added) was 3.2. Figure 3 presents the number of literals selected as correct corresponding literal equivalents for a specific concept, number of English literals in the corresponding PWN synset, and Serbian literals offered by the system. It can be seen that there was a large number of candidates that were rejected, which we plan to reduce by introducing additional constraints.

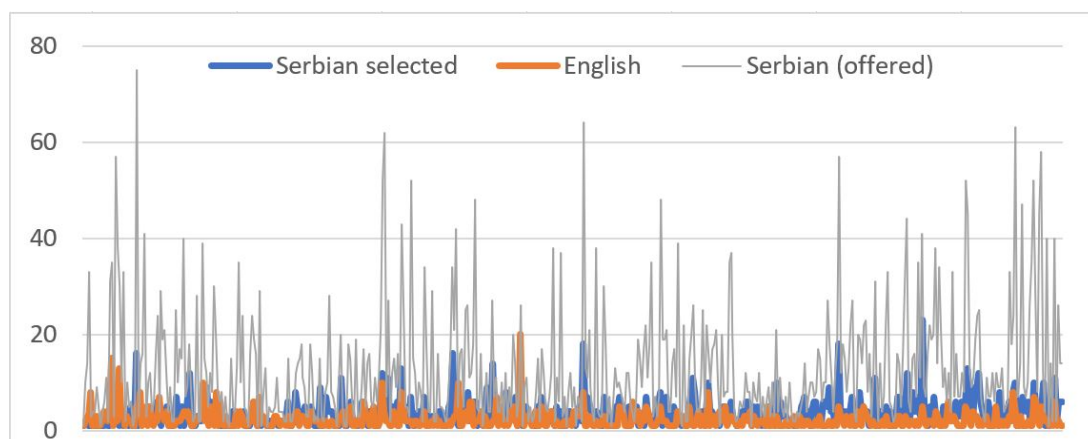


Figure 3: Number of literals per synset

We also evaluated the usability of the Translation API based on Google Language Translation API

¹¹<http://sm.jerteh.rs>

¹²SentiWordNet – <http://sentiwordnet.isti.cnr.it>; SUMO - <https://github.com/ontologyportal/sumo>; WordNet Domains – <http://wndomains.fbk.eu/>; PWN – <http://wordnet.princeton.edu/wordnet/download/current-version/>

for automatic translation of synset definitions. All definitions for the analyzed synsets, automatically translated from English into Serbian by Google were checked and post-edited, if needed. Semantical similarity between automatically obtained and post-edited translations was measured by Levenshtein distance (LD). As to the number of corrections during post-editing, it was found that 21% of automatic definitions did not have corrections, 12% had correction up to 5 characters, 17% up to 10, etc. It is thus safe to say that for approximately half of the definitions obtained by automatic translation, there were no corrections or corrections were moderate. Only 15% definitions had corrections of over 30 characters (Figure 4).

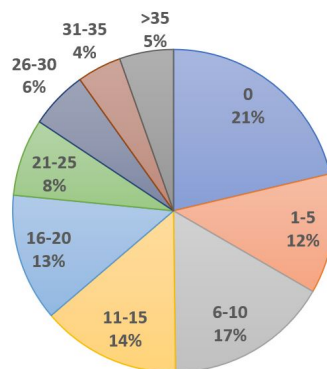


Figure 4: Number of corrections in definition during post-editing phase

5. Concluding remarks and future work

In this paper we presented an approach to speed up WordNet enlargement using bilingual resources and machine translation APIs. PWN synset definitions from a selected set of non-adjusted PWN synsets were translated to Serbian using Google API and manually post-edited. For synset literals, the approach used a compilation of a bilingual list of en-sr pairs of terms, which were aligned with literals from PWN synsets. Since our approach is aimed at semi-automatic production and evaluation of SerWN synsets, the procedure includes obligatory post-editing of automatically prepared datasets. In our first case-study we applied our approach to 1248 synsets, and found out that expert had that 56.6% of literals in produced Serbian synset were literals offered by the system. Also, approximately one half of synset definitions in Serbian offered by the system did not need corrections, or the corrections were moderate. We can thus conclude that the system developed in our approach offers considerable help in augmentation and enrichment of SerWN. However, there are several opportunities for improvement.

As the average number of literals per synset in SerWN (1.66) is about 20% lower than the average number of literals per synset in PWN (2.01), in the following period not only the addition of new synsets, but also the addition of literals to existing synsets will be considered, to ensure a better representation of concepts in SerWN. In addition to that, we plan to broaden the set of parallel resources, and search for new pairs of aligned literals for synsets, which will then be manually post-edited. We also plan to use parallel corpus based methodologies relying on two strategies proposed in ((Oliver et al., 2015)) for automatic construction of the required corpora: by machine translation of sense-tagged corpora and by automatic sense-tagging of English-Serbian parallel corpora. POS tag annotation of bilingual en-sr parallel list is also envisaged, with the aim of increasing precision of offered translational equivalents.

We will investigate possibilities for expanding the information offered to experts by including examples of usage and hypernyms, which would further clarify the concept context. In addition to the Translation API, we plan to include other APIs, like BING API. The bottleneck of the procedure, manual evaluation, will be further supported by crowdsourcing techniques and by the development of a user-friendly web application for post-editing of automatically adapted Serbian synsets. We also plan to include further restrictions using semantic and domain markers from Serbian morphological dictionaries and other resources, in order to improve precision of system.

Acknowledgements

This research was partially supported by Serbian Ministry of Education and Science under the grants #III 47003 and 178006.

References

- Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., Eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Bentivogli, L. and Pianta, E. (2003). Beyond lexical units: Enriching wordnets with phrasets. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2, EACL '03*, pages 67–70, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Bentivogli, L., Forner, P., Magnini, B., and Pianta, E. (2004). Revising the wordnet domains hierarchy: Semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, MLR '04, pages 101–108, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Bhingardive, S., Ajotikar, T., Kulkarni, I., Kulkarni, M., and Bhattacharyya, P. (2014). Beyond lexical units: Enriching wordnets with phrasets. In *Global WordNet Conference (GWC)*.
- Chugur, I., Peñas, A., Gonzalo, J., and Verdejo, F. (2001). Monolingual and bilingual dictionary approaches to the enrichment of the spanish wordnet with adjectives. In *Carnegie Mellon University, Pittsburgh*.
- Fellbaum, C., Ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Horák, A. and Smrž, P. (2004). VisDic – Wordnet browsing and editing tool. In *Proceedings of the GWC 2004*, pages 136–141.
- Hristea, F. (2007). Semiautomatic generation of wordnet type synsets and clusters using class methods. an overview. *Revue roumaine de linguistique*, 52:97–133.
- Jurgens, D. and Pilehvar, M. T. (2016). Semeval-2016 task 14: Semantic taxonomy enrichment. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1092–1102, San Diego, California, June. Association for Computational Linguistics.
- Kolonja, L., Stanković, R., Obradović, I., Kitanović, O., and Cvjetić, A. (2016). Development of terminological resources for expert knowledge: a case study in mining. *Knowledge Management Research & Practice*, 14(4):445–456.
- Kovačević, L., Injac, V., and Begenišić, D. (2004). *Bibliotekarski terminološki rečnik: englesko-srpski, srpsko-engleski*. Narodna biblioteka Srbije.
- Krstev, C., Stanković, R., Vitas, D., and Obradović, I. (2006). WS4LR: A Workstation for Lexical Resources. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006*, pages 1692–1697.
- Krstev, C., Stanković, R., and Vitas, D. (2010). A Description of Morphological Features of Serbian: a Revision using Feature System Declaration. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., and Tapias, D., Eds., *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).
- Matuschek, M. and Gurevych, I. (2013). Dijkstra-WSA: A graph-based approach to word sense alignment. *TACL*, 1:151–164.
- Mladenović, M. and Mitrović, J. (2014). *Natural Language Processing for Serbian – Resources and Application*, chapter Semantic Networks for Serbian: New Functionalities of Developing and Maintaining a WordNet Tool. University of Belgrade, Mathematical Faculty.
- Mladenović, M., Mitrović, J., and Krstev, C. (2014). Developing and maintaining a wordnet: Procedures and tools. In *Proceedings of the GWC 2014*, pages 55–62.

- Niles, I. and Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, pages 412–416.
- Oliver, A. and Climent, S. (2014). Automatic creation of wordnets from parallel corpora. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., Eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).
- Oliver, A., Šojat, K., and Srebačić, M. (2015). Enlarging the croatian wordnet with wn-toolkit and cro-deriv. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 480–487.
- Simões, A., Gómez, X. G., and Almeida, J. J. (2016). Enriching a portuguese wordnet using synonyms from a monolingual dictionary. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., Eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).
- Stanković, R., Trivić, B., Kitanović, O., Blagojević, B., and Nikolić, V. (2011). The Development of the GeolIS-STERM Terminological Dictionary. *INFOtheca*, 12(1):49a–63a.
- Tufis, D., Cristea, D., and Stamou, S. (2004). Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information science and technology*, 7(1-2):9–43.
- Vintar, Š. and Fišer, D. (2017). Enriching Slovene wordnet with domain-specific terms. *Annotation, exploitation and evaluation of parallel corpora: TC3 I: TC3 I, 3*.
- Vossen, P. (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Norwell, MA, USA.

Classifying Verbs in WordNet by Harnessing Semantic Resources

Svetlozara Leseva
Institute for Bulgarian
Language – BAS
zarka@dcl.bas.bg

Ivelina Stoyanova
Institute for Bulgarian
Language – BAS
iva@dcl.bas.bg

Maria Todorova
Institute for Bulgarian
Language – BAS
maria@dcl.bas.bg

Abstract

This paper presents the principles and procedures involved in the construction of a classification of verbs using information from 3 semantic resources – WordNet, FrameNet and VerbNet. We adopt the FrameNet frames as the primary categories of the proposed classification and transfer them to WordNet synsets. The hierarchical relationships between the categories are projected both from the hypernymy relation in WordNet and from the hierarchy of some of the frame-to-frame relations in FrameNet. The semantic classes and their hierarchical organisation in WordNet are thus made explicit and allow for linguistic generalisations on the inheritance of semantic features and structures.

We then select the beginners of the separate hierarchies and assign classification categories recursively to their hyponyms using a battery of procedures based on generalisations over the semantic primes and the hierarchical structure of WordNet and FrameNet and correspondences between VerbNet superclasses and FrameNet frames. The so-obtained suggestions are ranked according to probability. As a result, 13,465 out of 14,206 verb synsets are accommodated in the classification hierarchy at least through a general category, which provides a point of departure towards further refinement of categories.

The resulting system of classification categories is initially derived from the WordNet hierarchy and is further validated against the hierarchy of frames within FrameNet. A set of procedures is established to address inconsistencies and heterogeneity of categories. The classification is subject to ongoing extensive manual verification, essential for ensuring the quality of the resource.

1. Introduction

This paper outlines the principles and procedures involved in the elaboration of a hierarchical classification of verbs through the mapping of three semantic resources – WordNet, FrameNet and VerbNet. The classification is induced from the relational hierarchies of WordNet and FrameNet and derives its categories from FrameNet frames and VerbNet superclasses which are assigned to WordNet synsets. More specifically, we adopt the FrameNet frames as the primary categories of the classification and transfer them to WordNet synsets through a set of procedures involving either (i) exact mapping or (ii) generalisations over the hierarchical structure of WordNet and FrameNet and correspondences between VerbNet superclasses and FrameNet frames. The hierarchical relationships between the categories are projected both from the hypernymy relation in WordNet and from the hierarchy of some of the frame-to-frame relations in FrameNet.

The classification exploits previously interconnected resources in a way that enables the study and use of structured representations of salient semantic and syntactic properties as realised in the hierarchical verb lexicon, the validation of semantic and syntactic generalisations derived from each of these

resources against the data encoded in the other resources, the mutual enhancement and the expansion of coverage through generalisations over combinations of features of the different resources.

2. Linguistic prerequisites

The creation and mutual integration of complementary lexical semantic resources have presented a great interest in the research community. Notable efforts on mapping semantic resources include the work of Shi and Mihalcea on mapping WordNet, FrameNet and VerbNet (Shi and Mihalcea, 2005), Tonelli and Pighin's enrichment of FrameNet with WordNet mappings (Tonelli and Pighin, 2009), the system Semlink (Palmer, 2009) that unites these three resources with PropBank, and its follow-up Semlink+ that brings in mapping to Ontonotes (Palmer et al., 2014). While these efforts give rise to databases of integrated semantic knowledge, most of them deal with mapping of the units of the original resources to each other. Much less attention has been paid to exploring and exploiting the internal structure of these resources, especially with respect to how these structures relate and correspond to each other, how they can be mapped to each other, etc.

WordNet (Miller, 1995; Fellbaum, 1998b) is a large lexical database which represents conceptual and lexical knowledge in the form of a network whose nodes represent cognitive synonyms (synsets) interlinked through a number of conceptual-semantic and lexical relations. While the relations of hypernymy/hyponymy that provide the basic hierarchical organisation of verbs and nouns in WordNet are explicit, the membership of a synset to a hypernym tree only gives a very general idea about the semantic class to which this synset's members belong. Consider the tree dominated by the root *change:1*, *alter:1*, *modify:3* (cause to change; make different; cause a transformation). We can only infer that the subordinate members of the tree denote some kind of change brought about by an entity and affecting another entity regardless of the depth of the hierarchy at which a particular synset is found.

On the other hand, more detailed classificatory features emerge from the mappings with VerbNet verb classes and superclasses and FrameNet frames. Both resources provide linguistic abstractions that either specify or translate into semantic classes of different granularity and different level of generalisation and in addition propose a certain hierarchical organisation.

FrameNet frames represent conceptual structures describing particular types of objects, situations, etc. along with their participants, or frame elements, FEs (Baker et al., 1998; Ruppenhofer et al., 2016). As such, they are abstract representations of lexical units that lexicalise these situations or objects. Though not exhaustively, FrameNet frames are related into a network through frame-to-frame relations part of which also provide a hierarchical organisation, the most prominent being Inheritance in which the child frame is a subtype of the parent frame, e.g. *Change_of_temperature* and *Proliferating_in_number* inherit from *Change_position_on_a_scale*. Other relations include Using, Perspective on, Subframe, Precedes, Inchoative_of, Causative_of, etc.

We adopt the FrameNet frames as the primary categories of the proposed classification which are to be further explored and enhanced. The hierarchical relationships between these categories are projected both from (i) the hypernymy (troponymy) relation in WordNet and (ii) the hierarchies formed by 2 frame-to-frame relations in FrameNet. As an illustration of (i), consider the hypernym-hyponym pair: *change:1*, *alter:1*, *modify:3* > *heat:1*, *heat up:2* where the verbs are mapped to the frames *Cause_change* and *Cause_temperature_change*, respectively, which are adopted as classification categories. Given that, we posit a relation of hierarchy between the frame-derived classification categories: *Cause_change* > *Cause_temperature_change*. The FrameNet hierarchies are employed in augmenting the coverage of the mapping between frames and synsets as described in detail in Section 5.

The VerbNet (Kipper-Schuler, 2005) classes, which represent explicit natural groupings of verbs with shared semantic and syntactic properties, are structured in a shallow hierarchy of types (herein called superclasses), classes and subclasses (if any). Superclasses unite classes related to a particular type of eventualities, e.g. 'Verbs of putting', 'Verbs of removing', and provide semantically grounded linguistic generalisations. They are employed in addition to the semantic information derived from the FrameNet frames to support the mapping between FrameNet and WordNet, as well as to help resolving ambiguities and inconsistencies in the classification.

3. Related work and motivation of the proposal

While remaining a less explored area, FrameNet’s frame structure and frame-to-frame relations has been employed in various domains, such as text understanding (Fillmore and Baker, 2001), semantic analysis (Burchardt et al., 2005), generation of lexical entailment rules (Coyne and Rambow, 2009; Aharon et al., 2010), paraphrase extraction (Hasegawa et al., 2011), construction of event ontologies (Palmer et al., 2014), role linking of implicit semantic arguments (Li et al., 2015).

Research into the enhancement of frame-to-frame relations has been proposed by a number of studies. Pennacchiotti and Wirth (2009) offer a definition of the notion of frame relatedness and different types of automatic measures to compute it. Ovchinnikova et al. (2010) describe a methodology for improving FrameNet’s conceptual organisation through restructuring and axiomatisation of the frame relations. Frame relations have also been used in augmenting FrameNet’s coverage with paraphrases (Rastogi and Van Durme, 2014).

Extension of frame relations has been another emerging area. Virk et al. (2016) propose a supervised model for enriching FrameNet’s relational structure through predicting new frame-to-frame relations using structural features from the existing FrameNet network, information from the WordNet relations between synsets, and corpus-collected lexical associations. In devising the WordNet features, the authors employ similar logic to the one adopted in this paper by transferring the relational knowledge for pairs of related synsets to matching lexical units and frames in FrameNet. Botschen et al. (2017) present systems for predicting frame-to-frame relations based on text-based frame embeddings; the best-performing one uses the FrameNet hierarchy.

The research into verb classifications in terms of verbs’ syntactic properties and behaviour (Levin, 1993; Pinker, 1989; Goldberg, 1994), among others, thematic structure (Chafe, 1970; Cook, 1979; Longacre, 1976; Foley and Van Valin, 1984; Van Valin Jr. and LaPolla, 1997), lexical conceptual structure (Gruber, 1965; Jackendoff, 1990), frame semantics (Fillmore, 1982) has culminated in resources such as VerbNet, FrameNet and WordNet and subsequent efforts at linking them in such a way as to maximise the merits of each resource (see Section 1).

Other researchers have brought together the ideas of a detailed description of verb semantic classes and a hierarchical classification of these classes. Hlaváčková et al. (2009) have proposed an outline of a verb ontology based on the data in the verb valency dictionary for Czech VerbaLex (Hlaváčková and Horák, 2005). A shallow hierarchy of Spanish verbs based on their semantic and syntactic properties has been described and implemented within the ADESSE project (García-Miguel and Albertuz, 2005).

The classification we propose aims at harnessing these already existing and interlinked resources while trying to bridge the divide between their use as a source of knowledge about the lexical items and natural semantic groupings and their combined potential to explore semantic and syntactic properties and generalisations, the distribution of these properties in the lexicon, the relations between these features, and so forth.

The classification combines information from the 3 outlined resources, each of which contains diverse lexical, semantic and syntactic information: detailed conceptual structures of lexical units with shared semantic and syntactic properties organised through frame-to-frame relations (FrameNet); more explicit verb semantics paired with a detailed representation of syntactic behaviour and linking between semantics and syntax (VerbNet); a hierarchical organisation of word senses within semantic domains connected in a lexicon-wide coverage network through a variety of conceptual and lexical relations (WordNet).

Due to its features, which are essential for the structuring of the proposed classification, WordNet is used as the base resource, providing the classification’s backbone. FrameNet frames are used as the classification categories for the grouping of WordNet verbs into semantic classes and their taxonomic organisation. The application of VerbNet superclasses lends a new dimension to the mapping among the resources and enables the direct exploitation of knowledge from the other two resources.

4. Mapping the resources

The efforts at linking lexical resources generally suffer from limited coverage and compatibility issues due to multiple release versions of the original resources. This reduces considerably their applicability and further development. A feature of the proposed classification is that it attempts at resolving this shortcoming by translating semantically salient groupings into classification categories, which, if not as exhaustive as appropriately assigned frames or verb classes, provide feasible semantic generalisations.

The mappings and mapping procedures implemented have been adopted from the works of Baker and Fellbaum (2009), Shi and Mihalcea (2005) and Laparra and Rigau (2010).

VerbNet 3.3¹ provides an integrated mapping between members of verb classes and WordNet literals in synsets with corresponding senses. Using this direct mapping we have assigned verb classes to 4,885 out of 14,206 verb synsets in WordNet 3.0. Verb classes are further combined into superclasses (cf. VerbNet Annotation Guidelines²) which add a more abstract semantically grounded level of description. The membership to a particular superclass is indicated by a common class number: thus, 'Verbs of ingesting', assigned the number 39 include the classes: eat-39.1, chew-39.2, gobble-39.3, devour-39.4, dine-39.5, gorge-39.6, feeding-39.7. We have mapped the superclass names proposed in the Guidelines to the respective number and assigned them to the corresponding synsets.

With respect to the linking with FrameNet, we have employed two existing mappings. The Sem-Link project³ distributes a many-to-many mapping of VerbNet classes and FrameNet frames where a FrameNet lexical unit may be mapped to more than one VerbNet verbs, and more often, a VerbNet verb is mapped to more than one frame. We use this mapping to assign frames to synsets indirectly by using the FrameNet to VerbNet mapping first, and then the VerbNet to WordNet mapping. In order to increase the limited coverage (only 2,630 verb synsets are assigned frames from FrameNet), a direct FrameNet-to-WordNet mapping called WordFrameNet (Laparra and Rigau, 2010)⁴ is also applied using a set of intermediate mappings between release versions. As a result the number of verb synsets with assigned frames has increased to 3,134. WordFrameNet also provides frame assignments to over 4,000 words of other parts of speech (adjectives and nouns).

5. Building the classification

We have devised a number of procedures for the augmentation of the mapping coverage, which in the scope of the classification translates as increasing the number of classification categories and extending the coverage of the individual classification categories.

These procedures are aimed at: (i) discovering existing but unmapped relations between synset members and FrameNet frames; (ii) transferring frames between synsets through relations of inheritance derived from WordNet and FrameNet; (iii) adopting additional classification categories from VerbNet.

5.1. Extending the coverage using FrameNet frames

Nouns and verbs in WordNet are classified into distinct semantic domains that roughly correspond to semantic primitives, such as person, artifact, act for nouns or change, motion, cognition for verbs. This information is made explicit through special synset labels, such as noun.person, noun.artifact, noun.act, verb.change, verb.motion, verb.cognition. There are 25 such categories for nouns and 15 for verbs (Miller, 1998). Unlike nouns which form several hierarchies, each originating from a different unique beginner (Miller, 1998), verbs pertaining to a given semantic domain do not belong to a single hierarchy, but often represent several independent trees (Fellbaum, 1998a). The picture is more complex in reality as WordNet 3.0 has 559 verbs with no hypernym, of which 225 have no hyponyms (single verbs) and only 254 have more than 1 direct or indirect hyponym (Richens, 2008).

A tree often does not consist entirely of synsets that have the same semantic primitive, but one primitive is clearly predominant. As a very preliminary approximation at a viable classification criterion,

¹<http://verbs.colorado.edu/verbnet/index.html>

²https://verbs.colorado.edu/verb-index/VerbNet_Guidelines.pdf

³<https://verbs.colorado.edu/semlink/>

⁴<http://adimen.si.ehu.es/web/WordFrameNet>

we have sorted the trees according to the predominant prime.

As the frames of the beginners are (the most) abstract nodes in a particular hierarchy and their properties are inherited by the subordinate nodes, it is important that we: (i) identify the beginners within each semantic domain; (ii) check if they are mapped to frames, correct wrong mappings and assign a frame where the automatic procedure has failed; (iii) attach those roots that do not qualify as beginners to a given hierarchy.

5.1.1. Identifying the beginners in each semantic domain

There is no explicit information which root verbs represent the right level of abstraction so as to qualify as beginners of verb hierarchies, but the best candidates are those that are homonymous or near-synonymous with the semantic primes assigned to them and/or have a multitude of hyponyms, e.g., *change:2* (undergo a change...) and *change:1*, *alter:1*, *modify:3* (cause to change) (semantic prime: *verb.change*). Beginners are or should be assigned a FrameNet frame that is high in the frame-to-frame hierarchy. The second-best candidates are other trees of considerable size (in terms of number of members) and level of abstraction (in terms of the FrameNet frames assigned). If not identified as beginners, they should be attached to the tree of a beginner, preferably one with the same dominant prime. Single verbs and verb roots with a couple of hyponyms tend to be lower in the conceptual hierarchy and should be attached to a larger tree.

5.1.2. Checking of the mapping and manual assignment of frames

We have checked manually the mapping of the root verbs to FrameNet frames, focusing on the beginners as errors tend to propagate down the tree. For instance, during the automatic procedure described in Section 4 *change:2* was mapped to the *Cause_change* frame instead of *Undergo_change* and so, if the mapping was left uncorrected, all its hyponyms would be classified in a wrong hierarchy.

The procedure for validation and assignment of frames to verb roots includes the following steps: (i) check the correctness of the assigned frame(s) if any – at this point the frames are assigned from the existing mapping distributions (cf. Section 4); if no frame is assigned (ii) check the frames in which the literals of the synset in question are found as lexical units and select an appropriate one; if no (appropriate) match is found, (iii) try to assign a suitable frame to the given synset using other information.

In particular, step (i) has resulted in the manual check of the frames of 1,300 synsets which includes: (a) frames that in the course of the analysis of the data have been found to be erroneously assigned; these have been individually corrected (total of 139); (b) frames which have been consistently assigned instead of other semantically related frames, e.g. *Attaching* covers examples of the frames *Attaching*, *Becoming_attached* and *Detaching*; all the synsets to which such frames have been automatically assigned have been analysed and priority has been given to the most suitable frame (total of 373 examples); and (c) synsets which have been assigned more than one frame; in such cases, either one of them has been selected or corrected, or in case more than one frame is considered suitable, they have been ranked (total of 788 examples).

In step (ii) and step (iii) the process has been facilitated by exploring semantic and structural features, such as:

(a) the frames of the synset's hyponyms – this is a check if a hyponym's frame (especially one predominant among the hyponyms) describes appropriately the conceptual structure of the verb synset in question. An additional verification procedure keeps track of whether verbs are consistently assigned causative or inchoative frames depending on their membership to one or the other class. This is done by verifying that in a tree dominated by a causative root the causative counterparts of homophonous or similar pairs of frames are assigned; and respectively, that in an inchoative tree the inchoative counterparts are mapped. This procedure is aimed at removing the errors in the automatic mapping;

(b) the mapping of the VerbNet (super)class to FrameNet – check if a frame homonymous or similar to the (super)class of a given synset describes appropriately the conceptual structure of the verb;

(c) the gloss – check whether the generic term which is elaborated in the definition is or may be mapped to a frame and if this frame describes appropriately the conceptual structure of the verb;

(d) structural features – these are based on the relations of the root synset to other verbs or nouns. In the former case, we check if the FrameNet frame of (i) a semantically related verb synset, such as an

antonym or a causative/inchoative counterpart, or (ii) a direct or an indirect hypernym of such a synset describes appropriately the verb in question. Consider, for instance, *back:4* (cause to travel backward), which is an antonym of *advance:5*, *bring forward:1* (cause to move forward). The latter's hypernym is *move:2*, *displace:4* (cause to move or shift into a new position or place, both in a concrete and in an abstract sense) which is mapped to the frame *Cause_motion*. *Cause_motion* is thus suggested as a mapping of *back:4*, as well.

Another strategy employs the connectedness of nouns in WordNet (all the nouns except for the root (entity:1) have at least one hypernym) and the relations between eventive deverbal nouns and derivationally related verbs. Thus, given the Event relation between *fall:17* (lose one's chastity) and *fall:5* (a lapse into sin...), we establish that the latter's hypernym *sin:2*, *sinning:1* and its Event derivative *sin:1*, *transgress:3*, *trespass:4* are mapped to the frame *Misdeed*, which is also a suitable mapping for *fall:17*.

Finally, if no appropriate frame exists, we posit a new classification category (and a frame) provided that it is predictable from the FrameNet's frame structure. For instance, while *Motion* is linked to *Cause_motion*, *Self_motion* (e.g. *jump:1*, *leap:1*, *bound:1*, *spring:1* – move forward by leaps and bounds) does not have a causative counterpart to which verbs such as *jump:1*, *leap:4* (cause to jump or leap) can be mapped, so we formulate one.

5.1.3. Attaching non-beginner roots into the hierarchy (hypernym assignment)

It is most likely to find both the beginners and suitable positions for attachment of smaller trees into a particular hierarchy among synsets that pertain to the same semantic domain (semantic prime) as the one we explore. Hypernym assignment involves techniques similar to the ones presented in Section 5.1.2. with certain differences. Thus we look at:

- (a) the frame of the synset – check if a verb homonymous or similar to the frame assigned to a given synset or other verbs evoking this frame represent a suitable hypernym;
- (b) the mapping of the synset to a VerbNet (super)class – check if a verb homonymous or similar to the assigned (super)class is a suitable hypernym;
- (c) the gloss – check whether the generic term elaborated in the definition is a suitable hypernym;
- (d) structural features – these are more or less the same as in (d) above, but in this case we check if a semantically related verb (a hypernym of an antonym, a derivative of a derivative's hypernym, etc.) is a suitable hypernym; these features also involve analysing a tree structure more globally.

Consider the root verb *look:1* (perceive with attention; direct one's gaze towards): it is assigned the frame *Perception_active* and its definition contains as a generic term the verb 'perceive' so an obvious choice of hypernym is *perceive:1*, *comprehend:2* (to become aware of through the senses). The tree structure of the latter synset confirms this choice as among its hyponyms one finds both active perception verbs such as *listen:1* and passive perception verbs such as *see:1* and *hear:1*.

As a result of the procedures in 5.1, incorrect frames assigned to the original set of roots have been replaced by appropriate ones, frames have been assigned to those roots for which the automatic mapping has failed and several new frames have been defined. Most of the original beginner roots have been assigned a FrameNet frame. In addition, those of them that could be successfully mapped to a larger tree have been assigned a suitable hypernym.

5.2. Procedures for automatic classification

We have developed 4 procedures for frame identification (STEPS 1–4) aimed at expanding the frame-to-synset coverage for the purposes of the classification, which we then apply recursively down the WordNet trees. We have also implemented a procedure for ranking candidate frames in order to facilitate their manual verification and selection.

5.2.1. Employing hierarchical relations in WordNet (hypernymy–hyponymy)

FrameNet frames may correspond to different hierarchical levels in the hypernym hierarchy in WordNet which results in the fact that lexical units evoking the same frame may be in a hypernymy relation to each other. For instance, *amble:1*, *totter:2*, *wade:1* are found in hyponyms of *walk:1*, but all of them, including

walk:1, evoke the frame *Self_motion*. This asymmetry in the grouping of lexical units and synsets has been used in mapping procedures, cf. Tonelli and Pighin (2009).

We therefore assume that, by default, a hyponym synset inherits the frame(s) of the hypernym synset. So as a first approximation, given a hypernym–hyponym pair, we consider the set of all the frames assigned to the hypernym and transfer them to the hyponym (STEP 1).

We also employ the hierarchical relations in FrameNet (frame-to-frame inheritance); more particularly, we make use of 2 hierarchical frame-to-frame relations which to a different degree correspond to the structure of WordNet. These are: *Is_Inherited_by* – the child frame is a subtype of the parent frame (e.g. *Motion* *Is_Inherited_by* *Fluidic_motion*); *Is_Used_by* – the child frame presupposes the parent frame as background (e.g. *Theft* *Is_Used_by* *Robbery*).

These two relations may be construed as potentially corresponding to the hypernymy relation: for instance, the lexical unit ‘influence’ evokes the frame *Objective_influence* which is inherited by the frame *Manipulate_into_doing*, and the latter is evoked by the lexical unit ‘manipulate’. In WordNet the synset *influence:1, act upon:1, work:15* (have and exert influence or effect) is the hypernym of *manipulate:1, pull strings:1, pull wires:1* (influence or control shrewdly or deviously). Similarly, the lexical unit ‘arrive’ and ‘reach’ evoke the frame *Arriving* which is in the relation *Is_Used_by* with the frame *Having_or_lacking_access*, and the latter is evoked by the lexical unit ‘access’, among others. In the corresponding synset pair *reach:1, make:22, attain:4, hit:4, arrive at:1, gain:4* (reach a destination, either real or abstract) is the hypernym of *access:2, get at:1* (reach or gain access to).

Based on the above considerations, given a synset *S*, we select as a mapping of *S* any frame *F* if *S* contains a literal *L* coinciding with a lexical unit *LU* in FrameNet, such that *LU* evokes the frame *F1*, where *F1* is assigned to the hypernym of *S* and *F1* is related to *F* through one of the following set of frame-to-frame relations $R\{\textit{Is_Inherited_by}, \textit{Is_Used_by}\}$ (STEP 2).

5.2.2. Combining hierarchical relations in WordNet and FrameNet

We also apply the reverse operation up the hypernym-hyponym tree so as to identify frames that generalise over the frames assigned to a synset’s hyponyms. Thus, for each synset *S* with no frame assigned, we add to the mapping any frame *F* such that there is a synset *S1* which is a direct hyponym of *S* and *S1* is assigned the frame *F1* where *F1* is a direct descendant of *F* (STEP 3).

5.2.3. Employing VerbNet superclasses to identify possible frame candidates

We use the VerbNet to FrameNet mapping to identify indirectly frames to be assigned as classification categories to synsets based on the correspondences between the VerbNet superclasses and FrameNet frames. For each synset *S* with an assigned VerbNet class and thus, superclass *C*, but no frame assigned, we add to the mapping any frame *F* such that there is a synset *S1* which has been assigned both superclass *C* and frame *F* (STEP 4).

5.2.4. Ranking of candidate frames

After the implementation of the above procedures, we apply ranking to all new candidate frames in order to identify the most likely frames and to facilitate their manual verification and selection. The scoring system reflects the source of the candidate: score ranges indicate the procedure via which the frame has been assigned, including information whether the assignment has been made through the application of more than one procedure, which increases the probability of a candidate being a valid suggestion as a classification category for that synset. The actual ranking then combines these into a single score and orders the candidates in decreasing order.

The ranking is based on the following factors:

- (i) Priority is given to direct over indirect inheritance of frames from hypernym to hyponyms;
- (ii) The score of frames assigned by more than one procedure is a sum of the individual scores given to the frame individually via each procedure;
- (iii) Frame score is adjusted with respect to the collective frequency of the literals belonging to the respective synset found in the candidate frame’s lexical units list.

Example 1. presents the result of the assignment and ranking of frames to the synset *spy:3, sight:1*.

Example 1. eng-30-02163746-v spy:3, sight:1 'catch sight of; to perceive with the eyes'

WordNet semantic prime: verb.perception

VerbNet class: sight-30.2 (Verbs of Perception)

Ranking of possible FrameNet frames in descending order: Perception (104); Becoming_aware (100); Awareness (100); Perception_experience (100); Categorization (100); Sensation (20); Perception_active (20); Intentionally_act (1);

The similar scores of Perception, Becoming_aware, Awareness and Perception_experience reflect the similarity among these conceptual structures in the FrameNet hierarchy: Perception is inherited by Becoming_aware and Perception_experience. The values of these particular frames correspond very neatly to the WordNet structure, as well: perceive:1, comprehend:2, the hypernym of spy:3, sight:1, is mapped to Perception_experience and most of the hyponyms of spy:3, sight:1 are mapped to Becoming_aware. These two frames are therefore very likely candidates.

In addition, while Awareness and Becoming_aware are not linked through a relation, it is obvious that such a relation needs to be specified as Becoming_aware (referring to a Cogniser's adding a Phenomenon to their model of the world) is the inchoative counterpart of the stative Awareness (referring to a Cogniser's having a piece of Content in their model of the world)⁵, so this merits further analysis.

6. Classification of verbs in WordNet

6.1. Resulting classification of verbs in WordNet

Using the above procedures, we obtain a detailed classification of verbs in WordNet based on a number of distinctive features: (a) verb semantic primes; (b) FrameNet frame(s); (c) WordNet hierarchical relations; and (d) VerbNet classes and superclasses. The classification categories (FrameNet frames) are assigned within each general semantic category of verbs (as defined by their semantic prime) and form a hierarchy determined by the WordNet hypernymy/hyponymy hierarchy (where the respective synsets have been assigned FrameNet frames in the existing mappings between the resources, cf. Section 4) in conjunction with the hierarchy of FrameNet frames based on the Is_Inherited_by / Inherits_from and Is_Used_by / Uses relations and the VerbNet (super)class assignments to fill the gaps in the cases where frames have not been assigned to synsets in the existing mappings. These procedures align well, as the inheritance between frames roughly follows the route from more general to more specific conceptual structures, which corresponds to the branching of the WordNet hierarchy from more general to more specific senses.

Initially, there are 3,134 classified synsets (i.e synsets assigned a frame as a classification category through the mapping procedure described in Section 4) out of a total of 14,206 verb synsets in WordNet. Using the algorithm in Section 5, we have been able to assign at least one classification category derived from the inventory of FrameNet frames to 10,331 synsets. At STEP 1, by transferring frames from the hypernym to the hyponym, 10,217 candidate frames have been assigned. Another 1,192 candidate frames have been assigned at STEP 2, and further 1,311 frames have been mapped at STEP 3. More than half of the synsets (5,606) are assigned more than one candidate frame, about a third (3,012) – more than two candidate frames, and about 10% (981) – more than five candidate frames.

The candidate frames for each synset obtained through the different procedures are ranked according to their scores in descending order. A total of 1,210 candidate frames have been assigned using more than one STEP rendering them more probable and bringing them forward in the ranking. The results are then subjected to manual validation so as to ensure high quality of the assignments, including through removing errors stemming from the automatic FrameNet to WordNet mapping. The classification is available at: <http://dcl.bas.bg/en/verb-classification/>.

6.2. Towards the validation and streamlining of the classification

Proper automatic validation of the method is not possible as there is no gold standard against which to evaluate the classification. We propose partial validation of the resulting structure against the relational

⁵cf. <https://framenet2.icsi.berkeley.edu/>

structure of FrameNet. The comparison is focused on the consistency between the hierarchy of categories (frames) automatically built from the relational structure of WordNet and the hierarchy on which FrameNet itself is constructed. Consistency is measured in terms of identical or near-identical paths in the two hierarchies according to a particular relation: the hypernymy/hyponymy relation in WordNet and the Inheritance and Uses relations (and their combinations) in FrameNet. Each category is evaluated in relation to its direct parent in the hierarchy and is labeled either: (a) OK – the direction of inheritance of the frames coincides with the direction of the hypernym-to-hyponym relation in WordNet (Example 2a); (b) UPPER – the direction of inheritance of the frames is reversed (Example 2b); or (c) DIFFERENT – the frame appears elsewhere in the FrameNet hierarchy (Example 2c).

Ideally, the two structures coincide (with possible omissions or additions of nodes in either of them) which means that for a pair of categories in the classification where F1 (evoked by synset S1) is a parent of F2 (evoked by synset S2) and S1 is a hypernym of S2, F1 is a direct or an indirect parent of F2 in FrameNet (F1 *Is_Inherited_by* / *Is_Used_by* F2).

Example 2.

(a) OK

eng-30-01835496-v travel:3; go:16; (verb.motion, Motion) is hypernym of
eng-30-01886488-v slither:1; slide:3 (verb.motion, Self_motion)
and the frame Motion *Is_Inherited_by* the frame Self_motion

(b) UPPER

eng-30-01463963-v arrange:4; set up:5 (verb.contact, Arranging) is hypernym of
eng-30-01543000-v drape:2 (verb.contact, Placing)
and the frame Placing *Is_Inherited_by* the frame Arranging

(c) DIFFERENT

eng-30-00983824-v utter:4; emit:2 (verb.communication, Communication_noise) is hypernym of
eng-30-01197208-v smack:4 (verb.consumption, Body_movement)
and the frame Communication_noise is not related to the frame Body_movement neither directly nor indirectly via the *Is_Inherited_by* / *Inherits_from* or *Is_Used_by* / *Uses* relation.

Deviations from the ideal as presented in (b) and (c) may be indicative of: (i) different logic underlying the hierarchical organisation of the conceptual structures and the synsets; (ii) inconsistencies or errors in the frame assignment or in the WordNet structure, e.g. mingling transitive/intransitive synsets in a single subtree or inconsistencies in the semantic prime and/or the WordNet hierarchy (as in (c)).

Further, in order to deal with the inconsistencies, an intermediate level of abstraction may be introduced in the hierarchy based on the (super)classes in VerbNet. This information can be used to combine heterogeneous examples in more general categories and to further ensure the validity of the classification.

7. Future development

An immediate task to be pursued is to identify inconsistencies in the data through exploring the inheritance among classification categories. Future work will be focused on elaborating techniques for increasing the depth of the classification hierarchy through maximising the weight of the similarity between the WordNet hypernym-hyponym structure and the FrameNet frame structure, through correlating VerbNet (super)classes with frames in such a way as to reflect the degree of semantic generality/specificity, etc.

Another venue of research is extending FrameNet’s hierarchy with new frame relations and relation instances identified in the proposed WordNet classification, as well as mutually enhancing the two hierarchies on the basis of linguistic information retrievable from them.

Acknowledgements

This study has been undertaken within the project *Towards a Semantic Network Enriched with a Variety of Semantic Relations* funded by the National Scientific Fund of the Republic of Bulgaria under the Fundamental Scientific Research Programme (Grant Agreement No. 10/3/2016). We would like to thank three anonymous reviewers for their valuable comments.

References

- Aharon, R. B., Szpektor, I., and Dagan, I. (2010). Generating Entailment Rules from FrameNet. In *Proceedings of the ACL 2010 Conference Short Papers, Uppsala, Sweden, 11-16 July 2010*, pages 241–246. Association for Computational Linguistics, Stroudsburg, PA, USA.
- Baker, C. F. and Fellbaum, C. (2009). WordNet and FrameNet as Complementary Resources for Annotation. In *Proceedings of the Third Linguistic Annotation Workshop (ACL-IJCNLP '09), Association for Computational Linguistics, Stroudsburg, PA, USA*, pages 125–129.
- Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference. Montreal, Canada*, pages 86–90.
- Botschen, T., Mousselly-Sergieh, H., and Gurevych, I. (2017). Prediction of Frame-to-Frame Relations in the FrameNet Hierarchy with Frame Embeddings. In *Proceedings of the 2nd Workshop on Representation Learning for NLP, Vancouver, Canada, August 3, 2017*, pages 146–156. Association for Computational Linguistics, Stroudsburg, PA, USA.
- Burchardt, A., Frank, A., and Pinkal, M. (2005). Building Text Meaning Representations from Contextually Related Frames – A Case Study. In *Proceedings of the Sixth International Workshop on Computational Semantics*.
- Chafe, W. L. (1970). *Meaning and the Structure of Language*. University Press, Chicago.
- Cook, W. A. (1979). *Case Grammar: Development of the Matrix Model (1979-1978)*. Georgetown University Press.
- Coyne, R. and Rambow, O. (2009). Lexpar: A Freely Available English Paraphrase Lexicon Automatically Extracted from FrameNet. In *Proceedings of the Third IEEE International Conference on Semantic Computing*.
- Fellbaum, C. (1998a). A Semantic Network of English Verbs. In Fellbaum, C., Ed., *WordNet: An Electronic Lexical Database*, pages 69–104. MIT Press, Cambridge, MA.
- Fellbaum, C., Ed. (1998b). *WordNet: an Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Fillmore, C. J. and Baker, C. F. (2001). Frame Semantics for Text Understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*.
- Fillmore, C. J. (1982). Frame Semantics. In *Linguistics in the Morning Calm*. The Linguistic Society of Korea, Seoul: Hanshin.
- Foley, W. and Van Valin, R. (1984). *Functional syntax and Universal grammar*. CUP.
- García-Miguel, J. M. and Albertuz, F. J. (2005). Verbs, Semantic Classes and Semantic Roles in the ADESSE project. In Erk, K., Melinger, A., and Schulte im Walde, S., Eds., *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes, Saarbrücken, 28 February – 1 March 2005*.
- Goldberg, A. (1994). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.
- Gruber, J. (1965). *Studies in Lexical Relations, Doctoral dissertation, MIT, Cambridge, MA*. Also published in J. Gruber (1976) *Lexical Structures in Syntax and Semantics*, North-Holland, Amsterdam, 1–210.
- Hasegawa, Y., Lee-Goldman, R., Kong, A., and Akita, K. (2011). FrameNet as a Resource for Paraphrase Research. *Constructions and Frames*, 3:104–127.
- Hlaváčková, D. and Horák, A. (2005). VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In *Proceedings of the Computer Treatment of Slavic and East European Languages 2005, Bratislava, Slovakia*, pages 107–115.
- Hlaváčková, D., Khokhlova, M., and Pala, K. (2009). Semantic Classes of Czech Verbs. In *Recent Advances in Intelligent Information Systems*, pages 207–217.
- Jackendoff, R. S. (1990). *Semantic Structures*. Cambridge, Mass., The MIT Press.
- Kipper-Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon. PhD Thesis*. Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA.

- Laparra, E. and Rigau, G. (2010). eXtended WordFrameNet. In *Proceedings of LREC 2010*, pages 1214–1219.
- Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago and London: The University of Chicago Press.
- Li, R., Wu, J., Wang, Z., and Chai, Q. (2015). Implicit Role Linking on Chinese Discourse: Exploiting Explicit Roles and Frame-to-frame Relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China, July, 2015, pages 1263–1271. Association for Computational Linguistics.
- Longacre, R. E. (1976). *An Anatomy of Speech Notions*. Peter de Ridder Press.
- Miller, G. A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41.
- Miller, G. A. (1998). Nouns in Wordnet. In Fellbaum, C., Ed., *WordNet: An Electronic Lexical Database*, pages 21–46. MIT Press, Cambridge, MA.
- Ovchinnikova, E., Vieu, L., Oltramari, R., Borgo, S., and Alex, T. (2010). Data-driven and Ontological Analysis of FrameNet for Natural Language Reasoning. In *Proceedings of LREC 2010*.
- Palmer, M., Bonial, C., and McCarthy, D. (2014). SemLink+: FrameNet, VerbNet and Event Ontologies. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929–2014)*, Baltimore, Maryland USA, June 27, 2014, pages 13–17. Association for Computational Linguistics.
- Palmer, M. (2009). Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*. 9–15.
- Pennacchiotti, M. and Wirth, M. (2009). Measuring Frame Relatedness. In Lascarides, A., Gardent, C., and Nivre, J., Eds., *Proceedings of EACL 2009*, pages 657–665. Association for Computer Linguistics.
- Pinker, S. (1989). *Learnability and Cognition: The acquisition of argument structure*. MIT Press.
- Rastogi, P. and Van Durme, B. (2014). Augmenting FrameNet Via PPDB. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, Baltimore, Maryland, USA, pages 1–5. Association for Computational Linguistics.
- Richens, T. (2008). Anomalies in the WordNet Verb Hierarchy. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester, August 2008, pages 729–736.
- Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Baker, C. F., and Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California.
- Shi, L. and Mihalcea, R. (2005). Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In Gelbukh, A., Ed., *Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science*, volume 3406. Springer, Berlin, Heidelberg.
- Tonelli, S. and Pighin, D. (2009). New Features for Framenet – Wordnet Mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL’09)*, Boulder, USA.
- Van Valin Jr., R. D. and LaPolla, R. J. (1997). *Syntax: Structure, meaning and function*. Cambridge University Press.
- Virk, S. M., Muller, P., and Conrath, J. (2016). A Supervised Approach for Enriching the Relational Structure of Frame Semantics in FrameNet. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, pages 3542–3552.

A Pilot Study for Enriching the Romanian WordNet with Medical Terms

Maria Mitrofan
ICIA
Romanian Academy
maria@racai.ro

Verginica Barbu Mititelu
ICIA
Romanian Academy
vergi@racai.ro

Grigorina Mitrofan
National Institute of Diabetes
and Metabolic Diseases
mitrofan.grigorina@gmail.com

Abstract

This paper presents the preliminary investigations in the process of integrating a specialized vocabulary, namely medical terminology, into the Romanian wordnet. We focus here on four classes from this vocabulary: anatomy (or body parts), disorders, medical procedures and chemicals. In this pilot study we selected two large concepts from each class and created the Romanian terminological (sub)trees for each of them, starting from a medical thesaurus (SNOMED CT) and translating the terms, process which raised various challenges, all of them asking for the expertise of a specialist in the health care domain. The integration of these (sub)trees in the Romanian wordnet also required careful decision making, given the structural differences between a wordnet and a terminological thesaurus. They are presented and discussed herein.

1. Introduction

The experiment presented here is to be understood within the larger context of medical terms identification, automatic extraction and classification of relations between them.

For the automatic identification and classification of the relations between terms occurring in a corpus, we need access to a knowledge source. This can be either a corpus annotated with medical terms and relations between them or a resource containing terms and relations between them. For Romanian, a medical corpus (of more than 100,000 tokens) was annotated with medical entities (Mitrofan, 2017). This corpus contains four top level entity classes (Anatomy, Chemicals and Drugs, Disorders, Procedures) corresponding to UMLS (Unified Medical Language System) semantic groups, but no semantic relation is annotated between these entities.

We present here a pilot study of the steps necessary for the development of an ontology-like resource for Romanian medical terminology. Translating SNOMED CT is a huge effort, which takes too long and also requires funding for finding the professionals to do this. However, taking advantage of the existence of a wordnet for Romanian (see below, section 3.1.1.), we decided to expand its set of medical terms with structured information from SNOMED CT (see below, section 3.1.2.) in the form of subtrees containing medical terms of four semantic groups, corresponding to the ones with which the Romanian medical corpus is annotated: anatomical ones, terms designating disorders, medical procedures and chemicals. Given our interest in the diabetes domain, the terms we selected are specific to it, as shown below. Our aims were to establish a work methodology and to discover the difficulties encountered during the translation process and methods to address them. By no means are the selected terms and all their subtrees (when existent) representative for all possible cases of translation equivalents.

2. Related Work

Even though the current version of Princeton WordNet has a broad coverage of the medical terminology, it has certain issues which reflect both the fact that the WordNet was not built for domain-specific ap-

plications and the need of specialized expertise when compliance with medical terminologies is needed (Bodenreider et al., 2003).

In the literature several reports on enriching wordnets with medical terms are available. Barry and Fellbaum (2004) presented an important initiative in this direction. They attempted to create a free-standing lexical resource designed for natural language processing applications in the medical domain, Medical WordNet (MWN). This lexical database contains medical terms used by non-expert subjects and two sub-corpora, Medical FactNet (MFN) and Medical BeliefNet (MBN). The former contains validated sentences that illustrate medically relevant vocabulary and the latter contains statements which are believed to be true by non-experts.

Recently, another attempt at integrating medical terms from the National Cancer Institute Thesaurus (NCIt) into Open Multilingual Wordnet (OMW) was presented (Hicks et al., 2018). Two methods were investigated: one which automatically calculates mappings between the literals and their definitions in the two resources, and another one, manual, was the analysis of the Princeton WordNet (PWN) v.3.0 coverage of the thesaurus. Both methods involved literals and glosses alike.

Toumouth et al. (2006) addressed the possibility of mapping a general ontology (PWN) to a specific domain (medicine) by extracting 57887 pairs of nouns separated by a conjunction from Ohsumed corpus (W. Hersh and Hickam, 1994) and automatically introducing them into PWN. For the domain-specific words with multiple senses the accuracy was 46% and the recall 51%.

As far as the Romanian wordnet is concerned, we are not aware of any such initiative.

3. The Experiment

This section contains the description of the resources used and the methodology observed in our pilot study of investigating the necessary steps to follow and decisions to be made for an appropriate integration of the medical terms in wordnet, so as to serve the aim of helping in the automatic identification of terms and of relations between them in a medical corpus.

3.1. Resources

Two types of resources were used: two wordnets and a hierarchical thesaurus of medical terms. One wordnet is the Romanian one, which is to undergo enrichment. The other one was the English wordnet: it is richer than the former and, consequently, it can be referred to when drawing a comparison between common language hierarchies and the domain-specific ones (see below, section 4.2.).

3.1.1. Wordnets

Princeton WordNet

This lexical resource (Fellbaum, 1998) contains words mainly from the general language, although some lexical items specific to various domains do occur in it: consider the literal *oxidized LDL cholesterol*, specific to the medical domain, or *levodopa* from pharmacology. The words are nouns, verbs, adjectives and adverbs, each class with its own internal organization: nouns and verbs are distributed in hierarchies, descriptive adjectives are organized in bipolar clusters, while adverbs and relational adjectives lack an organization. For the present study we are interested only in the noun component of this resource.

PWN is a semantic network in which the synonymy relation between word senses is the condition for grouping them together in the nodes of the network: the nodes are called *synsets*, as they represent sets of synonymous words. The structure of the network of nouns is given by the semantic relation of hyperonymy between word senses in the nodes¹. One hyperonym can have several hyponyms and a hyponym can also have more than one hyperonym.

The Romanian WordNet

The wordnet for Romanian (RoWN) (Tufiş and Barbu Mititelu, 2014) has been developed following the expand method, which implied its alignment to PWN, that is importing the structure and creating the Romanian equivalents of the English synsets. At the moment, RoWN is aligned to PWN v.3.0 and contains almost 60,000 synsets.

¹Another relation linking the noun synsets is meronymy, with scarce interest for the present study.

3.1.2. Thesaurus

SNOMED CT

Over the past decades the need for common, controlled, sharable and reusable medical terminologies has been widely recognized. Therefore, there is a growing interest in comprehensive medical terminologies that allow consistent ways of storing, indexing, and retrieving medical data. In order to fulfill these requirements several concept representation systems were created. For example, Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) is a comprehensive clinical terminology covering procedures, clinical findings and diseases, which was created by health-care specialists in order to structure and to reduce the variability of the way medical data is used. SNOMED CT is a multiaxial nomenclature that contains more than 1 million distinct medical terms, 326,734 active concepts (2017 release), 19 upper level hierarchies ².

A SNOMED CT concept is described by a unique name, a unique numeric code and descriptions (one more frequent term and one or more synonyms). SNOMED CT is officially released in English³ and Spanish⁴. A UMLS License is required to access SNOMED CT and to use the UMLS SNOMED CT Browser. Free of charge accounts are possible for scientific research in medical informatics.

3.2. Methodology

Making a pilot study, we did not target a complete coverage of the medical domain or at least of the diabetes subdomain. We worked with medical terms from four large categories extracted from the UMLS semantic groups, which are important for our further work: body parts (anatomy), disorders, procedures and chemicals. These terms are: body parts: *pancreas* (En. *pancreas*), *glandă endocrină* (En. *endocrine gland*); disorders: *boală metabolică* (En. *metabolic disorder*), *nefropatie diabetică* (En. *diabetic nephropathy*); procedures: *glicemie* (En. *blood glucose*), *hemoleucogramă* (En. *complete blood count*); chemicals: *colesterol* (En. *cholesterol*), *sânge* (En. *blood*).

We started from SNOMED CT, so from the English terms, and we translated them and all their direct and indirect hyponyms into Romanian. Given that the terms involved are not easily understood by a non-specialist, we had a physician assist the translation. Whenever a series of terms can be used as synonyms, i.e. to refer to the same concept, they are all added to the same node, in random order.

In order to enhance the RoWN with medical terminology, a translation methodology should be taken into consideration (Reynoso et al., 2000), because very often a term-to-term translation is either not possible or inappropriate.

This section is not meant as a thorough presentation of the challenges of terms translations. That is why we discuss only several of the problems we confronted with when translating the SNOMED CT selected terms.

In general, in domains where discoveries and innovations easily spread from one language community to another, specialized lexicons often incorporate new terms by borrowing them from the language of origin, either for a short, transitional period, during which an indigenous word is created and spreads among the specialists, or for ever. Such borrowing are, in Romanian, *by-pass*, *gastric-sleeve*, *pacemaker*, which are used in the medical community. Even though these terms have Romanian equivalents, they need to be included in the lexicon based on the frequency of their usage. Moreover, some of the borrowed terms are usually adapted to the pronunciation and the morphology of the Romanian language *pic* for *peak*, *șuntare* from *shunt*. So, adaptation should also be taken into consideration in the process of enhancing the RoWN with medical terminology.

In the medical domain, the usage of acronyms is a widely accepted practice. As far as English and Romanian are concerned, some acronyms are used in both languages, although they reflect the term structure only in one of the languages: see the use of *HDL* in both the English *HDL cholesterol* and the Romanian *colesterol HDL*. *HDL* is the acronym for *high-density lipoprotein*, so an English structure. Its

²<https://www.snomed.org/snomed-ct/snomed-ct-worldwide>

³<https://www.snomed.org/about>

⁴<https://confluence.ihtsdotools.org/display/RMT/2017/10/12/October+2017+SNOMED+CT+Spanish+Edition+Beta+release+available+to+Members>

Romanian translation is *lipoproteina cu densitate înaltă*. However, the Romanian medical terminology does not include a term such as *colesterol LDI*, which would be the normal translation of the English term. Instead, only the word is translated and is used with the English acronym, that is *colesterol HDL*. What is more, the most frequently occurring form of the term is, in fact, *HDL colesterol*, which displays a word order that is not specific to Romanian (in which the modifier usually occurs postnominally), but to English (in which the modifiers occur before the modified noun). On the one hand, such examples show that the acronym transfer can be obligatory and must be encoded as a synonym of the term, that is as members of the same synset in wordnet. It can be doubled by the Romanian acronym when existing and being used in the literature. Some other times the terms in each language have their own acronym which is used by specialists and thus the acronym borrowing becomes unnecessary: consider *AIDS* in English and *SIDA* in Romanian, where the English abbreviation is not used. On the other hand, the example shows that structural differences between languages can be overridden in terminology translation.

Turns of phrases occur in term translation: a noun phrase like *120 minute blood glucose measurement* is reorganized: the equivalent of the modifier *blood glucose* becomes the head of the Romanian term and the English head noun *measurement* gets translated as a participle (*măsurată*) modifying this noun: *glucoză măsurată la 120'*.

Such an example is also illustrative of the fact that terms may not always translate in the same way: some of the occurrences of the noun *measurement* in the hyponyms of *blood glucose* are translated by the adjective *măsurată* (e.g., *120 minute blood glucose measurement* translates as *glicemia măsurată la 120 de minute*, etc.), others are translated by the noun *determinarea* (e.g., *capillary blood glucose measurement* translates as *determinarea glicemiei capilare*, etc.).

In conclusion, a well-developed methodology for translating the medical terms is needed mostly because they may be translated incorrectly when following a term-to-term approach: consider also the alteration, be it even slight, of the adjectives order between the corresponding terms *idiopathic transient neonatal hyperinsulinemia* and *hiperinsulinemie idiopatică neonatală tranzitorie*. Usually, more adjectives modifying the same English noun are translated in reverse linear order in Romanian, but it is not the case here.

4. Data Analysis

Data analysis implies discussing to what extent the selected terms and the trees whose roots they are occur in the wordnets and whether their structure is similar or not in the two types of resources, so as to help us to decide on the procedure for their integration in the RoWN.

4.1. Wordnet Coverage of Selected Terms

A first point of interest is the existence of the selected terms in the resource that is to be enriched, namely the RoWN. We notice that three out of the eight terms have not been implemented in RoWN yet: *hemoleucogramă*, *nefropatie diabetică*, *glicemie*. However, for the last two of them their English counterparts do not exist in PWN either: the term *diabetic nephropathy* is not recorded in PWN, while the term *blood glucose* is recorded, but with a different meaning from the one that is of interest for us, namely the medical procedure, not the chemical substance.

For the terms occurring in RoWN, we extracted their hyponymy and mero-part relations and compared them to the hierarchical structure of the corresponding terms in SNOMED CT. The reason for choosing two relations in the wordnet, but only one in the thesaurus will become clear from the trees analysis presented below.

Out of the five terms occurring in RoWN, two are leaves in the network: *colesterol* and *pancreas*. Whereas the former has two children in PWN (*HDL cholesterol* and *LDL cholesterol*⁵), which have not been implemented in RoWN yet, the latter is also a leaf in PWN.

As a consequence, there are only three terms that have subtrees in RoWN: an anatomical one (*glandă endocrină*), one designating a disorder (*boală metabolică*) and one from the category chemicals (*sânge*).

⁵Comparing *cholesterol*, on the one hand, and *HDL cholesterol* and *LDL cholesterol*, on the other hand, from the perspective of terminology formation, we notice that the hyponyms are formed, in this case, by modifying the hyperonym. As other terms in this paper show, this is not the only way of creating specialized terms.

The trees headed by the synsets they belong to were extracted from RoWN and compared to the trees headed by their equivalent synsets in PWN (for an online synchronous visualization of PWN and RoWN synsets visit <http://dcl.bas.bg/bulnet/>).

For all these terms we notice that the Romanian coverage of the English equivalents is quite high: *endocrine gland* has 16 hyponyms and 13 of them are implemented as hyponyms of *glandă endocrină*, *blood* has 10 hyponyms and 8 of them are implemented as hyponyms of *sânge*, and *metabolic disorder* has 6 hyponyms and 3 of them are implemented as hyponyms of *boală metabolică*.

These (sub)trees depths are quite low and identical for the two wordnets: *endocrine gland* and *glandă endocrină* have a depth of 3 levels, *blood* and *sânge* have 2 levels, and *metabolic disorder* and *boală metabolică* have a depth of 3, respectively 2, levels. The trees for the last two terms are rendered in Figure 1. The identically coloured edges are used for highlighting the hyperonymy relations of the same child. For example, the node *abetalipoproteinemia* has two hyperonyms: *hypobetalipoproteinemia* and *lipidosis*. The edges between the hyponym and each of its hyperonyms are the same colour.

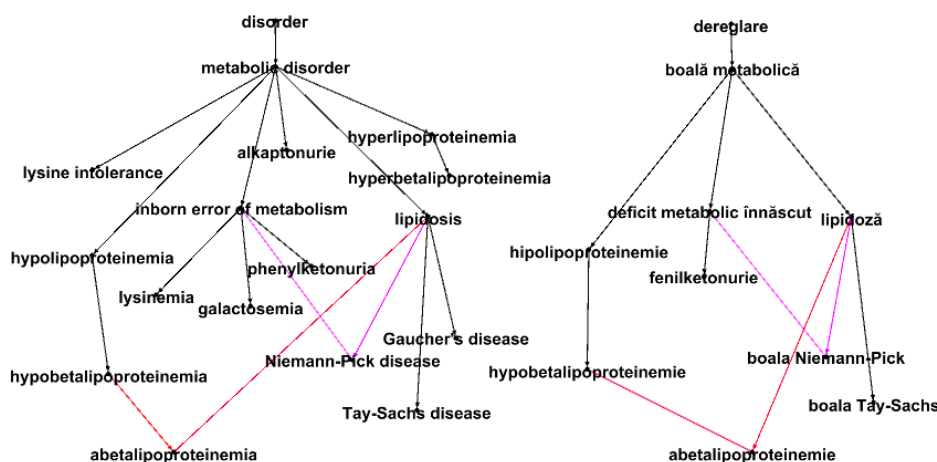


Figure 1: The subtrees of *metabolic disorder* in PWN (left) and of *boală metabolică* in RoWN (right)

Analyzing the direct and indirect hyponyms of these terms, as well as the relations in which these hyponyms are involved in the wordnet we can make several remarks:

- some hyperonyms share a part of their hyponyms: consider the term *boala Niemann-Pick* (En. *Niemann-Pick disease*) which is a hyponym of *deficit metabolic înnăscut* (En. *inborn error of metabolism*) and also of *lipidoză* (En. *lipidosis*). These two hyperonym terms are, in their turn, cohyponyms whose hyperonym is *boală metabolică* (En. *metabolic disorder*) - see Figure 1;
- some cohyponyms are also involved in the mero-part relation: this is the case with several terms in the (sub)tree of *glandă endocrină*: see Figure 2. For example, the term *cortex adrenal* (En. *adrenal cortex*) is both a cohyponym of *glandă suprarenală* (En. *adrenal gland*) and a meronym of it. There is one case when even a part and a subpart of this part are cohyponyms with the whole: *pars intermedia*, *posterior pituitary* and *pituitary gland* are cohyponyms of the hypernym *endocrine gland*. However, *pars intermedia* is a meronym of *posterior pituitary*, and *posterior pituitary* is a meronym of *pituitary gland* - see the same figure.

4.2. Comparison between Wordnet Hierarchies and SNOMED CT Hierarchy

A first remark on the corresponding (sub)trees in wordnets and in SNOMED CT concerns their depth: in the case of the terms with hyponyms in wordnet (*endocrine gland*, *blood* and *metabolic disorder*) the difference is rather small: with the exception of *endocrine gland*, which has 3 levels in wordnet and 4 in SNOMED CT, the other (sub)trees have the same depth in both resources: *blood* has 2 levels and *metabolic disorder* has 3.

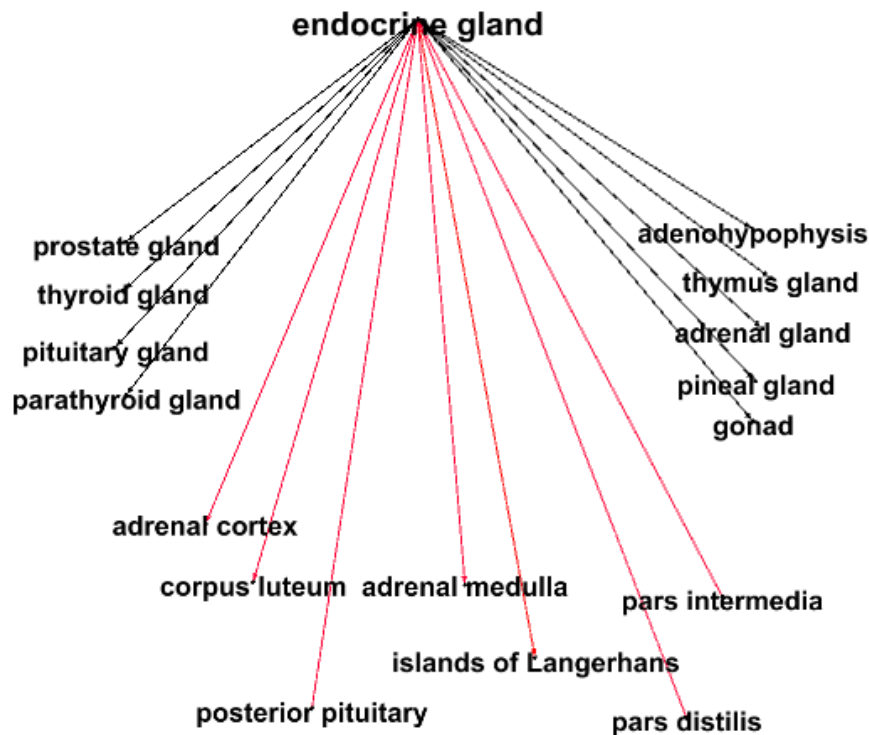


Figure 2: The subtree of *endocrine glands* in PWN

The terms that are not present in wordnets, namely *blood glucose* and *diabetic nephropathy*, have a 3-level depth in SNOMED CT. Moreover, these terms have a large number of children in the SNOMED CT hierarchy: each of them has 11 children. In Figure 3 we illustrate the hierarchy for *blood glucose* in SNOMED CT.

From a medical point of view the assessment of blood glucose is the most important paraclinical measurement, being used for screening, diagnosis and follow up of the patients with diabetic disease. Therefore if one needs to perform named entity recognition (NER) or relation extraction (RE) using PWN or RoWN the annotation of this term is essential. Consequently, because *blood glucose* is one of the most frequently used medical term in the diabetes domain, it should be introduced in both RoWN and PWN.

The terms that are present in PWN but lack any hyponyms are *complete blood count* and *pancreas*. In SNOMED CT, the same terms have different depths: the former has a 2-level depth, while the latter has a 7-level depth. They have three, respectively six children. We notice again the richness of the subtrees headed by these concepts, thus proving their importance in the medical domain.

The case of *cholesterol* is suggestive of the selectivity of wordnets, which, as resources aiming at reflecting the general language rather than the domains of activity, out of the big number of specialized terms encode only those that are accessible to all people: in PWN there are only two children for *cholesterol*: *LDL cholesterol* and *HDL cholesterol*. A further proof of the accessibility of these terms to non-specialists and of their use in the general language is the creation, in the common language, of two synonyms whose meaning is more transparent for patients given the use of very common adjectives in the structure of these terms: they are *colesterol rău* (En. *bad cholesterol*) and, respectively, *colesterol bun* (En. *good cholesterol*). In SNOMED CT, however, the concept *cholesterol* has 8 other children besides these two. Moreover, among all these ten children, the two mentioned above have the highest number of children, in their turn: *HDL cholesterol* has 7 children and *LDL cholesterol* has 5 children.

All these terms discussed so far in this section show that wordnet hierarchies tend to be simpler, thus more accessible to non-specialist people and useful in Natural Language Processing when dealing with

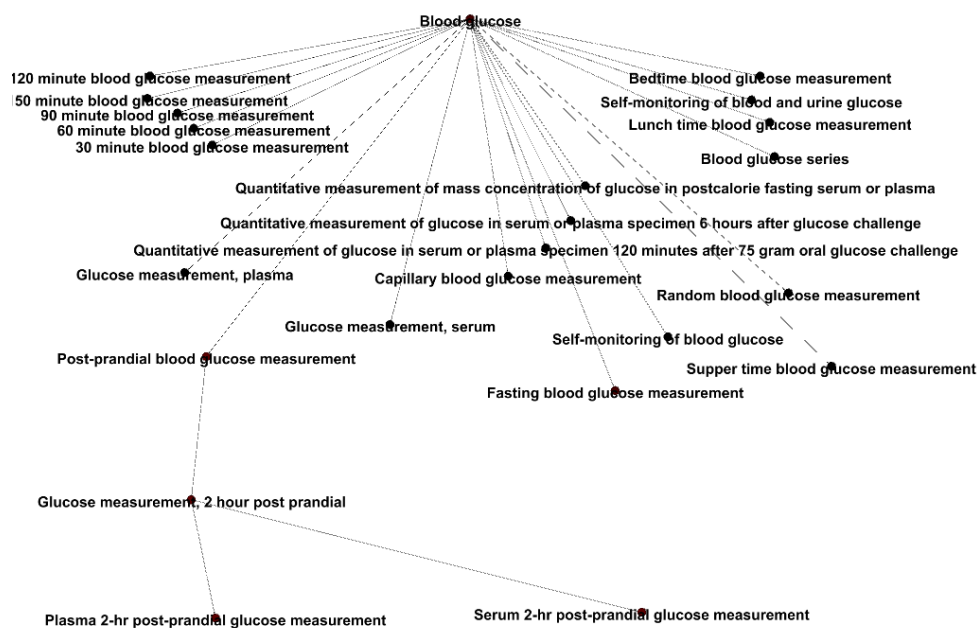


Figure 3: The subtree of *blood glucose* in SNOMED CT

texts from general language, rather than domain-specific ones (Bodenreider et al., 2003). The term that we discuss below is meant to highlight further shortcomings of wordnets when dealing with (at least) medical texts and terms, namely conceptual organization of terms in partial contradiction to medical knowledge.

When comparing the subtrees⁶ for *endocrine gland* in PWN (see Figure 2) and in SNOMED CT (see Figure 4), we notice several mismatches of the same kind: several parts of endocrine glands are encoded as such (with the help of the mero-part relation), but also as co-hyponyms of the nouns designating these glands. Consider the *adrenal cortex*. In PWN it is in mero-part relation with *adrenal gland*, but also its co-hyponym, both sharing the hyperonym *endocrine gland*. However, in SNOMED CT it is a hyponym of *layer of adrenal gland*, which has *endocrine gland* as hyperonym; at the same time, *adrenal cortex* is a part of the *adrenal*. A similar, although slightly more complicated case, is that of *pars intermedia*: in PWN it is a part of *posterior pituitary* which, in its turn, is a part of *pituitary gland*; both *posterior pituitary* and *pars intermedia* are co-hyponyms of *pituitary gland*, all having the hyperonym *endocrine gland*. In SNOMED CT *pars intermedia* is a hyponym of *adenohypophysis* which is a hyponym of *pituitary part*. So, we notice the wrong attachment of *pars intermedia* as a part of the *posterior pituitary* instead of the *adenohypophysis* in PWN, alongside its inappropriate attachment as a hyponym of *endocrine gland*. All the red edges in Figure 2 mark (PWN) relations that are non-conformant to the ones in green in Figure 4 (in SNOMED CT).

We only mention now, but leave it unexplained given space constraints, that the hierarchy of *metabolic disorder* in wordnets (as represented in Figure 1) does not conform with the SNOMED CT hierarchy.

5. Integrating Medical Terms in RoWN

Analyzing the eight different cases of diabetes related general terms we dealt with in this pilot study, we can identify several scenarios for their implementation in RoWN, together with all their hyponyms. These scenarios depend on the current status of the implementation of these terms in RoWN and PWN and the cases are: (i) the term is not implemented in RoWN and it is not in PWN either: this is the case

⁶These subtrees are quite dense in both these resources and, given space constraints, we needed to simplify them so as to serve the purpose of the discussion in this section.

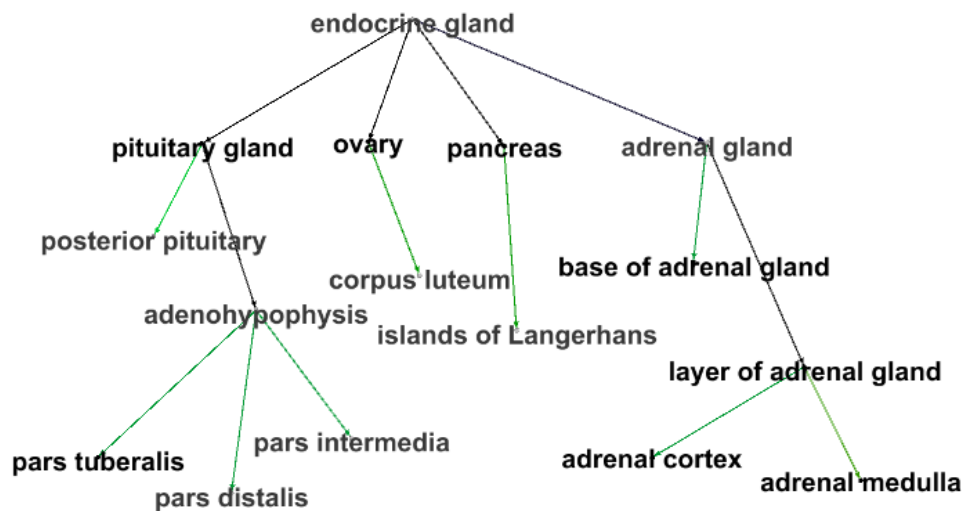


Figure 4: The subtree of *endocrine gland* in SNOMED CT

of *glicemie* (En. *blood glucose*) and of *nefropatie diabetică* (En. *diabetic nephropathy*); (ii) the term is not implemented in RoWN but it is in PWN, where it is a leaf: this is the case of *hemoleucogramă* (En. *complete blood count*); (iii) the term is implemented both in RoWN and in PWN as a leaf: see the case of *pancreas*; (iv) the term is implemented both in RoWN and in PWN, but it has children only in PWN, in RoWN being a leaf: this is the case of *cholesterol*; (v) the term is implemented in both RoWN and PWN and has children in both of them: see the cases *boală metabolică* (En. *metabolic disorder*), *sânge* (En. *blood*) and *glandă endocrină* (En. *endocrine gland*).

For these cases we propose the following respective scenarios: (i) go up, in parallel, in the PWN and SNOMED CT hierarchies until corresponding nodes are found; translate and implement in the RoWN the terms and their hyponymic organization in SNOMED CT starting from this common node and going down until the leaves level; keeping track of them is ensured by assigning them a unique distinctive ID⁷; (ii) translate the general term from PWN and implement it in RoWN with the same ID as in PWN; find its equivalent in SNOMED CT and translate and implement all children down to the leaves level, importing also their hyponymic structure; all children get a unique distinctive ID; (iii) the same as at (ii) above, but skipping the implementation of the general term; (iv) if the children that are already in PWN are also children of the same general concept in SNOMED CT, then implement them also in RoWN, with the same ID as in PWN; if their structure is different from the one in SNOMED CT, see the next scenario; translate and implement all the other children (if any) and all children's children down to the leaves level, importing also their hyponymic structure; all children that were not in PWN get a unique distinctive ID; (v) this is the most challenging situation, because it has to deal with the discrepancies between the two types of resources, some of which are presented above. Modifying the existent RoWN structure and completely replacing it with the one from SNOMED CT is excluded because the alignment of RoWN and PWN confers the former (but also the latter) a great value for bi-/multilingual applications involving natural language processing (Tufiş et al., 2004). As a consequence, we can create a parallel medical structure in the RoWN: the nodes that do not conform with the SNOMED CT organization will be doubled and marked distinctively. The result will be the existence of two synsets containing the same literals, but establishing different relations in the network. Applications requiring access to medical data will ignore the wordnet-specific relations and make use of the SNOMED CT-specific ones in such cases.

⁷The synsets in RoWN have an ID identical with the one in PWN, thus ensuring the alignment of the two resources.

6. Conclusions

Resources such as SNOMED CT are expensive both time-wise and money-wise. For languages such as Romanian, lacking such a thesaurus, an alternative solution can be the integration of necessary knowledge from SNOMED CT in the existing wordnet. However, as we showed here, the process involves making important decisions at two levels: finding the Romanian equivalents of the English terms in SNOMED CT (see section 3.2.) and establishing various scenarios for their integration in the wordnet (see section 5.). These are not trivial, given the importance of having aligned wordnets for different languages, thus the inefficiency of a solution that would modify the wordnet structure.

Even though the wordnet was not built for domain-specific applications it can be enriched with specialized terminologies (medical) extracted from already existing specialized ontologies (SNOMED CT) in order to perform terms identification or relations extraction.

References

- Barry, S. and Fellbaum, C. (2004). Medical wordnet: a new methodology for the construction and validation of information resources for consumer health. In *Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics*.
- Bodenreider, O., Burgun, A., and Joyce, A. M. (2003). *Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: a feasibility study*.
- Fellbaum, C., Ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Hicks, A., Seppälä, S., and Bond, F. (2018). Toward Constructing the National Cancer Institute Thesaurus Derived WordNet (ncitWN). In *Proceedings of the 9th Global WordNet Conference, Singapore*.
- Mitrofan, M. (2017). Bootstrapping a romanian corpus for medical named entity recognition. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 501–509.
- Reynoso, G., March, A., M Berra, C., P Strobietto, R., Barani, M., Iubatti, M., P Chiaradio, M., Serebrisky, D., Kahn, A., Vaccarezza, O., L Leguiza, J., Ceitlin, M., Luna, D., Quirós, F., I Otegui, M., Puga, M., and Vallejos, M. (2000). Development of the spanish version of the systematized nomenclature of medicine: methodology and main issues. In *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 694–8, 02.
- Toumouh, A., Lehireche, A., Widdows, A., and Malki, M. (2006). Adapting wordnet to the medical domain using lexicosyntactic patterns in the ohsumed corpus. In *Computer Systems and Applications, 2006.*, pages 1029–1036.
- Tufts, D. and Barbu Mititelu, V. (2014). *Language Production, Cognition, and the Lexicon*, chapter The Lexical Ontology for Romanian. Springer.
- Tufts, D., Ion, R., and Ide, N. (2004). Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING2004*, Geneva.
- W. Hersh, C. Buckley, T. J. L. and Hickam, D. (1994). Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *In Proceedings of the 17th annual conference on Research and Development in Information Retrieval (SIGIR-94)*, pages 192–201.

Factors and Features Determining the Inheritance of Semantic Primes between Verbs and Nouns within WordNet

Ivelina Stoyanova

Department of Computational Linguistics

Institute for Bulgarian Language

Bulgarian Academy of Sciences

iva@dcl.bas.bg

Abstract

The paper outlines the mechanisms of inheriting semantic content between verbs and nouns as a result of derivational relations. The main factors determining the inheritance are: (1) the semantic class of the verb as represented by the noun; (2) the subcategorisation frame and argument structure of the verb predicate; (3) the derivational relation between the verb and the noun, as well as the resulting semantic relation made explicit through the derivation; (4) hierarchical relations within WordNet.

The paper explores three types of verb-noun prime inheritance relations: (a) universal – not depending on the argument structure, which are eventive or circumstantial; (b) general – specific to classes of verbs, for example agentive or non-agentive; (c) verb-specific – depending on the specific subcategorisation frame of the verb as presented in VerbNet and/or FrameNet. The paper presents a possibility for extended coverage of semantic relations based on information about the argument structure of verbs.

Further, the work focuses on the regularities in the way in which derivationally related nouns inherit semantic characteristics of the predicate. These regularities can be applied for the purpose of predicting derivationally and semantically related synsets within WordNet, as well as for the creation of language specific synsets, for consistency checks and verification.

1. Introduction

The study explores the ways in which the derivationally based semantic relations between a verb and a set of nouns are predetermined by the features of the verb, its semantics, subcategorisation frame and set of arguments. Both nouns and verbs within WordNet are classified into semantic classes (defined by their corresponding semantic primes). The nouns inherit certain semantic characteristics from the verb with which they have a derivational relation and we call this process inheritance of semantic primes between verbs and nouns within WordNet. The study does not take into account the direction of derivation (verb derived from a noun or vice versa) but considers the process of inheritance of semantic primes as directed from the predicate to the corresponding nouns.

The observations presented here include examples from Princeton WordNet¹ and the Bulgarian WordNet (BulNet)². The methodology and conclusions are based on generalised semantic features and are thus largely language independent. In addition to the semantic description of verbs and nouns in WordNet based on the semantic primes and the hierarchical organisation of the lexical-semantic network, we also use information about the FrameNet frames and verb arguments lists from VerbNet to the end of studying the inheritance of semantic features and lexical conceptual structures.

¹<https://wordnet.princeton.edu/>

²<http://dcl.bas.bg/bulnet/>

Section 2 outlines some previous work in the field. Section 3 describes resources and results that have been applied in the study. Factors that determine the inheritance of semantic primes are discussed in section 4, followed by the main results in terms of deduced regularities in the process of inheritance in section 5. The paper concludes with some notes on the application of the research and possible future developments in section 6.

2. Related work

Most attention semantic primes attract with respect to semantic decomposition and interpretation (Goddard and Wierzbicka, 1994; Gomez, 1998; Fähndrich et al., 2014). Extensive classifications of verbs have also been proposed (Wilks, 1987; Levin, 1993; Korhonen, 2002; Van Valin, 2006), which are based on the notion of semantic primes (also called primitives or universals) as generalised concepts.

The purpose of the classification of verbs introduced in WordNet (Fellbaum, 1999a) is to reflect their regular syntactic behavior, including in terms of possible alternations. Additionally, semantic primes contribute to the network of semantic relations by assigning the verb to a semantic class with stable features which are inherited by the hyponym from its hypernym. Some semantic primitives in WordNet are also introduced as relations, e.g. CAUSE.

Recent studies focus on shared semantic features between derivationally related literals within synsets in the context of morphosemantic relations (Fellbaum et al., 2009; Koeva, 2008; Mititelu, 2012). Pala and Hlaváčková (2007) use a limited number of derivational relations between verbs and nouns (agent, patient, instrument, action, property) and consider derivational nests of words around a certain root. Dziob et al. (2017) describe two main types of derivational relations between nouns and verbs – role inclusion (fine-grained into subject, instrument, result, location, object, time) and circumstance.

Research on derivational and morphosemantic relations for Bulgarian has recently been presented by Stoyanova et al. (2013), Leseva et al. (2014), Tarpomanova et al. (2014), Dimitrova et al. (2014), and Leseva et al. (2015). They analyse the semantic information carried by various derivational models. The set of morphosemantic relations used in Princeton WordNet is used and semantic primes of nouns are considered in order to identify inconsistencies with the semantic roles (Leseva et al., 2015).

The topic of cross-POS inheritance of semantic primes and semantic features, in particular between verbs and nouns connected by a derivational relation within WordNet, has rarely been discussed in the literature. Pustejovsky (1991), Pustejovsky (1995), Copestake (1992), among others, discuss lexical inheritance structure which defines how one lexical structure is related to other lexical structures in the lexicon. Inheritance of semantic features from verbs to nouns as presented in this paper is similar to the projective inheritance (Pustejovsky, 1995) since it also relies on linking the conceptual information from syntactic-based realisation of lexical items.

3. Prerequisites

WordNet (Miller, 1995; Fellbaum, 1999b) is a large lexical-semantic resource that groups word senses rather than lexical units into a large network. The individual senses correspond to synonym sets (synsets). WordNet provides multilingual support through a unique identification index. The relational structure of WordNet relies on: (a) conceptual relations such as hypernymy/hyponymy, holonymy/ meronymy, etc.; (b) lexical relations between members of synsets (literals) such as antonymy; (c) derivational relations between literals; (d) morphosemantic relations between verbs and nouns where the derivational relation (at the literal level) reflects a semantic relation (at the synset level). The hierarchical structure of WordNet is based on the relations of hypernymy/hyponymy. The root (top node in the tree) represents the most generalised meaning and its hyponyms down the tree inherit this meaning and make it more specific.

All the verb and noun synsets in Princeton WordNet are classified into a number of language-independent semantic primes. The nouns are categorised into 25 groups, such as noun.act (acts or actions), noun.artifact (man-made objects), noun.person, etc. The verbs fall into 15 groups, such as verb.body (verbs of grooming, dressing and bodily care), verb.change (verbs of size, temperature change, intensifying, etc.), verb.cognition (verbs of mental operations), as defined in the PWN lexicographer

files³. Semantic primes within WordNet are aimed to represent universal semantic-conceptual categories of verbs and nouns which provide a generalised description of their semantic features and syntactic behavior.

Morphosemantic relations link verb–noun pairs of synsets that contain derivationally related literals. For the purposes of this study, we do not consider the direction of the derivation (source and derivative) and assume that derivational relations are symmetric. As semantic and morphosemantic relations refer to concepts, they are universal, and such a relation must hold between the relevant concepts in any language, regardless of whether it is morphologically expressed or not. This has enabled the automatic transfer of the relations to other languages, e.g. from Princeton WordNet to Bulnet (Koeva, 2008; Stoyanova et al., 2013; Leseva et al., 2014; Leseva et al., 2015).

We use the inventory of morphosemantic relations from the Princeton WordNet 3.0. morphosemantic database⁴: Agent, By-means-of (inanimate Agents or Causes but also Means and possibly other relations), Instrument, Material, Body-part, Uses (intended purpose or function), Vehicle (means of transportation), Location, Result, State, Undergoer, Destination, Property, and Event (linking a verb to its eventive nominalisation).

On the other hand, VerbNet verb classes and FrameNet frames provide more detailed features for the classification of verbs with respect to their semantic and syntactic properties and function. Both resources can be used to group verbs into semantic classes of different granularity and different level of generalisation as part of a hierarchical organisation.

FrameNet (Baker et al., 1998) represents conceptual structures called frames which describe particular types of objects, situations, etc. along with their participants, or frame elements (Ruppenhofer et al., 2016). FrameNet is internally hierarchically structured using a set of frame-to-frame relations, in particular *Inheritance* – the child frame is a subtype of the parent frame, e.g. *Change_position_on_a_scale* inherits from *Event* and is inherited by *Change_of_temperature*.

The VerbNet (Kipper-Schuler, 2005; Kipper et al., 2008) classes represent formations of verbs with shared semantic and syntactic properties and behaviour. They are organised in a shallow hierarchy grouping classes into generalised types such as *Verbs of Creation and Transformation*, *Verbs of Communication*, *Verbs of Social Interactions*, etc.

However, linking VerbNet and FrameNet to WordNet is not straightforward. There are two main types of mappings that have already been applied on the lexical resources discussed herein: (a) lexical mapping – lexical units (from one resource) have been assigned categories from another resource; and (b) structural mapping – classification categories from one resources have been aligned to categories from another. Previous efforts at linking these resources include Shi and Mihalcea (2005), Baker and Fellbaum (2009), Laparra and Rigau (2010), as well as the SemLink project⁵ and WordFrameNet⁶. These result in limited coverage of WordNet synsets and further efforts are required in order to improve the mappings. More details on linking WordNet, VerbNet and FrameNet are presented by Leseva et al. (2018).

4. Factors determining the inheritance of semantic primes between verbs and nouns

This section outlines the process of analysis of derivation and the resulting semantic relations between the verb and a set of derivationally related nouns. We briefly discuss the types of inheritance (as reflected by the morphosemantic relations adopted in Princeton WordNet). These, coupled with the semantic and syntactic features of a given verb, play a significant role in determining the semantic primes of the associated nouns.

4.1. Features of the verbs

The following features of the description of the verbs are essential for the analysis of inheritance:

³<https://wordnet.princeton.edu/man/lexnames.5WN.html>

⁴<http://wordnetcode.princeton.edu/standoff-files/morphosemantic-links.xls>

⁵<https://verbs.colorado.edu/semlink/>

⁶<http://adimen.si.ehu.es/web/WordFrameNet>

- (a) Semantic prime of the verb in WordNet. The semantic prime describes the abstract semantic category of the verb. We use the following list of verb semantic primes: verb.body, verb.change, verb.cognition, verb.communication, verb.competition, verb.consumption, verb.contact, verb.creation, verb.emotion, verb.motion, verb.perception, verb.possession, verb.social, verb.stative, verb.weather.
- (b) Semantic frame from FrameNet. Each frame is described by the following components: a general definition of the frame; a list of frame elements (core, peripheral and extra-thematic elements of the frame) corresponding to arguments or adjuncts, which may also contain relevant semantic restrictions; a set of relations to other frames in FrameNet (in their entirety these frame-to-frame relations establish the internal structure of FrameNet).
- (c) The VerbNet class and the set of arguments associated with it. The description provided by the class and the superclass in VerbNet partially overlaps with the information encoded in the semantic frame from FrameNet. We enrich the descriptions from the two sources by combining them. A VerbNet class is typically associated with more than one syntactic frame which shows possible syntactic variations of the usage of the verb.
- (d) Known morphosemantic relations in WordNet. When we analyse the set of possible derivational and semantic relations, we consider the ones that have already been encoded. These are also useful in making observations about the frequency of pairs of semantic primes and corresponding morphosemantic relations.
- (e) Hierarchical relations within WordNet. The structure of WordNet can be helpful since the hyponyms inherit the semantic properties of the hypernym (a more concrete concept inherits the properties of a more general one and adds on a more specific meaning).

The semantic primes of verbs impose restrictions on the possible semantic roles and the semantic frame of the verb. For example, verb.weather or verb.phenomenon are not compatible with the frame element or the semantic role Agent since the meaning of the verb implies the action of natural forces.

4.2. Semantic primes of the nouns

We work with the following set of noun semantic primes: noun.act, noun.animal, noun.artifact, noun.attribute, noun.body, noun.cognition, noun.communication, noun.event, noun.feeling, noun.food, noun.group, noun.location, noun.person, noun.phenomenon, noun.possession, noun.process, noun.relation, noun.shape, noun.state, noun.time, noun.vehicle. Also, the generalised label noun.Tops is used to signify top-level (root) abstract nouns (e.g. *entity*) which are not relevant for the present analysis.

There are various restrictions imposed on the possible relations and verb arguments which stem from the noun semantic class as expressed by the semantic prime. For example, Agent is associated with persons (noun.person), social entities, e.g., organisations (noun.group) and animals (noun.animal) that are capable of acting. Instruments are concrete man-made objects (noun.artifact), but nouns with the prime noun.communication may also function as instruments, e.g. *дебъгър:1* – *debugger:1* (*n*).

Inanimate causes (Fellbaum et al., 2009) – non-living (and non-volitional) entities that bring about a certain effect or result – are expressed by the morphosemantic relations Body-part, Material, Vehicle, and By-means-of. The relation Body-part may be an inanimate cause that is an inalienable part of an Actor and is expressed by nouns with the prime noun.body (rarely noun.animal or noun.plant). The relation Material denotes a subclass of inanimate causes – substances that may bring about a certain effect, e.g. *inhibitor:1* (a substance that retards or stops an activity). Beside noun.substance, noun.artifacts (synthetic substances or products) also qualify for the relation, e.g. *depilatory:2* (hair removal cosmetics). The relation Vehicle represents a subclass of artifacts (means of transportation); consequently the respective synsets have the prime noun.artifact and are generally hyponyms of the synset *conveyance:3*; *transport:8*. Inanimate causes whose semantics differ from that of the other three relations, are assigned the generic relation By-means-of, e.g. *geyser:2* ('a spring that discharges hot water and steam') (noun.object), etc.

The relation Event denotes processual nominalisation and involves nouns such as noun.act, noun.event, noun.phenomenon, and rules out concrete entities such as animate beings, natural (noun.object) and man-made (noun.artifact) objects, etc. The relation State denotes abstract entities such as feelings, cognitive states, etc. Undergoer is assigned to entities which are affected by the event or state. The relation Result involves entities that are produced or have come to existence as a result of the event or state. Property signifies various attributes and qualities. These relations involve nouns with various primes.

The relation Location denotes a concrete (natural or man-made) or an abstract location where an event takes place and therefore relates verbs with nouns with various primes – noun.location, but also noun.object, noun.plant, noun.artifact, noun.cognition, etc. The relation Destination is associated with the primes noun.person, noun.location and noun.artifact, which corresponds to two distinct interpretations of the relation – Recipient (noun.person) and Goal (noun.artifact, noun.location). The relation Uses denotes a function or purpose, e.g. *lipstick:1 – lipstick:3*. The relation allows nouns with various primes, both concrete and abstract.

4.3. Types of inheritance

Derivational relations between verbs and nouns (regardless of the direction of derivation) result in semantic relations which depend on the semantic characteristics of the verb. We analyse the typology of derivationally-based inheritance of semantic properties as a factor for the realisation of the semantic relations.

We recognise three types of verb-to-noun inheritance of semantic characteristics:

- (1) **Universal inheritance** potentially can apply to all verbs regardless of their semantic prime and argument structure. However, not all verbs exhibit these relations: firstly, it is a matter of linguistic choice whether to lexicalise certain concepts, and secondly, there may be no derivational relation even if a semantic relation is present.

Universal inheritance is carried out by two types of relations: (1) Eventive relations such as EVENT, STATE or PROCESS – nominalisations of the action, state or process signified by the verb, e.g. EVENT *готвя:2 / cook:1 (v) – готвене:1 / cooking:1 (n)*, STATE *завиждам:1 / envy:2 (v) – завист:1 / envy:1 (n)*, etc.; and (2) circumstantial relations such as LOCATION, e.g. *печатам:1 / print:1 (v) – печатница:1 / printing press:1 (n)*; TIME, e.g. *вечерям:1 / dine:2 (v) – вечер:3 / evening:1 (n)*; ATTRIBUTE/ABSTRACT, e.g. *издържам:1 / endure:1 (v) – издръжливост:4 / endurance:1 (n)*; etc.

Cases where LOCATION or TIME are part of the subcategorisation frame of the verb, e.g. LOCATION *лагерувам:1 / camp:4 (v) – лагер:6 / camp:6 (n)* (noun.location), are not regarded as universal but as verb-specific (see below).

- (2) **General inheritance** is determined from the verb's membership to the general semantic class defined by its semantic prime. The properties of the semantic class often influence the set of possible arguments of the verb and thus, the set of semantic relations that can be manifested through a derivational relation.

General inheritance mostly refers to the division between agentive and non-agentive verbs. Some classes, such as verb.cognition, verb.possession, verb.consumption, associate with an AGENT, e.g. *продавам:5 / sell:4 (v) – продавач:1 / seller:1 (n)*, while other classes, such as verb.weather, associate with an inanimate ACTOR, e.g. *гърмя:6 / thunder:4 (v) – гръмотевица:1 / thunderbolt:2 (n)*. There are also verb classes whose members can take either an animate or an inanimate subject. However, for a better classification of verbs with a view to their syntactic behavior, these classes need to be subdivided into relevant subcategories that reflect these differences.

General inheritance also has to do with the division between causative and inchoative verbs. Analysis of material from WordNet shows that large verb groupings determined by a common semantic prime contain both causative and inchoative members, e.g. the prime verb.change is assigned to both *превърщам:3 / convert:5 (v)* and *превърщам се:2 / convert:1 (v)*, and even that causative

and inchoative verbs may be found in a single synset, e.g. *blacken:1 (v) 'make or become black'*. Clearly, a more finely grained classification of verbs with respect to their syntactic behavior will enforce a clear-cut distinction between these two types of verbs since they exhibit diametrically different semantic relations and inheritance capabilities.

It can also be possible to introduce further granularity of verb classes with respect to the semantic relations of RESULT (resultative verbs as part of semantic classes such as verb.change or verb.perception). Another relation is INSTRUMENT / BY_MEANS_OF / USES where a distinction can be made between concrete verbs of actions such as verb.body or verb.contact, which can involve instruments, unlike abstract verbs such as verb.cognition or verb.communication, which will more likely be associated with BY_MEANS_OF. However, these categories need further analysis and consideration in order to provide a clear-cut classification.

- (3) **Verb-specific inheritance** depends on the semantic frame and the set of arguments of the particular verb. This type of inheritance can be influenced by the hierarchical relations within WordNet, e.g. the frame of the direct hypernym of the verb synset.

For example, the verb *кърмя:1 / breastfeed:1 (v)* realises universal inheritance through the semantic relation of EVENT *кърмене:2 / breast feeding:1 (n)* and general inheritance by being an agentive verb of the class verb.consumption through the semantic relation of AGENT *кърмачка (not in WordNet) 'breastfeeding mum' (n)*. In addition, the verb has arguments that are inherited through the specific meaning which determines its membership to the VerbNet class feeding-39.7 (Verbs of Ingesting): Recipient (+animate), which is realised through the semantic relation of BENEFICIENT *кърмаче:1 / nursling:1 (n)*, and Theme (+comestible) which is realised through BY_MEANS_OF *кърма:2 / milk:4 (n)*.

4.4. Potential extended coverage of semantic relations of verb synsets

The process of gradually extending the number of possible semantic relations of verbs is also illustrated in Figures 1–3. Figure 1 shows the verb *завиждам:1 / envy:2 (v)* with its single derivationally related counterpart in WordNet, the noun *завист:1 / envy:1 (n)*, which exhibits the morphosemantic relation of STATE. Further, we extend the number of potential relations by considering the different sources of relations (Figure 2) – the universal eventive relation has been saturated, but it is potentially possible to have some circumstantial relations such as ATTRIBUTE. The verb belongs to the class verb.feeling which entails an animate Experiencer. The verb belongs to the VerbNet class admire-31.2 (Psych Verbs) which can have arguments such as Stimulus, Experiencer (+animate) and Attribute. Figure 3 shows that some of these potential relations have been realised.

Relations can either be: (a) lexicalised, presented by a synset in WordNet and encoded with explicit morphosemantic relation to the verb (e.g. STATE *завист:1 / envy:1 (n)*); (b) lexicalised, presented by a synset in WordNet, but not encoded with a morphosemantic relation to the verb (e.g.); (c) lexicalised but not present in WordNet (e.g. EXPERIENCER *завистник, завистливец 'envious person' (n)*); or (d) not be lexicalised in a given language (e.g. in Bulgarian, STIMULUS *обект на завист 'object of envy' (n)*).

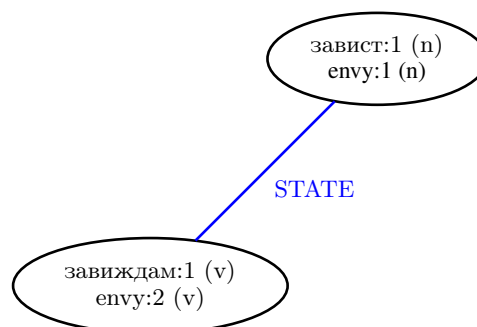


Figure 1: Verb synset connected to a set of noun synsets via morphosemantic relations in WordNet.

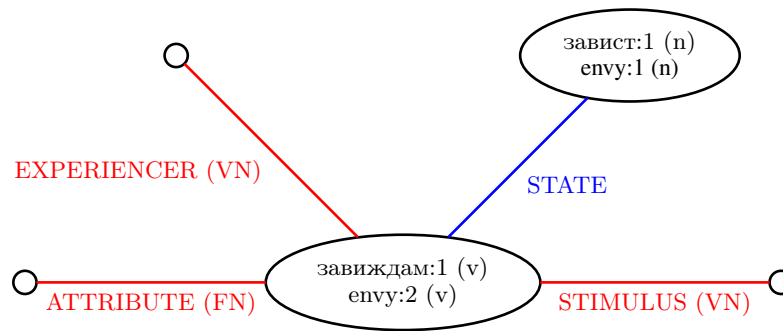


Figure 2: Extending the set of potential semantic relations using universal, general and verb-specific inheritance (non-exhaustive). Source: VN – VerbNet; FN – FrameNet.

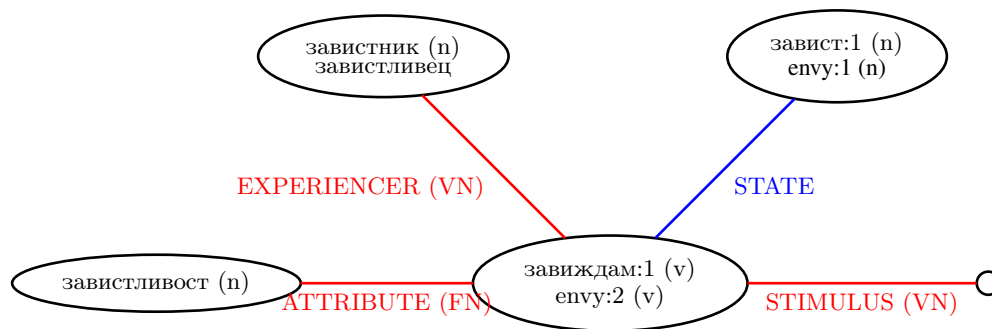


Figure 3: Filling (some of) the potential semantic relations slots (non-exhaustive).

5. Regularities in the inheritance of semantic primes

The main result of the study presented in the paper consists of discovering regularities in the process of inheriting semantic features from verbs to nouns as exhibited through derivational (and consequently) semantic relations.

Initial observations are performed on the morphosemantic relations already present in BulNet. The data are presented in the following format: for each pair <Verb_sem_prime, Morphosem_relation> we find all corresponding Noun_sem_prime (NSP) values with their respective frequencies within WordNet. The data show that within each verb semantic class, depending on the relation, there are a limited number of dominating noun semantic primes, while others are underrepresented either because they show rare cases or because they are due to errors or inconsistencies in the semantic prime assignments.

Table 1 shows the distribution of relations across verbs in BulNet, and the verb primes the highest percentage of which enter this relation. For certain relations reliable conclusions cannot be drawn due to the limited amount of data and these are not presented in the table. Moreover, the last column of Table 1 shows the potential extended coverage obtained by the method in section 4.4. Further, the coverage can be extended by introducing a more fine-grained classification of semantic relations aligning them to the thematic roles as presented in VerbNet and FrameNet, e.g. introducing the subrelation *Product* within *Result*, or *Material* within *By_Means_of*, etc. This will complement the semantic description of verbs within WordNet and will facilitate further investigation of semantic inheritance.

Table 2 shows the results from the observations on the most common (universal) relation, *EVENT*, which is realised for 42% of the verbs in WordNet. The distribution of some semantic primes follow a similar pattern being dominated by a small number of primes: *AGENT* (where all verb categories are dominated by noun.person), *INSTRUMENT* (dominated largely by noun.artifact), *LOCATION* (dominated by noun.location and noun.artifact), *STATE* (dominated largely by noun.state with the exception of verb.emotion dominated by noun.feeling and verb.cognition dominated by noun.cognition). More varied is the distribution of relations such as *RESULT* (Table 3), *BY_MEANS_OF*, *USES*, *UNDERGOER*, which result in more diverse set of noun semantic primes.

Semantic relation	Coverage across all verb synsets	Most coverage among the following verb primes	Potential extended coverage
Event	42%	verb.communication (67%) verb.perception (65%) verb.social (63%)	65%
Agent	16%	verb.social (33%) verb.competition (28%) verb.communication (28%)	35%
Result	8%	verb.creation (25%) verb.change (14%) verb.contact (12%)	14%
By_Means_of	6.5%	verb.communication (12%) verb.emotion (11%) verb.cognition (10%)	6.5%
Instrument	4.5%	verb.contact (16%) verb.creation (6%) verb.change (5%)	9%
State	3%	verb.emotion (30%) verb.cognition (6%) verb.stative (5%)	3%
Location	2%	verb.motion (4%) verb.stative (4%) verb.contact (3%)	6%

Table 1: Distribution of semantic relations across WordNet (as percentage of verb synsets) and the verb semantic primes for which the highest coverage of the relation occurs. Last column shows potential extended coverage obtained by the method.

Verb_semantic_prime	Number of different NSPs	Predominant primes	Coverage
verb.change	13	noun.act noun.process noun.event	83.8%
verb.cognition	12	noun.act noun.cognition	77.3%
verb.communication	13	noun.communication noun.act	82.8%
verb.consumption	11	noun.act	70.7%
verb.contact	11	noun.act noun.event	86.1%
verb.emotion	8	noun.feeling noun.act	70.0%
verb.motion	12	noun.motion noun.act noun.event	85.1%
verb.weather	7	noun.phenomenon noun.event noun.process	66.0%

Table 2: Distribution of resulting noun semantic primes from the EVENT relation across verb semantic primes (non-exhaustive). In the 3rd column the most frequent noun primes are listed corresponding to each verb prime which accumulatively account for over 2/3 of the cases.

Verb_semantic_prime	Number of different NSPs	Predominant primes	Coverage
verb.creation	15	noun.artifact, noun.communication, noun.cognition, noun.attribute, noun.food	77.2%
verb.change	20	noun.attribute, noun.substance, noun.object, noun.state, noun.food, noun.shape, noun.communication	71.7%
verb.contact	16	noun.artifact, noun.shape noun.object, noun.attribute noun.group	70.3%

Table 3: Distribution of noun semantic primes from the RESULT relation across verb semantic primes (non-exhaustive, for demonstration purposes only) demonstrating a variety of noun primes with no clear dominance.

These observations confirm the need for further refining of the semantic relations in order to capture better the variety of arising inheritance between a verb and the set of derivationally related nouns. By introducing subrelations such as BY_MEANS_OF_ACTOR, e.g. облекчавам:3 / palliate:1 (v) – успокоително:2 / palliative:2 (n), BY_MEANS_OF_INSTRUMENT, e.g. пека на скара:1 / grill:1 (v) – скара / grill:3 (n), this will lead to more consistent results. Moreover, it will allow further refining of inheritance within complex semantic primes such as noun.artifact or noun.communication.

6. Applications and further development

The research presented in this paper can be applied for the description of language-specific synsets in WordNet which at present are not part of the semantic classes. Moreover, as mentioned above, a more detailed classification of verbs and verb primes within WordNet will be beneficial for distinguishing groups of verbs with distinct syntactic features, such as causative and inchoative verbs, personal and impersonal verbs, etc. An improved mapping of VerbNet classes and FrameNet frames to WordNet synsets will be essential in obtaining more data and performing further analyses.

The results of the study can be used for consistency checks and verification of existing semantic relations. Further analysis is needed in order to distinguish rare but regular cases of inheritance from inconsistencies and mistakes, based on the semantic frames and the semantic and syntactic properties of both the verb and the noun, as well their respective place in the WordNet hierarchy and the relations with other synsets.

One of the most significant applications of the results is in extending WordNet with new semantic relations stemming from the argument structure of verb predicates. Correspondence between verb semantic primes and noun semantic primes in a derivational relation can help limit the scope of the search for possible new relations which will significantly improve the quality of automatic identification of relations. Further, the detailed classifications will be beneficial in identifying and defining new relations that have not been considered before, and may be used to further fine-grain the scope of relations and to enhance WordNet with richer semantic description.

Acknowledgement

This study has been undertaken within the project *Towards a Semantic Network Enriched with a Variety of Semantic Relations* funded by the National Scientific Fund of the Republic of Bulgaria under the Fundamental Scientific Research Programme (Grant Agreement No. 10/3/2016).

References

- Baker, C. F. and Fellbaum, C. (2009). WordNet and FrameNet as Complementary Resources for Annotation. In *Proceedings of the Third Linguistic Annotation Workshop (ACL-IJCNLP '09)*, Association for Computational Linguistics, Stroudsburg, PA, USA, pages 125–129.
- Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference. Montreal, Canada*, pages 86–90.
- Copstake, A. (1992). *The Representation of Lexical Semantic Information. Doctoral dissertation*. University of Sussex. Cognitive Science Research Paper CSRP.
- 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.
- Dziob, A., Piasecki, M., Maziarz, M., Wieczorek, J., and Dobrowolska-Pigoń, M. (2017). Towards Revised System of Verb Wordnet Relations for Polish. In *Proceedings of the LDK Workshop 2017*, pages 174–187.
- Fähndrich, J., Ahrndt, S., and Albayrak, S. (2014). Formal language decomposition into semantic primes. *AD-CAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, 3:56–73, 10.
- Fellbaum, C., Osherson, A., and Clark, P. E. (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.
- Fellbaum, C. (1999a). The Organization of Verbs and Verb Concepts in a Semantic Net. In Saint-Dizier, P., Ed., *Predicative Forms in Natural Language and in Lexical Knowledge Bases*, volume 6 of *Text, Speech and Language Technology*. Springer, Dordrecht.
- Fellbaum, C., Ed. (1999b). *WordNet: an Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Goddard, C. and Wierzbicka, A. (1994). Introducing Lexical Primitives. In Goddard, C. and Wierzbicka, A., Eds., *Semantic and Lexical Universals: Theory and empirical findings*, Studies in Language Companion Series, pages 31–56. John Benjamins Publishing Company.
- Gomez, F. (1998). Linking WordNet Verb Classes to Semantic Interpretation. In *Proceedings of the COLING-ACL Workshop on the Usage of WordNet in NLP Systems*, pages 58–64, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2008). Language Resources and Evaluation. *Commun. ACM*, 42(1).
- Kipper-Schuler, K. (2005). *VerbNet: A broad-coverage, comprehensive verb lexicon. PhD Thesis*. Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA.
- Koeva, S. (2008). Derivational and morphosemantic relations in Bulgarian Wordnet. *Intelligent Information Systems*, pages 359–368.
- Korhonen, A. (2002). Assigning Verbs to Semantic Classes via WordNet. In *Proceedings of the 2002 Workshop on Building and Using Semantic Networks*, volume 11 of *SEMANET '02*, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Laparra, E. and Rigau, G. (2010). eXtended WordFrameNet. In *Proceedings of LREC 2010*, pages 1214–1219.
- Leseva, S., Stoyanova, I., Rizov, B., Todorova, M., and Tarpomanova, E. (2014). Automatic Semantic Filtering of Morphosemantic Relations in WordNet. In *Proceedings of CLIB 2014, Sofia, Bulgaria*, pages 14–22.
- Leseva, S., Todorova, M., Dimitrova, T., Rizov, B., Stoyanova, I., and Koeva, S. (2015). Automatic classification of wordnet morphosemantic relations. In *Proceedings of BSNLP 2015, Hissar, Bulgaria*, pages 59–64.
- Leseva, S., Stoyanova, I., and Todorova, M. (2018). Classifying Verbs in WordNet by Harnessing Semantic Resources. In *Proceedings of CLIB 2018, Sofia, Bulgaria*.
- Levin, B. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago and London: The University of Chicago Press.
- Miller, G. A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41.

- Mititelu, V. B. (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.
- Pala, K. and Hlaváčková, D. (2007). Derivational Relations in Czech WordNet. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies, ACL '07*, pages 75–81, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Pustejovsky, J. (1991). The Syntax of Event Structure. *Cognition*, 44:47–81.
- Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge MA: MIT Press.
- Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Baker, C. F., and Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California.
- Shi, L. and Mihalcea, R. (2005). Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In Gelbukh, A., Ed., *Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science*, volume 3406. Springer, Berlin, Heidelberg.
- Stoyanova, I., Koeva, S., and Leseva, S. (2013). WordNet-based Cross-language Identification of Semantic Relations. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavonic Natural Language Processing*, pages 119–128.
- Tarpomanova, E., Leseva, S., Todorova, M., Dimitrova, T., Rizov, B., Barbu-Mitelu, V., and Irimia, E. (2014). Noun-verb derivation in the Bulgarian and the Romanian WordNet – a comparative approach. In *Proceedings of CLIB 2014, Sofia, Bulgaria*, pages 23–31.
- Van Valin, R. (2006). Some Universals of Verb Semantics. In Mairal, R. and Gil, J., Eds., *Linguistic Universals*, pages 155–178. Cambridge University Press, Cambridge, UK.
- Wilks, Y. (1987). Primitives. In Shapiro, S., Ed., *Encyclopedia of AI*, volume 2, pages 759–761. Wiley.

Online Editor for WordNets

Borislav Rizov

Institute for Bulgarian Language
Bulgarian Academy of Sciences
boby@dcl.bas.bg

Tsvetana Dimitrova

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

Abstract

The paper presents an online editor for lexical-semantic databases with relational structure similar to the structure of WordNet – Hydra for Web. It supports functionalities for editing of relational data (including query, creation, change, and linking of relational objects), simultaneous access of multiple user profiles, parallel data visualization and editing of the data on top of single- and parallel mode visualization of the language data.

1. Introduction

Hydra for Web is a complex system application for WordNet editing and visualization of complex relational data including parallel data (of two or more wordnets). Due to its complexity, the data encoded into the relational wordnet format need tools that are flexible and easy to use in order to give the users transparent access to data editing.

There are popular user interfaces that allow for the development and visualization of the wordnet relationships as graphs. For example, WordNet Editor (Dusza et al., 2013) is a tool for cooperative development of wordnet database that uses graphical component for graph visualizations with interactive navigation. Unlike the WordNet Editor which follows the idea for a strategy for wikipedia-like editing distributed platform, Hydra for Web integrates an editing functionality and a more simple interface that keeps the structure of a synset with all the relations integrated into one hierarchical structure, and strives for clear visualization of the parallel data.

In our approach to wordnet data, we treat wordnet as a relational structure – consisting of a set of objects and a set of binary relations between them. The objects are of three types – Synset, Literal and Note. The Literals (i.e., the words) in one synset are connected with it via a relation called *literal*. The Notes objects represent the textual data in wordnet – usage examples and notes. Every usage example is connected to its synset by the relation *usage* in a way similar to literals.

2. Viewer

Hydra for Web ¹, is a single page web application that uses as backend the API of the open source modal logic tool for wordnet development Hydra (it is freely distributed; (Rizov, 2008), for the browser functionalities see also (Rizov and Dimitrova, 2008). ².

The search system provides results in all of the available languages (selected by the user) – the database currently contains (open source) wordnets for 23 languages. The navigation bar has a drop-down menu for switching between the wordnets to be worked with. Except for the default Princeton WordNet 3.0, Bulgarian wordnet (BulNet) and Romanian wordnet (RoWN), the selectional options of which are visualized in corresponding pairs, the user can enable additional wordnets or disable others by means of a modal dialog.

The interface is currently available in English, Bulgarian, and Romanian.

¹Available at <http://dcl.bas.bg/bulnet/>

²Hydra for Web is freely downloadable at: <http://dcl.bas.bg/hydra/>

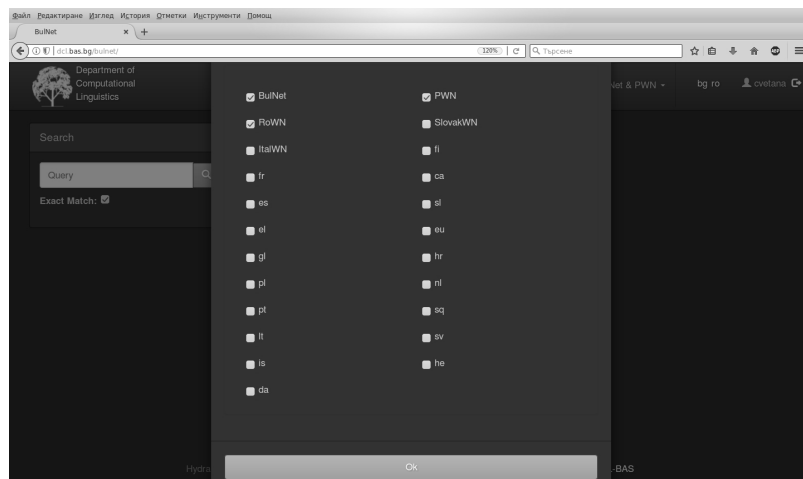


Figure 1: Hydra for Web – WordNet as a Multilingual Dictionary

The tool allows for searching into databases of different language wordnets with a single query as shown on Fig. 2 which illustrated how the word ‘**aspirin**’ can be searched for in Greek and Finnish.

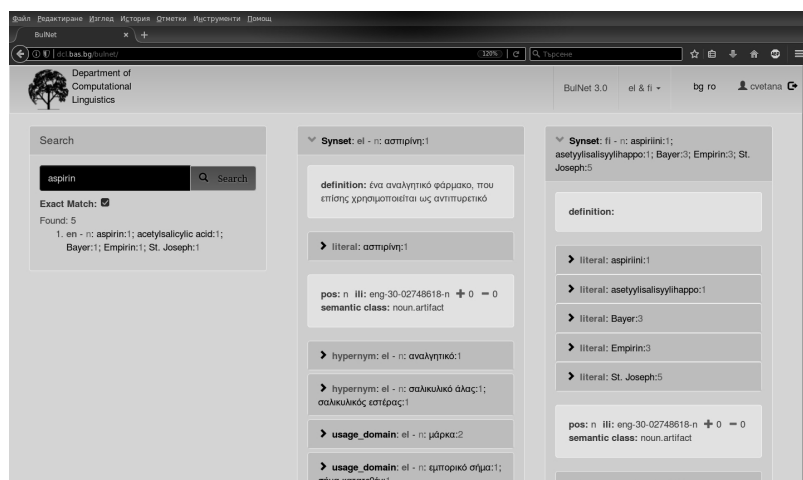


Figure 2: Hydra for Web – WordNet as a Multilingual Dictionary

The selected result – the first one from the list below the Search pane – is propagated to the right hand side visualizer(s). Hydra for Web supports two visualization modes: single mode, and bilingual mode (as on Fig. 2) where you see the correspondences of the selected synset (‘**aspirin**’ in English) in the mode’s languages (in Greek and in Finnish – to the left and to the right, respectively).

Every object visualisation is recursive in a sense that every relation (hypernym, holo_part, etc.) that leads to other object (i.e., synset) is expandable in the same way as the root one. The data in objects like pos, ILL, etc., are available immediately, while the relations are loaded by means of AJAX query, but asynchronously without blocking the UI.

2.1. Search

The tool allows searching for an exact match of a word string – a single word such as [**aspirin**], or a multiword unit, e.g., [**aspirin powder**] as in Fig. 3, or a non-exact match search which returns any synset where the searched word is found.

Although the three types of object in our approach to wordnet are fully-fledged, the search panel returns all the synsets that contain a literal matching the search query.

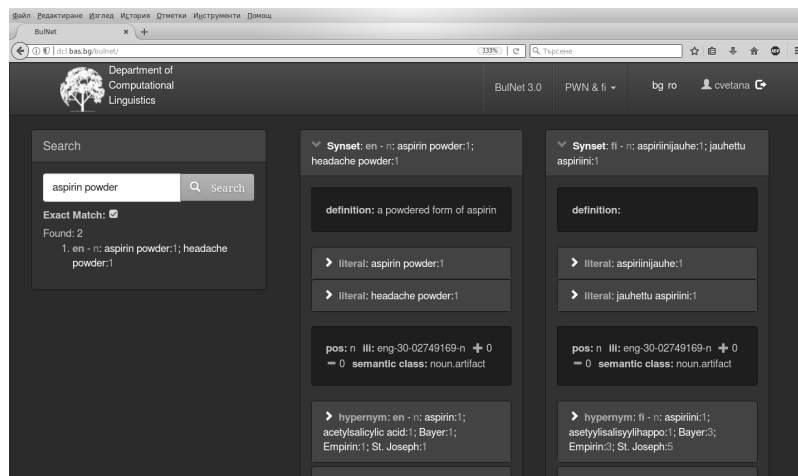


Figure 3: Hydra for Web – Multiword Unit

The search input is enhanced with autocomplete (with prefix match) as shown on Fig. 4 where synsets with the string ‘**powde**’ are shown while typing the word **powder**.

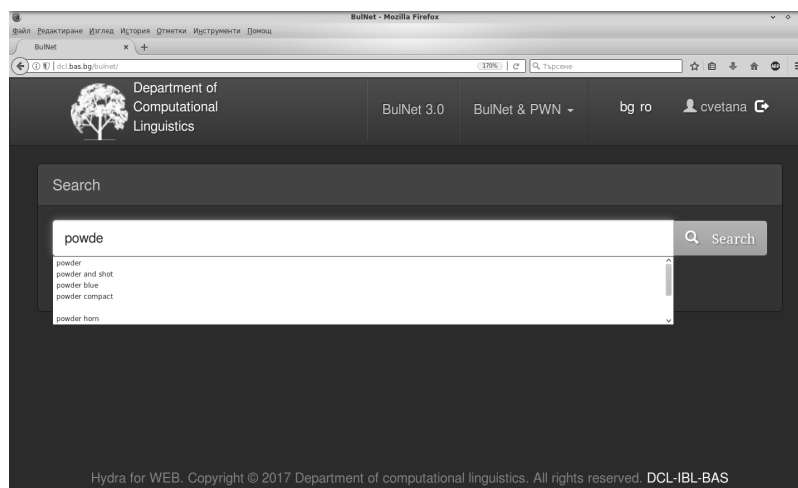


Figure 4: Hydra for web – autocomplete

The search returns a paginated list with the respective synsets in the database with the results shown at once being limited to 30 synsets in a list below the Search input but the user can navigate by using the button Next and Previous to browse between the pages with results.

To limit the results shown, the search respects word (string) boundaries, i.e., the user can search only for whole words but not parts of the words.

In the next section, we will present the editing functionalities of the tool.

2.2. Editing

Users can work on any wordnet (language) once it is added in the database.

The online editor Hydra for Web allows to: edit object’s data (some of the fields require free text like definition, while others are with predefined value list – f.ex. the part-of-speech); add object (literals and notes are added by button clicks in the parent objects); delete object; add binary relation between existing objects.

The editing functionalities of the tool will be illustrated by showing the process of editing the synset [kaysiya, kaysievo darvo] in the Bulgarian wordnet. This is automatically generated synset which has

not been validated yet as Bulgarian wordnet has been expanded by automatically translating a number of synsets (literals, definitions, and usage examples) to be further edited by experts.

When the user searches for a word (literal), the tool shows its status – literals that have been validated by an expert, are visualised in the standard color (white in Fig. 5), while those that have not been validated yet, are dimmed (muted).

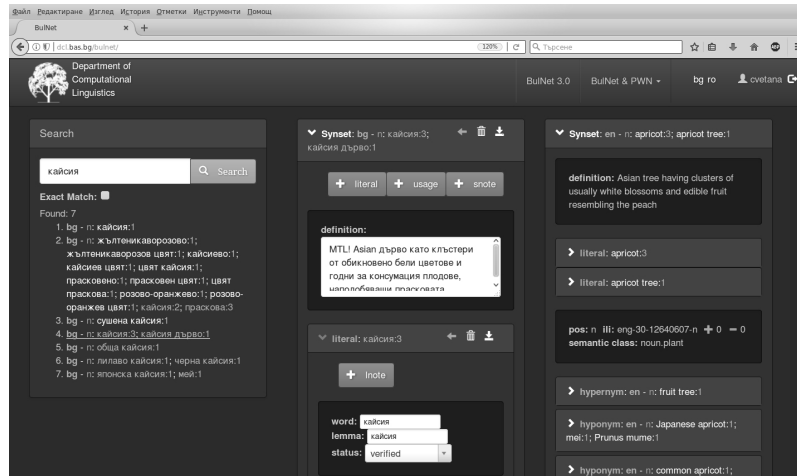


Figure 5: Hydra for Web – online editing

A synset can be edited by clicking on the top right-edge Edit button of the panel to put the linguistic unit (Synset, Literal or Note) panel in Edit mode – the data visualisation controls are replaced with those for editing.

The Edit panel for a synset consists of subpanels for the elements in the synset which are at least four: the set of literals constituting a synset; the definition; the literals visualised as a list one below the other (each literal can be edited as an independent object as shown on Fig. 5); and information for the current synset only – part-of-speech (pos), ILI, sentiment values according to SentiWordNet, semantic prime.

From top to bottom, the following elements are part of the editor panel (for Synset object) as shown on Fig. 5:

1. Panel header – textual representation of the synset – all the literals to the left, followed by buttons for canceling (the arrow sign), deleting (the ‘bin’ sign), and saving the synset.
2. Three buttons for adding (with the plus sign) [literal], [usage] and [snote] relations of the synset.
3. The definition.
4. The literals – each can be edited independently by clicking on the Edit button and opening an Edit panel which is much like the Editor panel of the parent synset. By clicking on the literal – without opening the Edit panel – the user can view the whole information about the literal at hand (word, lemma, status, and [Inote] plus the entire synset it pertains to below).
5. Information about: pos, ILI, sentiment values according to SentiWordNet, and semantic class. All values of these categories are editable – pos, SentiWordNet values, and semantic class are available as a list with fixed values (these are shown on the English synsets to the right in Fig. 5).

Other elements are: usage ((snote)), relations such as hypernym, hyponyms, derivational relations, morphosemantic relations, and others (see the example with the verb ‘cook’ on Fig. 6 below).

The synsets to which a currently edited synset is linked to via a relation (hypernym, hyponym, etc.) are given as a list after the subpanel (5) and each of the linked synsets can be edited further on its own.

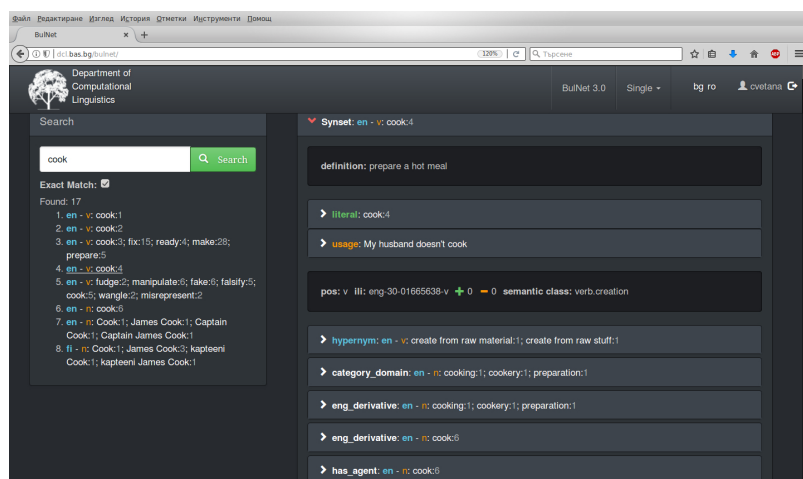


Figure 6: Hydra for web – whole synset, single mode

2.3. Linking

The Linguistic Units can be connected by introducing a relation between the two connected LUs. It is accomplished by means of a Wizard. To start it, the user clicks on the Connect button to the left of the Edit button on the unit panel. The procedure requires the following steps:

Step 1: A new Select Relation panel is opened to replace the Search panel. The new panel offers a list of all the relations available for the selected type of LU.

Step 2: The target LU of the relation is shown via a Search panel identical to the main Search panel. The search returns a list of synsets to be linked to the selected synset. Fig. 7 shows a selection of the $\langle is_agent_of \rangle$ relation that has to link the synset [kovach:1; zhelezar:1] ‘blacksmith’ to the synset [kova:2] ‘forge, hammer’. The selection of a target synset from the searched for list in the Search panel shows the whole synset below the list in the Search panel. If this is the intended synset, the user clicks on the button Connect and the link is visualized on the panel to right.

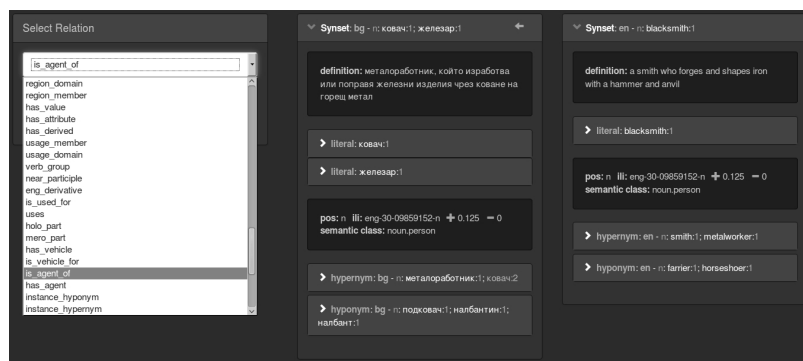


Figure 7: Hydra for web – selection of elements

3. Concurrent Editing

All modified data is propagated to the other connected users immediately by means of notifications by the wordnet server. In case of a conflict (when, for example, the same object is edited by more than one user), the last user is responsible for merging the data. When receiving a notification that some data is in edit mode, Hydra puts it in merge mode. If there is a possibility for not receiving a notification (i.e., there is a network problem), different strategy can be implemented.

Every client (web page instance) stores locally the copy of the data to be edited, and this copy is

submitted to the server alongside with its modifications. The server detects the conflict if any, and notifies the client, which puts the data in merge mode.

4. Users

Hydra for Web is freely accessible to all. Anonymous users can view (search and browse) the language data in the 23 wordnets in the database (but cannot edit the data – this option is available only to users with specific privileges). Additionally, the system is enhanced with user management with the following privilege options for every given language/wordnet:

- None: The wordnet is unavailable to the user.
- View: The user can search and browse this wordnet.
- Edit: The user can edit the data and relations in this wordnet.

Hydra for Web is mobile-friendly on a small width (mobile) (exemplified on Fig. 8 with the Romanian synset **agă:1**), where the panels are ordered successively.

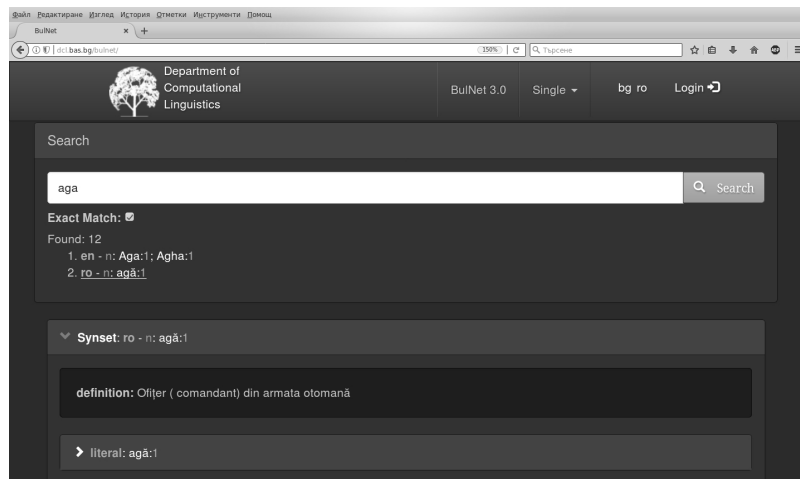


Figure 8: Hydra for web – selection of elements

5. Implementation

Hydra for Web is implemented by means of modern web technologies and libraries. Its source code is relatively small, straightforward and easy to maintain and extend *Hydra for web* with new features.

Hydra for Web is built with Node.js³ and Express⁴. It is a single page application and uses one of the most popular HTML, CSS and JS frameworks – Bootstrap⁵.

Hydra for Web is themed in Slate from Bootswatch⁶. Bootstrap makes easy the GUI to be responsive, and so it is mobile friendly.

For the html rendering, the very clean and elegant PUG template engine⁷ (formerly known as Jade) is used.

Many of the tasks in the GUI are solved in the client with the use of Knockout.js⁸ framework. It uses declarative bindings, dependency tracking and provides automatic UI refresh.

³Node.js® is a JavaScript runtime: <https://nodejs.org/>

⁴Web application framework for Node.js <http://expressjs.com/>

⁵<http://getbootstrap.com/>

⁶<https://bootswatch.com/>

⁷<http://jade-lang.com/>

⁸<http://knockoutjs.com/>

The wordnet data retrieval is made by means of the Wordnet Service. The retrieval uses AJAX and is completely asynchronous (non-blocking).

All the notifications of edited data are propagated to the the other users with SockJS which gives a coherent, cross-browser, Javascript API which creates a low latency, full duplex, cross-domain communication channel between the browser and the web server.

References

- Fellbaum, C. (1999). *WordNet: an Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Borislav Rizov. 2008. *Hydra: A Modal Logic Tool for Wordnet Development, Validation and Exploration.. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, ELRA.
- Borislav Rizov and Tsvetana Dimitrova. 2014. *Hydra for Web: A Browser for Easy Access to Wordnets.. Proceedings of the Global Wordnet Conference*, Tartu, pages 339–343.
- Dusza, K., Byczkowski, L., and Szymanski, J. (2007). Cooperative editing approach for building wordnet database. *Proceedings of the XVI international conference on system science*, 448–457.
- Miller, G. (1995). Wordnet: A Lexical Database for English. *Communications of the ACM*, November 1995, 38(11): 39–41.
- Vossen, P. (2002). WordNet, EuroWordNet and Global WordNet. *Revue française de linguistique appliquée*, 1/2002 (Vol. VII), 27–38.

The Effect of Unobserved Word-Context Co-occurrences on a Vector-Mixture Approach for Compositional Distributional Semantics

Amir Bakarov

The National Research University Higher School of Economics,
Federal Research Center ‘Computer Science and Control’ of Russian Academy of Sciences,
Moscow, Russia
amirbakarov@gmail.com

Abstract

Swivel (Submatrix-Wise Vector Embedding Learner) is a distributional semantic model based on counting point-wise mutual information values, capable of capturing word-context co-occurrences in the PMI matrix that were not noted in the training corpus. This model outperforms mainstream word embedding training algorithms such as Continuous Bag-of-Words, GloVe and Skip-Gram in word similarity and word analogy tasks. But the properness of these intrinsic tasks could be questioned, and it is unclear if the ability to count unobservable word-context co-occurrences could also be helpful for downstream tasks. In this work we propose a comparison of Word2Vec and Swivel for two downstream tasks based on natural language sentence matching: the paraphrase detection task and the textual entailment task. As a result, we reveal that Swivel outperforms Word2Vec in both cases, but the difference is minuscule. We can conclude, that the ability to learn embeddings for rarely co-occurring words is not so crucial for downstream tasks.

1. Introduction

Distributional semantic models (DSMs) are instruments which can represent words through real-valued vectors of fixed dimensions. The word *distributional* here is a reference to a *distributional hypothesis* saying that word semantics is distributed together with its contexts (Harris, 1954). DSMs can capture various functional or topical relations (for example, *semantic similarity* also known as *synonymy*) between words through collocation contexts observed in a corpus, and the strength of such relation between two words can be computed as a distance between vectors corresponding to these words.

There are two taxonomic classes of DSMs. The first one is based on counting word co-occurrences in a corpus (for example, through constructing a TF-IDF matrix or a PMI matrix). Each word could be represented as a sparse vector, and its dimensionality can be lowered by applying dimensionality reduction techniques (like SVD) to the matrix – this is how some ‘classic’ distributional models like LSA or PPMI matrix work (Landauer et al., 1998; Turney and Pantel, 2010). They are called *count-based models*.

Another option of capturing word semantics is based on sampling the training corpus with a sliding window while each word is initialized with a vector, which values are optimized for the task of predicting a word using its context (or, *vice versa*, predicting context using the word) (Collobert et al., 2011). DSMs that work in such a way are called *predictive models*, and the dense vectors produced by such models are called *word embeddings*.

Predictive models are the most mainstream class of DSMs since they proved their effectiveness in most NLP tasks. One of the most popular DSM models is *Word2Vec* (Mikolov et al., 2013) which provides two training architectures: *Continuous Bag-of-Words* (CBOW) that predicts words given their contexts, and *Skip-Gram* (SG) that predicts the contexts from the words.

The effectiveness of predictive models was revealed in the classic paper *Don’t Count, Predict!* (Baroni et al., 2014), which proved the benefits of predictive models against count-based models. However,

some researchers still try to propose novel approaches to count-based training that could pretend to outperform the predictive approach. One of the most recent interesting methods was introduced in a novel DSM called *Swivel* (which is an abbreviation for *Submatrix-Wise Vector Embedding Learner*) (Shazeer et al., 2016). This model is based on applying SVD for PMI matrix, and the main idea is to use a loss function which penalties depend on whether the word-context pair co-occurs in the corpus or not, so the algorithm could be trained to not to over-estimate PMI of common values whose co-occurrence is unobserved. Notably, Word2Vec with a negative sampling is also capable of taking unobserved co-occurrences into account, but it is done indirectly.

The central claim of the authors of *Swivel* is that none of the mainstream word embeddings provide any special treatment to unobserved word-context co-occurrences (Shazeer et al., 2016), so the ability to capture unobserved word-context co-occurrences helped to outperform other embedding training algorithms in word similarity and word analogy tasks. But the properness of the considered intrinsic evaluation tasks (word similarity and word analogy) could be questioned due to the recent negative critique of word similarity and word analogy as methods for evaluating DSMs (Batchkarov et al., 2016; Gladkova et al., 2016). Hence, it is unclear if the ability to count unobservable word-context co-occurrences could be also helpful in downstream tasks and allow the count-based models to outperform the predictive ones.

To this end, in this paper we want to consider a more proper evaluation method that introduces two downstream tasks based on *natural language sentence matching*. Our evaluation methods rely on testing the ability of the compared DSMs to build vectors of compositional linguistic units (like sentences and documents). Different techniques of vector composition allow to represent these units in a vector space and calculate a similarity value between vectors corresponding to them.

The scientific question that raised in this work is whether taking unobserved word-contexts co-occurrences into account could help count-based models to outperform the neural-based ones in a downstream task. *Our main contribution* is the comparison of performance of three different DSMs on three datasets in order to answer this question. Our work is the first towards an evaluation of *Swivel* and *Word2Vec* on extrinsic tasks.

We think that evaluation in a downstream task could also reveal differences between the methods that do not seem evident when only an intrinsic task is used. On the other hand, we see our next contribution in raising an issue of whether intrinsic evaluation is enough to make a conclusion about DSMs performance since its performance on an extrinsic task could differ. The results obtained in this study prove the existence of such issue.

The paper is organized as follows. Section 2 describes the background of tasks that we use for our downstream evaluation. Section 3 provides a brief introduction to the background of compositional distributional semantics and the approach for obtaining compositional unit representations which we consider in this paper. Section 4 is dedicated to the experimental setup, while section 5 covers the results of an evaluation and brings a discussion on them. Section 6 mentions recent studies that we find relevant to our topic. Section 7 concludes the article.

2. Downstream tasks

2.1. Paraphrase detection

Paraphrase is a restatement of a text giving the same semantic meaning in another textual form. *Paraphrase detection task* (which could be also mentioned as a *semantic similarity identification task*, *sentence similarity detection task* since the paraphrased sentences are, in substance, semantically similar expressions (Xu et al., 2015)) for a pair of natural language sentences T_1 and T_2 , is a task of identification whether T_1 and T_2 have the same semantic meaning. In this case, the pair of sentences could be only related to one of two labels:

- **semantic similarity exists:** *If you help the needy, God will reward you & You will receive the reward of God for your charity;*
- **semantic similarity does not exist:** *If you help the needy, God will reward you & You will receive the reward of God for your charity.*

2.2. Textual entailment

Textual entailment (also called *natural language inference*) is a concept describing a directional relation between two text fragments which holds whenever the truth of one text fragment (called hypothesis) follows from another text (called text) (Dagan et al., 2006). Therefore, a pair of natural language sentences T and H contains textual entailment if a human reading text would infer that H directly follows T , and T does not directly follow H . So one should notice that the textual entailment connection is directional unlike the connection of other types of semantical relatedness like paraphrasing. The pair of sentences can be labeled with three types of available entailment relations:

- **entailment:** the hypothesis entails the text if the relations or events described by the hypothesis are likely to be true given the relations or events described by the text (e.g. text: *If you help the needy, God will reward you*, hypothesis: *Giving money to a poor man has good consequences*);
- **contradiction:** the hypothesis contradicts the text if the relations or events described by the hypothesis are highly unlikely to be true given the relations or events described by the text (e.g. text: *If you help the needy, God will reward you*, hypothesis: *Giving money to a poor man has no consequences*);
- **neutral relation:** the hypothesis and the text are in a neutral relation if the relations or events between them are not semantically connected to each other (e.g. text: *If you help the needy, God will reward you*, hypothesis: *Giving money to a poor man will make you a better person*).

3. Vector mixture

Most of DSMs (particularly, the ones which we consider in this work) rely only on lexical semantics and do not build representations for other levels of text such as sentences or documents. Semantic modeling for composed linguistics units (sentences or documents) based on a distributional approach is usually considered as a separate subfield called *compositional distributional semantics* (CDS) (Mitchell and Lapata, 2010). The purpose of a compositional DSM is to provide a function that could produce a vector representation of the meaning of composed linguistic units from the distributional vectors of the words contained therein. The motivation of such approach is that if a sentence is a function of meaning of all its words, then word embeddings could also be treated as the building blocks of compositional representation. While it has been shown that semantic relations can be mapped to translations in the learned vector space, the claim could be made for sentence representations of the embeddings.

Mainly, approaches for constructing compositional distributional representations are divided into *vector-mixture based* (they rely on element-wise arithmetic operations on vectors), *tensor-based* (they additionally represent neighbor words as matrices in order to build a sentence representation) and *network-based* (they consider construction of weights for the sentence vector as a neural network's objective) (Sadrazadeh and Kartsaklis, 2016).

Since the downstream tasks which we consider rely on natural language sentence matching, they are actually linked with the task of construction of sentence vectors. Hence, the performance of a DSM hypothetically depends on the chosen vector composition algorithm. We consider *vector mixture* the most robust and interpretable technique of CDS modeling (Mitchell and Lapata, 2008), because while neural networks and additional matrices (from tensor-based and network-based approaches) could introduce bias in the obtained results due to their stochastic nature and abundance of new parameters in the algorithm (it is impossible to say how much of the conclusions would due to the neural network architecture itself, and how much to the optimization function, and how much to the first initialization of the weights), the beautiful simplicity of element-wise operations on word vectors guarantees that only the ability of the word-level distributional model to construct adequate representations is taken into account. Of course, vector mixture approach also has certain limitations (for example, it does not consider word order in the sentence, treating sentence like an unordered bag-of-vectors), and its linguistic justification could be possibly questioned, but the idea of relevance of operations on vectors for obtaining meaning transformation was justified by other tasks using vector operations like analogical reasoning (Le and

Mikolov, 2014). The possible justification of mixture models could be explained by the fact that if the vector of a word shows the extent to which this word is related to other words in the corpus, so will the compositional vectors show the extent to which things related to a certain vector can also be related to other vectors (Zhang et al., 2018).

Two main options of vector mixture include vector composition and vector multiplication. Feature-wise vector addition can be seen as feature union, and vector multiplication as feature intersection. In this work we will obtain compositional distributional representations through vector mixture approach based on element-wise vector addition. Treating a sentence as a bag of words, we will obtain its vector as an average of all vectors corresponding to all words it contains.

Our idea proposes that it is possible to use components V as features for learning the automatic classifier. So the list V_1, \dots, V_t of t sentence pairs could be used as a feature matrix for learning the classifier, and a vector i_1, \dots, i_t reporting the type of the relation between the sentences could be used as a target vector.

More formally, given two sentences, X with n words and Y with k words, one could obtain their vector representations, $S(X)$ and $S(Y)$, by averaging the vector of words that constitute the sentences:

$$S(X) = \frac{s(x_1)+s(x_2)+\dots+s(x_n)}{n}$$

$$S(Y) = \frac{s(y_1)+s(y_2)+\dots+s(y_k)}{k}$$

And then obtain vector representation of the two sentences:

$$V = \frac{S(X)+S(Y)}{2}$$

Therefore, such embedding of a sentence pair could be used to train the classifier to distinguish the existence of paraphrasing or the existence of entailment.

4. Experimental setup

As we mentioned before, our experimental setting includes the comparison of DSMs with different architectures trained on the same corpus with the same hyperparameters. As the training algorithms we use implementations of SG, CBOW and Swivel from the official repository of Tensorflow (Abadi et al., 2016), the most popular framework for deep learning. Our choice for the training data was the Gutenberg Project corpus (Lahiri, 2014) of 520 000 tokens containing English fiction. We used a filtered and lemmatized version of the corpus (lemmatization was done with UDPipe (Straka et al., 2016)). The main concern in the use of project Gutenberg corpus is that it contains a lot of words that are not used in our datasets, so we suppose that the use of fiction corpus may answer the question whether the ability of Swivel to use unobserved word-context co-occurrences could have an effect on the results since we consider that in all tasks related to distributional semantics the choice of the training corpus is highly decisive.

Each of the considered models was trained with context window of 10 and sub-sampling of 1e-3 (because these hyper parameters worked better among others that were compared); if it was possible, we turned on the negative sampling regime. For each of the training algorithms we trained a model with varying vector dimensions of 100, 150, 200, 250, 300, 350, 400, 450 and 500 (the highest boundary for vector size is explained by the fact that it is the threshold for which in our experiments most of the models stopped to increase performance significantly). We decided to check different vector dimensions since we suppose that in element-wise vector mixture the size of the vector could be crucial if we want to take into account all the data contained in vector components. All in all, we had 27 models: 3 different training algorithms and 9 vector sizes for each.

The datasets on which the models were evaluated are:

- **Sentences Involving Compositional Knowledge (SICK)**, 9 840 pairs of sentences assessed by textual entailment (Bentivogli et al., 2016);
- **Stanford Natural Language Inference (SNLI)**, 640 000 pairs of sentences assessed by textual entailment (Bowman et al., 2015);

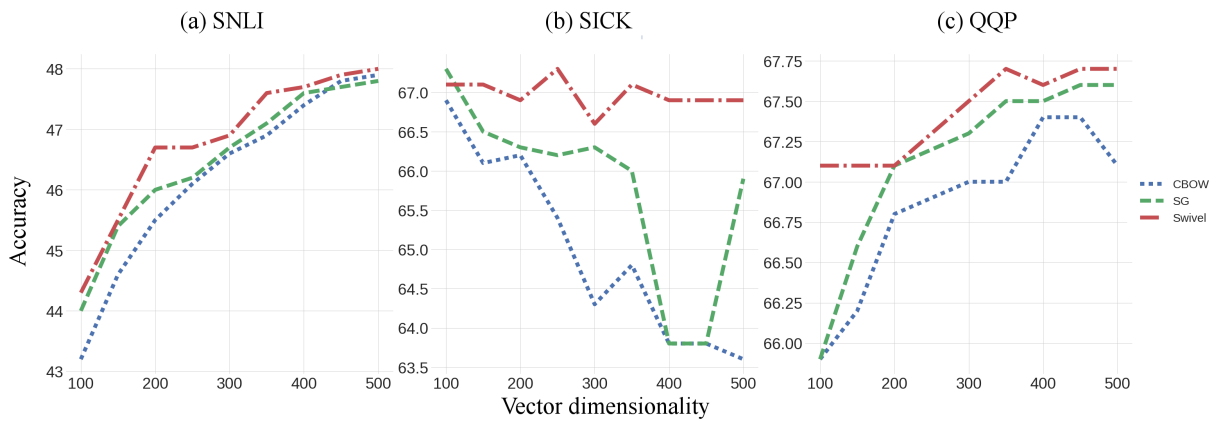


Figure 1: Accuracy of Swivel, CBOW and SG (in percents) on the considered datasets varying the dimensionality of word vectors.

- **Quora Question Pairs (QQP)**¹, 400 000 pairs of sentences assessed by semantic similarity.

The datasets were also lemmatized using the aforementioned UDPipe.

The method of capturing compositional distributional representation of the sentences included averaging all word embeddings of the sentence, as was mentioned in the previous section (out-of-vocabulary words were not taken into account). Then each dataset was represented as a labeled pair of vectors (binary labels in the case of the paraphrase detection task or multi-class nominal labels in the case of the textual entailment task).

These labels were used as target values, and each dataset was divided into two parts (in proportion $\frac{3}{1}$): one for training the classifier, and one for obtaining the final result of performance of given word embedding models by the classifier’s prediction. For classification we used the *Logistic Regression* model implemented in a *Scikit-learn* module (Pedregosa et al., 2011). We also tested other popular classification algorithms (*Naive Bayes*, *Random Forest*, *K-Nearest Neighbors*, *Support Vector Machine*, and so on), but they were not able to outperform Logistic Regression (we do not mention the obtained results on these algorithms here since we find that they could blur the focus of our paper).

5. Results and discussion

	SNLI	SICK	QQP
1	Swivel	Swivel	Swivel
2	SG	SG	SG
3	CBOW	CBOW	CBOW

Table 1: Ranking of the compared models across the considered datasets by the best result shown on any vector dimensionality.

We evaluated 27 models on the test chunk on SNLI (a), SICK (b) and QQP (c), and calculated accuracy varying the vector dimensionality (we used accuracy since the datasets were balanced, and this measure was able to report actual quality of the classification). Figure 1 illustrates the results of the evaluation while vector dimensionality of the compared models varies. The overall relative rankings of compared models and their best results on each of datasets are presented at Table 1. Swivel showed the best performance on all tasks (48% of correct answers on SNLI, 66.9% on SICK and 67.2% on Quora).

Such result shows that Swivel actually works better than Word2Vec (although the difference in the results is small). This could be explained by the fact that the style used in project Gutenberg corpus

¹<https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs>

differs from the one used in the proposed datasets. A claim could be made that this should just play to the strengths of Swivel, since it is supposedly strong for rarely co-occurring words. The results on all the datasets are far from state of the art (Conneau et al., 2017) (89.3% on SNLI (Tay et al., 2017), 80% on SICK (Bentivogli et al., 2016), 88.1% on QQP (Wang et al., 2017)). Anyway, the main aim of our work was not to propose the approach with the best performance, but to make a comparison of DSMs involving a robust and interpretable method, and we suppose that the performance gap is caused by the algorithm, not by the corpus – averaging word embeddings for obtaining sentence embeddings should work worse for long (more than approximately 10 words) sentences than sentence-level embeddings like *Skip-Thought* (Kiros et al., 2015) that were used in other papers. And this may be the cause why the SICK task demonstrates considerable changes in accuracy for different dimensions – sentences in SICK dataset are notably shorter than in two other datasets.

Another interesting result that we obtained is that models' performance in most cases increased with increasing vector dimensionality: for example, the dependence of accuracy of vector dimensionality for SNLI could be approximated with an ascending function. On the other hand, on SICK the performance of CBOW and SG sharply decreases on a threshold of 300 while the accuracy with Swivel did not decreased. The possible explanation is that SICK contains lexemes that rarely occurred in the training corpus, and the use of unobserved word-context co-occurrences helped Swivel to outperform other models.

The main limitation of Swivel that we see is that it requires notably more resources for training than Word2Vec. Training of Swivel on our small cluster equipped with ASUS GeForce GTX Titan X took up to 4 hours versus 10 minutes on CBOW and 30 minutes on SG. We can assume that in some tasks the difference in the results can be less notable feature than the time for training the model.

6. Related work

Exploratory research of DSMs was an object of big interest from the NLP community in recent years, and this interest is increasing with the popularity of word embeddings. Various researchers investigated different aspects of DSM, such as the underlying latent semantic structure (Senel et al., 2017), the effect of the chosen corpora (Kutuzov and Kunilovskaya, 2017) or the training algorithm (Bakarov and Gureenkova, 2017), the topology of gender (Bolukbasi et al., 2016), the nearest neighbors of frequent words (Radovanović et al., 2010), the types of semantic relations (Rei and Briscoe, 2014; Melamud et al., 2014) and so on. Particularly, in our work we actually put the effect of the training algorithm as a primary object of our research. To measure the performance we evaluated how well the considered model could work with two tasks based on natural language sentence matching: the paraphrase detection task and the textual entailment task.

The early approaches to a paraphrase detection task did not considered semantic structure at all, taking into account only word- and subword-levels (Dolan et al., 2004). Later, some researchers started to use manually constructed thesauri like WordNet (Lee and Cheah, 2016). Nowadays the state-of-the-art methods rely on bidirectional deep neural networks; for example, BiMPM (Wang et al., 2017). As the most comprehensive work on the task of paraphrase detection as well as textual entailment we consider (Androustopoulos and Malakasiotis, 2010). Due to the fact that the task of textual entailment is also strongly linked with the task of matching natural language sentences, a lot of approaches for this task are highly similar to the approaches for paraphrase detection. All in all, these two tasks are highly popular in word embeddings community and are used as subtasks of more global tasks, like information retrieval, plagiarism detection, and more (Zubarev and Sochenkov, 2017).

7. Conclusions

Our experiments demonstrate that Swivel outperforms Continuous Bag-of-Word and Skip-Gram in modeling compositional distributional semantics on the variety of tasks of natural language sentence matching (however, the difference in the performance in most considered cases is tenths of a percent). We conclude that taking into account unobserved word-context co-occurrences plays a certain role in downstream tasks like the ones considered in the presented study.

Since we have not concluded which method is basically better we think that it will be reasonable to

use some kind of retrofitting technique for refining the word vectors in future (Faruqui et al., 2015) to be able to prove our hypothesis that fiction corpus should provide better results, so one could unearth how much does the corpus size and choice affect the embeddings.

In the future work we also wish to reproduce the results on the non-English data: for instance, on *Bulgarian language*. It is interesting to see whether the ranking positions of the compared algorithms will be reproduced. However, now we are not able to make such comparison since we are not aware of any paraphrase detection (or textual entailment dataset) for Bulgarian (despite the task of textual entailment was considered at the *Computational Linguistics in Bulgaria* conference, but not from the perspective of Bulgarian (Dias and Moraliyski, 2014)). Creation of such benchmark for Bulgarian is also in our plans for future work on this topic. In contrast to English, Bulgarian is a highly fusional language, and we think that research of word embeddings evaluation on Bulgarian could give many interesting insights for researchers in the field of evaluation of word embeddings as well for the international NLP community.

Acknowledgements

We thank our colleague, Andrey Kutuzov, for productive discussion on this paper. We also want to thank three anonymous reviewers for attentive reviewing and helpful suggestions.

References

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.
- Androutsopoulos, I. and Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.
- Bakarov, A. and Gureenkova, O. (2017). Automated detection of non-relevant posts on the russian imageboard “2ch”: Importance of the choice of word representations. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 16–21. Springer.
- Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don’t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 238–247.
- Batchkarov, M., Kober, T., Reffin, J., Weeds, J., and Weir, D. (2016). A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 7–12.
- Bentivogli, L., Bernardi, R., Marelli, M., Menini, S., Baroni, M., and Zamparelli, R. (2016). Sick through the semeval glasses. lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. *Language Resources and Evaluation*, 50(1):95–124.
- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357.
- Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.
- Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.
- Dagan, I., Glickman, O., and Magnini, B. (2006). The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising textual entailment*, pages 177–190. Springer.

- Dias, S. P. G. and Moraliyski, R. (2014). Unsupervised and language-independent method to recognize textual entailment by generality. In *First International Conference Computational Linguistics in Bulgaria (CLIB 2014)*.
- Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics*, page 350. Association for Computational Linguistics.
- Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., and Smith, N. A. (2015). Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615.
- Gladkova, A., Drozd, A., and Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.
- Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.
- Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.
- Kutuzov, A. and Kunilovskaya, M. (2017). Size vs. structure in training corpora for word embedding models: Araneum rassicum maximum and russian national corpus. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 47–58. Springer.
- Lahiri, S. (2014). Complexity of word collocation networks: A preliminary structural analysis. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 96–105.
- Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.
- Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.
- Lee, J. C. and Cheah, Y.-N. (2016). Paraphrase detection using semantic relatedness based on synset shortest path in wordnet. In *Advanced Informatics: Concepts, Theory And Application (ICAICTA), 2016 International Conference On*, pages 1–5. IEEE.
- Melamud, O., Dagan, I., Goldberger, J., Szpektor, I., and Yuret, D. (2014). Probabilistic modeling of joint-context in distributional similarity. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 181–190.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. *proceedings of ACL-08: HLT*, pages 236–244.
- Mitchell, J. and Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.
- Radovanović, M., Nanopoulos, A., and Ivanović, M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531.
- Rei, M. and Briscoe, T. (2014). Looking for hyponyms in vector space. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 68–77.
- Sadrzadeh, M. and Kartsaklis, D. (2016). Compositional distributional models of meaning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 1–4.
- Senel, L. K., Utlu, I., Yucesoy, V., Koc, A., and Cukur, T. (2017). Semantic structure and interpretability of word embeddings. *arXiv preprint arXiv:1711.00331*.

- Shazeer, N., Doherty, R., Evans, C., and Waterson, C. (2016). Swivel: Improving embeddings by noticing what's missing. *arXiv preprint arXiv:1602.02215*.
- Straka, M., Hajic, J., and Straková, J. (2016). Udpipeline: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *LREC*.
- Tay, Y., Tuan, L. A., and Hui, S. C. (2017). A compare-propagate architecture with alignment factorization for natural language inference. *arXiv preprint arXiv:1801.00102*.
- Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.
- Wang, Z., Hamza, W., and Florian, R. (2017). Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814*.
- Xu, W., Callison-Burch, C., and Dolan, B. (2015). Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 1–11.
- Zhang, R., Guo, J., Lan, Y., Xu, J., and Cheng, X. (2018). Aggregating neural word embeddings for document representation. In *European Conference on Information Retrieval*, pages 303–315. Springer.
- Zubarev, D. and Sochenkov, I. (2017). Paraphrased plagiarism detection using sentence similarity. *Dialog*.

Introducing Computational Linguistics and NLP to High School Students

Rositsa Dekova

Plovdiv University Paisii
Hilendarski

rosdek@uni-plovdiv.bg

Adelina Radeva

Sofia University St.Climent
Ohridski

radeva.adelina@gmail.com

Abstract

The paper addresses a possible way of introducing core concepts of Computational Linguistics through problems given at the linguistic contests organized for high school students in Bulgaria and abroad. Following a brief presentation of the foundation and the underlying objective of these contests, we outline some of the types of problems as reflecting the different levels of language processing and the diversity of approaches and tasks to be solved. By presenting the variety of problems given so far through the years, we would like to attract the attention of the academic community to this captivating method through which high school students might be acquainted with the challenges and the main goals of Computational Linguistics (CL) and Natural Language Processing (NLP)¹.

1. Introduction

The Bulgarian linguistic contests for high school students² were founded back in the 1980s by Prof. Ruslan Mitkov (Mitkov, 2006a). Called initially “Competitions in Mathematical and Computational Linguistics”, these contests targeted highly motivated students, primarily from Mathematical and Language High Schools, with the primary goal of getting students interested in linguistics as a whole and introducing them to some of the core concepts of computational linguistics, thus providing “a promising springboard for future careers in NLP” (Mitkov, 2006a). And they did so, indeed. Many students have acquired their initial knowledge of what computational linguistics is mainly through solving such problems. Some of those students have in turn grown up to be computational linguists and carry on the work in other countries. Most notably, Dragomir Radev – a professor of Computer Science at Yale University (working on natural language processing and information retrieval), author of many problems in computational linguistics (Radev, 2013a, 2013b) given at the National Olympiads in the USA, Canada, Australia, etc. But there are many other Bulgarians, such as Yova Kementchedjhieva (MSc in Cognitive Science Informatics, University of Edinburgh, currently a PhD student in CL, University of Copenhagen), Nikolay Bogoychev (BSc in Artificial Intelligence and CS, University of Edinburgh, PhD student in Informatics at the University of Edinburgh, exploring the application of GPUs in NLP), Dimitar Hristov (BSc in Computer science, University of Southampton, Master’s Degree in CL, Sofia University, currently a researcher at the Department of Computational Linguistics, Bulgarian Academy of Sciences), Lyubomir Zlatkov, Pavel Sofroniev, and Stela Ilieva (Bachelor Degree in CL, University of Tübingen), Ivaylo Grozdev (Bachelor Degree in CL, University of Edinburgh), Todor Tchervenkov (Linguistics, Computer Science, Trinity College, Dublin), to mention just a few.

¹ We would like to thank the anonymous reviewers for their valuable remarks and comments.

² Throughout the paper, we will also use ‘students’ and ‘undergraduate students’ interchangeably to ‘high school students’.

Later, the name of the contests in Bulgaria was shortened to “Competitions in Mathematical Linguistics”, but CL oriented problems continued to be included whenever possible.

An important feature of all linguistic problems, designed for these competitions, is that they are created as self-sufficient, i.e. anyone could solve them with no prior theoretical knowledge, therefore allowing students to explore the fundamental concepts and guiding rules on their own, which is far more fascinating than learning theory by heart.

The paper aims at outlining the most general types of NLP tasks and CL applications presented as problems given at different events for undergraduate students such as seminars, summer schools, competitions, national and international Olympiads. Thus, we would like to showcase this alternative method of introducing computational linguistics to high-school students and engage more academics in the field.

2. Types of linguistic problems based on CL applications and NLP tasks

The problems are presented to the students in an accessible and entertaining way. Nevertheless, they outline samples of real NLP concepts, theoretic and practice examples of finite-state transducers, formal grammars and natural language generation, automatic text processing, incl. anaphora resolution, summarization, word sense disambiguation and word sense representation, machine translation, etc. Many linguists have contributed to authoring problems for these contests: Ruslan Mitkov, Dragomir Radev, Ivan Derzhanski, Tom McCoy, Harold Somers, Christiane Fellbaum, Jonathan May, Patrick Littell, Emily Bender, Jonathan Kummerfeld, Tom Payne, Daniel Lovsted, Richard Sproat, Andrea Schalley, Aleka Blackwell, Adam Hesterberg, Ben King, among others.

For the purposes of this article the authors have reviewed and categorized the most frequently given types of CL problems, including problems given at the Bulgarian National Olympiad in Linguistics, North American Computational Linguistics Olympiad (NACLO), established in 2006, Australian Computational and Linguistics Olympiad (OzCLO), established in 2008, All Ireland Linguistics Olympiad (AILO), established in 2009, United Kingdom Linguistics Olympiad (UKLO), established in 2010³.

2.1. Finite-State Transducers / Finite-State Automata

The problems related to Finite-State Transducers (FST) or Finite-State Automata (FSA) present examples of machines with finite set of states, including initial state and one or more final states. The FST transduces strings of symbols into other strings of symbols. The machines can make successful and unsuccessful transformations using different data – alphabet symbols, numbers, words, etc. The students are given examples of FST and how the transducer works, whereby the task is to repeat a transformation using a set of given data, or to create new transformations.

FST/FSA can be applied on different levels of language processing – phonology, morphology, syntax, or used for coding a text, as shown respectively in 2.1.1, 2.1.2, 2.1.3, and 2.1.4 below.

2.1.1. Phonemic FST and FSA

The concept of the limited system of phonemes in a natural language and therefore the possibility of defining rules for generating words is demonstrated quite well in the problem “aw-TOM-uh-tuh”, created by Patrick Littell (NACLO 2008, OzCLO 2008)⁴ and also available in Bulgarian (Derzhanski and Velinov, 2012). The problem illustrates an FSA that can distinguish between possible and impossible words in Rotokas⁵, which possesses one of the smallest phoneme inventories. The problem represents the FSA as a board game (see Fig. 1), allowing students to understand how to generate possible words with a given set of phonemes and rules for their combination.

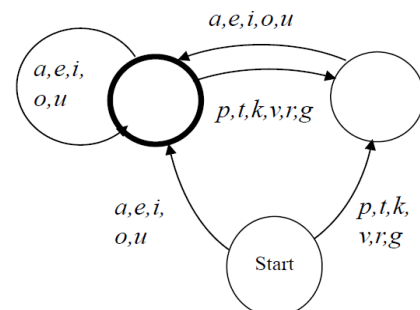


Figure 1: FSA for possible words in Rotokas

³ Past problems and practice samples could be found on most of the National Olympiads web-pages (cf. References).

⁴ Available online at: <http://www.nacloweb.org/resources/problems/2008/N2008-I.pdf>

⁵ An isolated language, spoken on the island of Bougainville, east of Papua New Guinea

Another problem demonstrating phonemic transformations with an FST is created by Tom Payne (NACLO Sample Practice Problems)⁶ and named “Computational Machines”. It shows a diagram of an FST that transforms the English word “cat” into the English word “dog” in three steps. It also provides an example of a machine that allows for an infinite number of inputs. Thus, the problem urges students to differentiate between the possible and the desired outputs. The task that should be accomplished by the students is to create a similar diagram that will transform “Tom Cruise” into “Ali Landry” using four circles or less.

2.1.2. Morphemic FST and FSA

The problem “Transition(al) numbers”, created by Harold Somers (NACLO 2017)⁷, illustrates the use of a morphemic FSA (presented as a “transitional network”) illustrating the set of rules generating the English numerals smaller than a hundred. The students also learn the concept of “overgeneration” and are asked to correct some of the rules or to create new ones in order to “fix” the network.

2.1.3. Syntactic FST and FSA

Finite-State Automata can also be used to illustrate simplified sentence generation. Three similar problems (Mitkov, 1989: 79-80) offer examples of such FSA, where every arc connecting two states represents a different word, and the generated sentences can be grammatically correct or incorrect. The students’ tasks are to propose an FSA using the same words but with different states and directions of the arcs, so that all the generated sentences are grammatically correct (compare the FTA presented to the students (Fig.2) with the FTA expected to be produced by the students as a solution to the problem in Fig. 3 below).

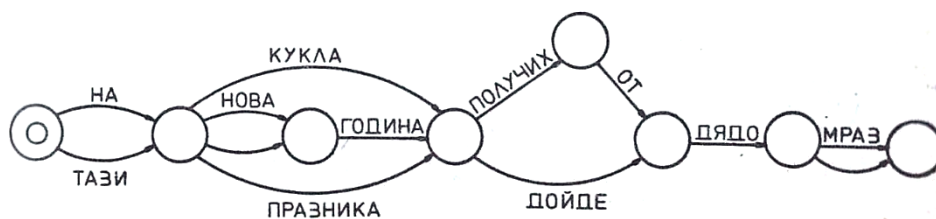


Figure 2: Sample syntactic FSA⁸

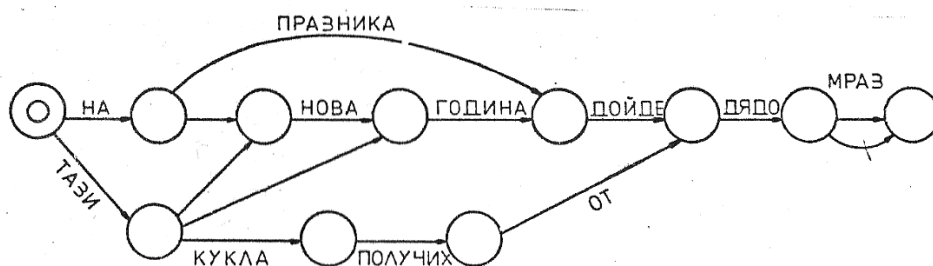


Figure 3: Problem Solution

2.1.4. FST and FSA for encoding text

Being an effective means in various types of automatic treatment of natural languages, Finite-State Transducers can also be used for coding and decoding texts. The problem “Finite-State Transducers” created by Richard Sproat (NACLO Sample Practice Problems)⁹ illustrates how the input alphabet describes a recognizable pattern that is transformed into the output alphabet, and how that can be applied for a text. The students’ task is to decode a sample of text, a simple kind of deciphering, using the provided output data, and a diagram of the FST used in the initial ciphering.

⁶ Available online at: <http://www.nacloweb.org/resources/problems/sample/FST-4.pdf>

⁷ Available online at: <http://www.nacloweb.org/resources/problems/2017/N2017-F.pdf>

⁸ Fig. 2 is used in a problem by I. Nenkova (Mitkov, 1989: 79-80) and Fig. 3 is given as a solution to the problem (ibid. 151).

⁹ Available online at: <http://www.nacloweb.org/resources/problems/sample/FST1.pdf>

2.2. Formal grammars and natural language generation

Problems presenting formal rules for generating words or sentences are another great illustration of introducing CL to students. The “Grammar Rules!” problem, created by Patrick Littell and Andrea Schaley (NACLO 2013, AILO 2013, OzCLO 2013)¹⁰ sets forth the notion of a Context free grammar presented as a set of phrase structure rules of the type ($S \rightarrow NP VP$, $NP \rightarrow N$, etc.). The students are then supplied with a number of sentences and their task is to select the sentences that are well formed according to this CFG, i.e. which can be generated by the given rules. Also, they must find the instances of overgenerated sentences, i.e. the ones that are well formed by the CFG rules but not grammatical (according to the rules of English). Finally, students must also detect the one rule that is redundant.

Another problem also employing a CFG is called “Fan Fiction” by Ben King (NACLO 2016, OzCLO 2016)¹¹. The story revolves around a fan-fiction writing robot who can use a few different methods for generating sentences, such as n-gram methods (unigrams, bigrams and trigrams) and a CFG. All the methods are illustrated by simple examples. The students are then provided with a collection of sentences, only three of which are real (written by a human author) and the rest are generated using one of the methods described. The students are to detect which sentences are real and which of the other methods has been used for each of the sentences generated by the bot.

Josh Falk introduces the students to the Horn clause notation (for ex. $S(xy) :- N(x), V(y)$.) in his problem “A Matter of Horn Clauses” (NACLO 2016)¹², where the students are supposed to use the notation to describe English and Swiss German sentences.

There are also other problems which apply formal rules for generating words, instead of sentences. Some examples of such problems are: “Text-o-matic” by Daniel Lovsted (NACLO 2017)¹³, which presents rules for generating the paradigm of French numerals; “Minimum Spelling Trees” by John de Nero (NACLO 2015)¹⁴, which involves encoding of German noun forms (generating the paradigm of a word); and “Lexicondensed” by Tom McCoy (NACLO 2014)¹⁵, which introduces formal lexicons for the task of creating Spelling Change Rules for generating a list of adjectival forms of country names.

2.3. Automatic text processing

2.3.1. Anaphora resolution

Anaphora resolution, on the one hand, still presents a challenge for NLP, but on the other hand, it can be transformed into a challenging problem for undergraduate students.

One of the oldest problems featuring this phenomenon was created by Ruslan Mitkov (Mitkov, 2006b: Sample Problem 5) and it was given at the 10th National Competition on Mathematical and Computational Linguistics (Bulgaria 1993). It illustrates a situation where all noun phrases (NPs) preceding the pronoun are regarded as potential antecedents, and the students’ task is to propose filters, which will allow the computer to define the correct antecedents. At the end of the problem, the author provides an additional point of view, asking the students whether this is possible only when the pronoun follows the NP (thus pointing towards the existence of cataphora without explicitly defining the notion).

The problem “Help my camera!”, created by Emily Bender (NACLO 2009, OzCLO 2009)¹⁶ and adapted for Bulgarian students by Ivan Derzhanski (Derzhanski and Velinov, 2012), presents three dialogues between a human and a computer, as the examples below:

Human: *There's this restaurant on Bancroft that's supposed to be really good that I heard about from my mother. Can you help me find it?*

Computer: *Where did you last see your mother?*

¹⁰ Available online at: <http://www.nacloweb.org/resources/problems/2013/N2013-F.pdf>

¹¹ Available online at: <http://www.nacloweb.org/resources/problems/2016/N2016-H.pdf>

¹² Available online at: <http://www.nacloweb.org/resources/problems/2016/N2016-P.pdf>

¹³ Available online at: <http://nacloweb.org/resources/problems/2017/N2017-R.pdf>

¹⁴ Available online at: <http://nacloweb.org/resources/problems/2015/N2015-M.pdf>

¹⁵ Available online at: <http://nacloweb.org/resources/problems/2014/N2014-J.pdf>

¹⁶ Available online at: <http://www.nacloweb.org/resources/problems/2009/N2009-F.pdf>

Human: *My parents are having a hard time choosing a new digital camera. How can I help them?*
Computer: *I'm afraid I don't understand your question. People don't help cameras.*

The computer misunderstands the human because it is misinterpreting certain words or phrases. The students' tasks are to explain why the computer cannot understand the human properly, and to propose better rules for the computer to interpret the anaphoric words and phrases. The problem shows the difference between human and machine way of understanding such relations.

The "Common Sense" problem by Jordan Ho (NACLO 2017)¹⁷ again provides a classic example of how anaphora relates to its antecedent. Students are given examples of anaphoric references, and their task is to relate the pronouns to the correct antecedent (all antecedents are words invented by the author).

2.3.2. Word sense disambiguation and word sense representation

As mentioned above, one of the main advantages of using linguistic problems for introducing CL and NLP tasks to high-school students is that these problems are often quite intriguing and entertaining. Sometimes, this allows for rather complex theoretical issues to be introduced to the students in a simple and enjoyable way. For example, Emily Bender's problem "The Old Man the Boats" (NACLO 2015)¹⁸ presents syntactic ambiguities with a sense of humour. The problem reviews a number of sentences (also known as garden-path sentences) each containing a local ambiguity to be solved, as in the examples:

The old train the young.
I convinced her children to do their homework.
The man who whistles tunes pianos.

After explaining concisely the nature of POS ambiguities resulting in different sentence structures, it requires the students to parse the sentences, to define their local ambiguity point and to provide a new ending after that point so that the other reading of the ambiguous word surfaces.

The problem "Kings, Queens and Counts" by Tom McCoy (NACLO 2016) introduces a method of automatically representing word meaning (as shaded graphs) based on the count of its collocations. The students are given a number of diagrams with the most common collocations defining a word based on a sample text, and the meaning of the word itself. Firstly, the students are asked to shade the graph representation of another word from the same sample text. Then their task is complicated – to match 11 mystery words to their definitions using only the information from the representations of 33 words (incl. the mystery words), obtained from a different sample text.

2.3.3. Word categorization

There are a lot of problems which ask students to categorize (unknown) words based on the context. Created by Dragomir Radev, Christiane Fellbaum and Jonathan May, the problem called "Zoink!" (NACLO 2015, UKLO 2015, AILO 2015)¹⁹ engages the students in helping the administrators of the website Zoink! to determine whether their reviews are written by bots or real people. To do that the administrators must write filtering software categorizing words into three groups based on a corpus (set of snippets from reviews written by humans). The students' task is to categorize the new data (another set of snippets) into the same three groups: written by human (correct), written by bots (wrong), and undefined (maybe). The problem explores the linguistic phenomenon of the specific structure of arranging multiple adjectives that describe different degrees of a quality (the correct 'good but not great' vs. the ungrammatical 'furious but not angry', and the undefined 'furious but not good').

"Gelda's House of Gelbergarg" by Patrick Littell (NACLO 2010, UKLO 2010, OzCLO 2010)²⁰, presents a similar model of categorizing unknown words into a) individual, discrete food items; b) liquids, undifferentiated masses or masses of uncountably small things; and c) containers or

¹⁷ Available online at: <http://www.nacloweb.org/resources/problems/2017/N2017-O.pdf>

¹⁸ Available online at: <http://nacloweb.org/resources/problems/2015/N2015-P.pdf>

¹⁹ Available online at: <http://www.nacloweb.org/resources/problems/2015/N2015-G.pdf>

²⁰ Available online at: <http://nacloweb.org/resources/problems/2010/A.pdf>

measurements. Again, decisions must be made based on context, presented in the form of customers' reviews, as in the examples below:

A hidden gem in Lower Uptown! Get the färsel-försel with gorse-weebel and you'll have a happy stomach for a week. And top it off with a flebba of sweet-bolger while you're at it!

The portions at this place are just too big! I'd rather have half the portions at a lower price – they just bring out too many göngerplose and too much meembel for me.

2.3.4. Summarization

Automatic summarization aims to create an abstract with the major points of the original document. In a problem given at the 11th National Competition on Mathematical and Computational Linguistics (Bulgaria 1994), Mitkov (2006b, Sample problem 6) displays a case where a computer program must summarize a given document without understanding it, using only a set of predefined “selection” and “rejection” rules. The students' task is to propose three rules of each kind. The rules should not include any morphological, syntactic, semantic or pragmatic analysis.

Another problem involving automated summarization is “Summer Eyes” by Dragomir Radev and Adam Hesterberg (NACLO 2009)²¹. The students are presented with the inputs and outputs of an extractive summarizer and the scores assigned by the summarizer for each sentence according to some criteria which mark it as a good summary sentence. The students are then asked to guess the criteria and rescore the sentences after a change in the story. The criteria which the students should discover include the primacy or recency of a sentence, the presence of named entities and words from the title, and choice between past- vs. present-tense verbs.

2.4. Machine translation

Even if originating in the middle of the 20th century, the idea of machine translation as the process of translation of one natural language to another, using computational software, still captivates our efforts as researchers and presents many challenges. Therefore, there is a variety of concerns to be addressed in the field. The linguistic problems related to MT evolve with each new approach. Some of the earliest problems regarded rule-based MT (Mitkov 1989: 71-75), while the ones that are more recent relate to different aspects of statistical MT. For example, “Running on MT” by Harold Sommers (NACLO 2011, UKLO 2011)²² points out the problem of word sense disambiguation for the purposes of machine translation. To simulate the effect of automatically selecting a word sense, a number of individual words from an ordinary English text were replaced with alternative words which share a meaning with the original word, but which were not correct in this context, as presented in the sample below:

Annie Jones sat ^{cross} ~~angry~~-legged on her Uncle John's facade porch; her favorite rag doll clutched under one supply. The deceased afternoon sun polished through the departs of the giant oak tree, casting its flickering ignite on the cabin. This entranced the child and she sat with her confront changed upward, as if hypnotized. A stabilize hum of conversation flowed from inside of the cabin.

As MT could be based on a number of different methods, applying a variety of approaches, and include numerous subtasks, problems may vary a lot. Future linguistic problems might as well include various notions of Machine Learning and analysis of simple inputs and outputs of Neural Networks, for example, as in fact does one of the latest NACLO problems “Nothing But Net(works)” created by Tom McCoy (NACLO 2017)²³.

Despite the impressive advance of new technologies, however, we believe that CL problems should remain self-sufficient in nature and technologically independent to reach out to students regardless of their prior knowledge and status.

²¹ Available online at: <http://www.nacloweb.org/resources/problems/2009/N2009-E.pdf>

²² Available online at: <http://www.naclo.cs.cmu.edu/problems2011/A.pdf>

²³ Available online at: <http://www.nacloweb.org/resources/problems/2017/N2017-H.pdf>

3. Conclusions

Outlining different problems designed for the needs of linguistic contests, we attempted to show a possible enthralling, yet effective way of introducing CL concepts and NLP tasks to high school students. Besides the examples presented in the paper, there are other NLP tasks which may (and in fact are) presented in a problem: spelling correction, optical character recognition and handwriting recognition, expansion of abbreviations, named entity classification, sentence boundary identification, etc.

A detailed statistical survey of the problems by year and type would not be very informative as the nature of the problems presupposes their relative uniqueness. Thus, a specific CL topic could be used just once unless a completely new scenario is suggested using the same underlying CL task or method.

Then a broader spectrum overview reveals the general tendencies – the more used a method is for solving different CL or NLP tasks, the more likely it is to appear in a new problem; and the other way round – the variety of methods employed to complete a specific CL or NLP task or application correlates directly to the variety of scenarios that can be created.

We believe that throwing some light on the issue will facilitate the involvement of more academics and researchers in this undertaking and will encourage their interest in creating new problems exploring recent methods used in CL and NLP.

References

- All Ireland Linguistics Olympiad (AILO). <https://ailo.adaptcentre.ie/>
- Australian Computational and Linguistics Olympiad (OzCLO). <http://ozclo.org.au/>
- Derzhanski, I. Velinov, A. (2010). *Lingvistichna mozaika*. Prosveta Publishing House.
- Derzhanski, I. Velinov, A. (2012). *Lingvistichen kaleidoskop*. Prosveta Publishing House.
- Mitkov, R. (1989). *V pomosht na izvanklasnata rabota po matematicheska i kompyutyurna lingvistika*. Natsionalen tsentar za uchenichesko i nauchno tvorchestvo. Sofia.
- Mitkov, R. (2006a). *Brief Outline of the School Activities in Mathematical and Computational Linguistics in Bulgaria in the 1980s and 1990s*. <http://pers-www.wlv.ac.uk/~le1825/ena/outline.pdf>
- Mitkov, R. (2006b) *Sample problems offered at Mathematical and Computational Linguistics Competitions in Bulgaria*. <http://pers-www.wlv.ac.uk/~le1825/ena/sample.pdf>
- North American Computational Linguistics Olympiad (NACLO). <http://nacloweb.org/>
- Radev, D. (2013a). *Puzzles in Logic, Languages and Computation: The Red Book (Recreational Linguistics 1)*. Springer.
- Radev, D. (2013b). *Puzzles in Logic, Languages and Computation: The Green Book (Recreational Linguistics 2)*. Springer.
- United Kingdom Linguistics Olympiad (UKLO). <http://www.uklo.org/>

Linguistic Problems on Number Names

Ivan Derzhanski

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

Milena Veneva

Independent Researcher
milena.p.veneva@gmail.com

Abstract

This paper presents a contrastive investigation of linguistic problems based on number names in different languages and intended for secondary-school students. We examine the eight problems of this type that have been assigned at the International Linguistics Olympiad throughout the years and compare the phenomena in the number systems featured there with those of the working languages of the Olympiad and other languages known to be familiar to the participants. On the basis of a statistical analysis of the results achieved by the contestants we draw conclusions regarding the ways in which the difficulty of a problem depends on its structure and the kinds of linguistic phenomena featured in it.

1. Introduction

Self-sufficient problems on number names have a long past and a tangible presence at linguistic olympiads for secondary-school students. Of the 75 problems that have been assigned at the individual contest of the first 15 instalments of the International Linguistic Olympiad (IOL) (www.ioling.org/), 8 (11%) are on number names. So are 9 (5%) problems out of 168 at the 11 instalments of the North American Computational Linguistics Olympiad (www.nacloweb.org/) and 3 (15%) out of the 20 problems in (Derzhanski, 2009).

Despite fitting comfortably into the general scheme, these problems are often seen as a maverick category, pertaining to a separate area of linguistics if not to another field of science altogether ('We did not know that linguistics could be mathematical'—from the confession of a team who had had little success with one such problem, and were attributing that to their pre-installed perception of linguistics as a branch of the humanities, unrelated to the exact sciences). And they have a reputation for being fiendishly difficult.

There is some (though not much) truth to the latter. These are the problems on number names that have been used at IOL to date and the languages featured in each, complete with their ISO 639-3 codes, families and countries where spoken:

1. IOL1.#2 (Ivan Derzhanski): Egyptian Arabic (arz: Afro-Asiatic, Egypt);
2. IOL3.#3 (Ivan Derzhanski): Mansi (mns: Uralic, Russian Federation);
3. IOL5.#4 (Ivan Derzhanski): Ndom (nqm: Trans-New Guinea, Indonesia);
4. IOL7.#1 (Evgenia Korovina and Ivan Derzhanski): Sulka (sua: isolate, Papua New Guinea);
5. IOL8.#2 (Ksenia Gilyarova): Drehu (dhv: Austronesian, New Caledonia);
6. IOL10.#2 (Ksenia Gilyarova): Umbu-Ungu (ubu: Trans-New Guinea, Papua New Guinea);

Keywords: number names, numerals, typology, linguistic problems

7. IOL13.#1 (Milena Veneva): Arammba (stk: South-Central Papuan, Papua New Guinea) and Classical Nahuatl (nci: Uto-Aztecan, Aztec Empire);
8. IOL15.#1 (Milena Veneva): Birom (bom: Atlantic-Congo, Nigeria).

Table 1 presents the average scores for the instalments of IOL where these problems (in **boldface**) appeared, together with their rank within each set (from hardest to easiest *a posteriori*, that is, 1 labels the lowest and 5 the highest average score).¹ One can see that the problem on number

No.	IOL1	IOL3	IOL5	IOL7	IOL8	IOL10	IOL13	IOL15
# 1	14.85: 4	12.91: 5	11.80: 3	14.77 : 5	15.49: 5	6.41: 2	3.43 : 1	7.66 : 3
# 2	6.88 : 1	11.98: 2	14.17: 4	11.29: 4	7.38 : 1	7.69 : 3	5.78: 4	1.68: 1
# 3	11.56: 2	10.66 : 4	3.43: 1	4.38: 2	14.29: 4	6.29: 1	5.51: 3	10.47: 4
# 4	15.24: 5	11.56: 3	3.80 : 2	1.33: 1	9.55: 3	8.92: 4	10.43: 5	11.22: 5
# 5	14.06: 3	4.84: 1	14.62: 5	9.28: 3	9.43: 2	9.60: 5	3.57: 2	7.35: 2

Table 1: IOLs with problems on number names: the average scores for all problems in the sets.

names has turned out to be the hardest one in its set three times out of eight, and the easiest only once.

The histograms in Figures 1, 2, 3, and 4 show the distribution of the points for each of the problems in question. One notices the many occasions when almost half of the participants scored zero. Problem #1 of IOL13 stands out as having been the hardest IOL problem on number names (its average score of 3.43 is the lowest one ever), and at the same time as the only problem of this type for which no solver got full score (20 points).

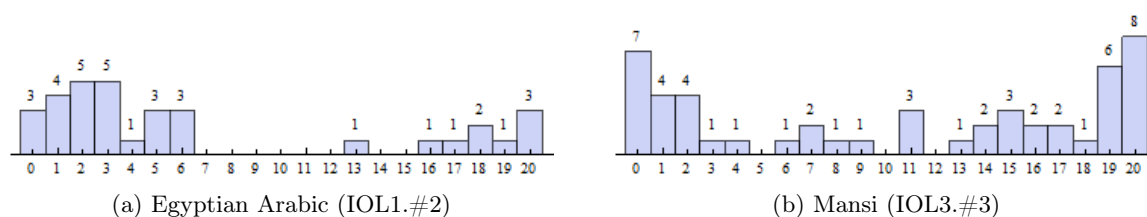


Figure 1: The distribution of scores for the first and the second problem.

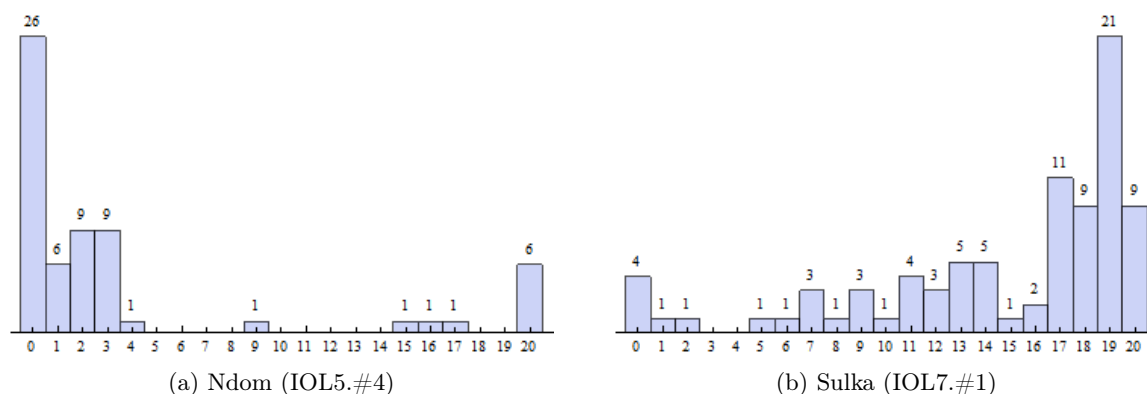


Figure 2: The distribution of scores for the third and the fourth problem.

¹At IOL every problem carries a maximal score of 20 points; the only exceptions were at IOL1 (2003), where Problems #2 and #3 were worth 25 and 15 points respectively, but in this paper the scores have been normalised in the all-20 system for ease of comparison.

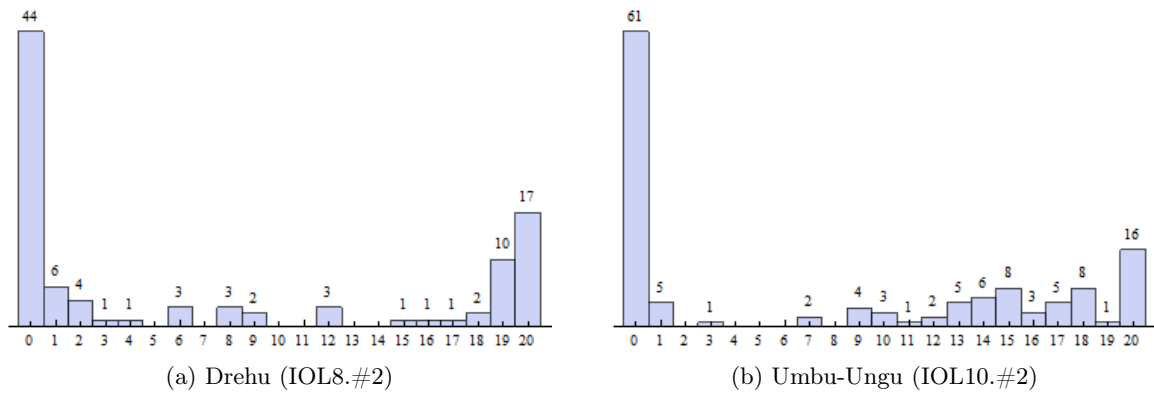


Figure 3: The distribution of scores for the fifth and the sixth problem.

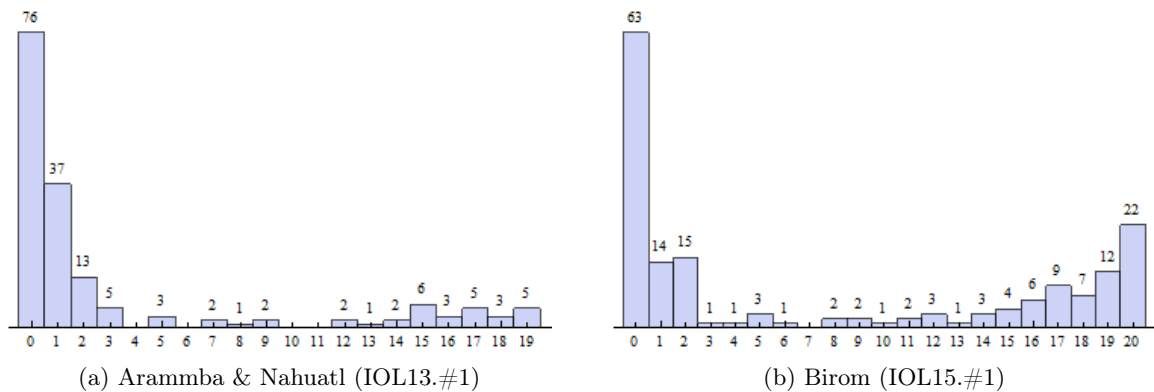


Figure 4: The distribution of scores for the seventh and the eighth problem.

The former claim, that about the pronouncedly mathematical character of problems on number names, is harder to defend. Although conspicuously absent in some languages,² number names are an integral part of language as numbers are of reality. It is true that languages have developed their systems of number names to different extents (depending on their speakers' need to count, which in turn depends on historical and social circumstances) and in different ways (Comrie, 2013), but the same can be said of any other domain of semantics. Arguably problems on number names are not significantly different from other Rosetta Stone or Chaos and Order³ problems which illustrate the crosslinguistic variety in the verbalisation of concepts (Derzhanski, 2007), (Bozhanov and Derzhanski, 2013).

In fact, thanks to the universality and the discreteness of numbers, more uniformity can be expected in number names than in many other areas of meaning. Let us see what happens.

2. Numeral Systems

In the most basic scenario languages build number names in accordance with the polynomial formula

$$N = k_n B^n + k_{n-1} B^{n-1} + \dots + k_1 B^1 + k_0 B^0,$$

²The most notorious (though controversial) case of a language with no numerals at all is Pirahã, spoken in Brazil. Some other (mostly Australian) languages seem not to get beyond 'one' and 'two', cf. Warlpiri *jinta* 'one', *jirrama* 'two', *panu* 'many (= three or more); all' (Bittner and Hale, 1995).

³The terms 'Rosetta Stone' and 'Chaos and Order' for problems in which the glosses of the words, phrases or sentences of the unfamiliar language are presented respectively in order or out of order were introduced by Ivan Derzhanski in 2004 and gained currency within IOL's Problem Committee.

where B is the base of the number system (which is 10 most of the time⁴) and $k_0, k_1, \dots, k_{n-1}, k_n$ are coefficients ($0 \leq k_i < B$). That is, a language tends to have underived names for the numbers from 1 to the base, and to express the other numbers through:

- **addition** of smaller numbers to the base (and its multiples), as in Turkish *on bir* 11 (lit. ‘ten [plus] one’); zero terms are usually omitted, though cf. Chinese *bā qiān líng yī* 8001 (lit. ‘eight thousand zero one’).
- **multiplication** of the base (and its powers) by smaller numbers, as in Bulgarian *pet-deset* 50 (lit. ‘five [times] ten’); 1 as a coefficient may or may not be explicit, cf. English *ten* but *one hundred*;
- **exponentiation**, which typically means having new underived words for the square, cube, etc. of the base, such as Italian *cento* 100 = 10^2 , *mille* 1000 = 10^3 .

The many number name systems based on the number 10 and on these three arithmetic operations vary in the details:

- the operations may be expressed by function words, inflexion or simply by juxtaposition, cf. Bulgarian *dvadeset i edno* 21 (lit. ‘20 and 1’), Hungarian *huszon-egy* 21 (lit. ‘on 20 1’, *huszon* being a modified superessive case form of *húsz* 20), and English *twenty-one*; not infrequently the sole indication of the operation is the order of its arguments, cf. Chinese *shí-sān* 13 (lit. ‘10 3’), *sān-shí* 30 (lit. ‘3 10’), *sān-shí-sān* 33;
- the grammatical expression of the operations may be motivated by the language’s gender, number or case system, cf. Czech *sto* 100, *dvě stě* 200, *tři sta* 300, *pět set* 500 with the round number *sto* ‘hundred’ in different forms (respectively singular, [obsolete] dual, paucal = nominative plural, and partitive = genitive plural) as required by the coefficient, or Russian *dve tysjači* 2000 but *dva miliona* 2000000, respectively with a feminine and a masculine form of the coefficient;
- the order of the factors (coefficient and power of the base) in the terms is language-specific, cf. Hawai’ian *kana-kolu* 30, *kana-hā* 40;
- so is the order of the terms in the sum, cf. Malagasy *iraika amby roapolo sy telonjato* 321 (lit. 1 over 20 and 300), German *dreihundert-eins-und-zwanzig* 321 (lit. 300 1 and 20), English *three hundred and twenty-one*;
- there may be a lesser or greater amount of (morpho)phonological change, syncretism or suppletion, cf.
 - Bulgarian *šest-deset* 60, transparently composed of *šest* 6 and *deset* 10,
 - Irish *aon déag* 11 but *dó dhéag* 12 with intervocalic lenition that is very characteristic of the language,
 - Colloquial Bulgarian *šejset* 60 (as above, but somewhat opaque because of the contraction),
 - Hindi *bāwān* 52 with no discernible relation to either *do* 2 or *pācās* 50, although historically it is compositional: *pācās* and *bāwān* go back to Sanskrit *pañcāśat* and *dvā-pañcāśat*, respectively (Berger, 1992: 272);
 - Turkish *kırk* 40, altogether unrelated to *dört* 4 and *on* 10;

⁴Of the 196 languages surveyed in (Comrie, 2013), 125 have a decimal number system, and these include at least nine of the world’s ten most spoken languages (French being the best-known one that breaks the pattern, though in a very limited way); decimal arithmetic also underlies the expression of numbers by Arabic numerals that are used worldwide.

Quite often a language switches to borrowed numerals beyond a certain threshold, which results in what looks like large-scale suppletion, e. g., Japanese *hitotsu* 1, *futatsu* 2, obsolete *hatachi* 20 (native) but *jū-ichi* 11, *jū-ni* 12, *ni-jū* 20 (Sino-Japanese).

Further, a language may use other operations to derive number names:

- **subtraction** (used for constructing numbers which are just a little smaller than the base or its multiples), as in Latin for numbers whose units are 8 or 9, e.g., *duo-de-viginti* 18 (lit. ‘2 [missing] from 20’); Hindi for numbers whose units are 9 (except for 89 and 99), e.g., *ūn-tīs* 29, *tīs* 30; Finnish and Estonian *kahdeksan/kaheksan* 8 and *yhdeksän/üheksa* 9 are transparently related to *kaksi/kaks* 2 and *yksi/üks* 1,⁵ though it is debatable whether the second half was originally a negative verb (Suihkonen, 2001) or a word for 10 borrowed from Iranian (Rätsep, 2003: 16);
- **overcounting** (meaning how many units are taken from the next multiple of the base), as in Finnish where the teens are constructed by adding *toista* (the partitive form of *toinen* ‘second’),⁶ or in Old Turkic, where the numbers in the range 11–89 are expressed as a number of ones from the next decade, e.g., *tört otuz* 24 (lit. 4 [from] 30);
- multiplication by **fractional coefficients**, specifically, by one half (Welsh *hanner cant* 50, lit. $\frac{1}{2}$ 100).

A significant minority of the world’s languages use bases other than 10, although very few have had the historical opportunity to construct a full-scale system (that is, one that gets at least to the square of the base) using a single non-10 base. Far more common is the situation where two or more numbers share the duties of the base in different parts of the system. In particular, languages with base-20 systems regularly form the names for 11–19 by adding 1–9 to 10, and often multiply 20 only by 2–4, whereafter a underived word for 100 makes an appearance, as a result of contact with languages with base-10 systems (this is how Basque and Georgian work).

Some languages use different number systems for counting different items, although this tends to mean that there are groups of different sizes used for counting (as in English 10 may be called *five brace* when referring to game birds).

Finally, authors of linguistic problems often use various technical complications in order to make the problem harder or more interesting, such as composing it as a Chaos and Order rather than a Rosetta Stone, including arithmetic equalities with gaps instead of just number names and their corresponding numerical values, or using material from more than one unfamiliar language.

3. The Problems

Let us now go back to the eight problems on number names assigned at IOL to this day. Table 2 summarises the principal features of the languages’ number systems and the problems, that is, the answers to the following questions:

1. Is the base of the number system 10, or 20, or 20 with supplementary bases (5, perhaps 10 and then perhaps 15), or something else?
2. Does the base have alternative (suppletive) names?
3. Are there any other numbers that play a base-like part in the number system?
4. Does the language use subtraction, or better, do the numbers just below the base behave – or are they formed – in an unusual way?

⁵Whether their present-day speakers are aware of this is another matter, and whether it helps them detect the same phenomenon in another language is a third; the results of the marking of the only problem at IOL where a similar thing happens (IOL15.#1 on Birom) do not suggest that the Estonian contestants had an advantage.

⁶In older Finnish the same system worked with larger decades as well, but now this usage is considered archaic beyond 20.

5. Does the language use overcounting?
6. Are all arithmetic operations marked, or only some of them, or none?
7. What is the word order within the polynom (+) and within every summand (×)?
8. Are there any (morpho)phonological changes in the derivation of number names?
9. What other peculiarities of the language or the number system are there that make it harder to crack, ignoring more or less straightforward morphophonological processes?
10. Does the problem present the numeric values of the expressions in an unordered list?
11. Does the problem present equalities or equations in addition to, or instead of, just numbers?
12. What other peculiarities of the problem are there that make it harder to solve?

Also the table restates the average score and the ranking of each problem within its set (where 1 is hardest and 5 is easiest). These two values jointly motivate the ordering of the columns of the table.

problem	13.#1		1.#2	5.#4	8.#2	15.#1	10.#2	3.#3	7.#1		
language	stk	nci	arz	nqm	dhv	bom	ubu	mns	sua		
1: base	other	20+	10	other	20+	other	other	10	20	20	20+
2: other names	yes	yes	no	no	yes	yes	no	no	no		
3: other bases	no	no	no	yes	no	no	yes	no	yes	yes	no
4: subtraction	no	no	no	no	no	yes	no	yes	no		
5: overcounting	no	no	no	no	no	no	yes	yes	no		
6: operations	no	+	×1, ×2	+, ×2	+	+, ×	no	no	+, ×2		
7: word order											
(a): +	↘	↘	—	↘	↗, ↘ (1)	↘	↘	↘	↗		
(b): ×	↗	↗	—	↘	↗	↘	↘	↘	↘		
8: phonology	no	yes	no	no	yes	yes	no	no	no		
9: other	—	—	(2)	(3)	(4)	—	(5)	—	(6)		
10: disorder	no		no	yes	yes	no	no	no	no		
11: equalities	yes		yes	yes	yes	yes	no	no	no		
12: other	(7)		(8)	—	—	—	—	—	—		
score	3.43		6.88	3.80	7.38	7.66	7.69	10.66	14.77		
difficulty within the set	1		1	2	1	3	3	4	5		

Table 2: Linguistic phenomena in the IOL problems on number names.

Notes to Table 2:

- (1) In Drehu the compound numbers in the intervals 6–9, 11–14, 16–19 are constructed as a sum of 5, 10, 15 and the remaining augend α , that augend coming first, while the compound numbers past 20 are formed as Γ *nge* Δ , where Γ is a multiple of 20, $1 \leq \Delta \leq 19$.
- (2) As a typical Semitic language, Egyptian Arabic has a templatic (non-concatenative) morphology, which many contestants found impenetrable. Also it expresses multiplication by 2 through a dual number form (*tumn* $\frac{1}{8}$, *tumn-ēn* $\frac{2}{8}$, *talat-t itmān* $\frac{3}{8}$).
- (3) Ndom operates a base-6 system, but 18 has a special name, which is then used to add smaller numbers to, so that 25 is not **mer an thonith abo sas* $6 \times 4 + 1$ but *tondor abo mer abo sas* $18 + 6 + 1$.
- (4) Drehu expresses the numbers 5, 10 and 15 by one set of morphemes when units are added to them and in an entirely different way when that is not the case, e.g., *caa-pi* $5 = 1 \times 5$, *caa-ngömen* $6 = 1 + 5$, *lue-pi* $10 = 2 \times 5$, *lua-ko* $12 = 2 + 10$.

- (5) Umbu-Ungu has special (unanalysable) names for all multiples of 4 (the secondary base beside 24, the primary one) up to 32, so that 56 is *tokapu polangipu* $24 + 32$ and 57 is *tokapu talu rurepo-nga telu* $24 \times 2 + 12 - 1$ (here ‘ $-$ ’ stands for overcounting; $12 - 1$ is 9 because it means ‘1 from the 4 that ends at 12’).
- (6) The Sulka number systems comes in three varieties (for counting coconuts, breadfruit and everything else), which are all featured in this problem. Some of the nouns have suppletive singular and plural forms (e. g., sg. *tu*, pl. *sngu* ‘yam’). There is also a dual number, although in this language it does not preclude the use of a numeral (*a tu a tgiang* ‘1 yam’, *a lo tu a lomin* ‘2 yams’, *o sngu a korlotge* ‘3 yams’; *lo* is the dual marker).
- (7) Besides featuring two completely unrelated number systems from different languages in such a way that the two have to be untangled in parallel, the problem also employs bigger numbers than are usually dealt with in linguistic problems.
- (8) The numbers in the problem are actually vulgar fractions.

Ordinal logistic regression (SPSS Inc., 2013) was conducted with the help of IBM SPSS Software (IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.). This method is used to predict an ordinal dependent variable (response) given a set of independent variables (predictors), which can be factors (categorical predictors) or covariates (continuous predictors). The design of the ordinal regression in SPSS is based on (McCullagh, 1980). The independent variables which have a statistically significant effect on the dependent variable were determined ($p < 0.001$). The following observations and conclusions can be made on the basis of the results:

1. Surprisingly, the extremes of the ranking are the two problems in which there is more than one number system, but Arammba–Nahuatl with its two unrelated languages is hardest and Sulka with its three counting methods is easiest.
2. The base of the number system (10, 20, or other) has no direct bearing on the difficulty of the problem.
3. The existence of an alternative name of the base makes a problem rather more difficult. The same is true for the phonological changes.
4. Neither subtraction nor overcounting make a problem difficult. Nor does the existence of an auxiliary base.
5. However, the explicit marking of the arithmetic operations does increase the difficulty.
6. A problem with different word orders in the polynom and in every summand is difficult.
7. The Chaos and Order format makes a problem harder, but not so much as large integers or vulgar fractions.
8. All problems that involve equalities prove harder than all problems that do not.

4. Conclusions

Our research, based on the results of IOL 1–15, has revealed some ways (mostly unexpected ones) in which the difficulty of a problem on number names depends on its structure and the kinds of linguistic phenomena featured in it. It would be interesting to conduct similar studies on the results of other linguistic olympiads and contests which are old long enough to have accumulated a statistically useful pool of problems on number names, and to compare the findings, which may shed further light on the effect of working languages.

References

- Berger, H. (1992). Modern Indo-Aryan. In Gvozdanović, J., Ed., *Indo-European Numerals*, pages 243–287. Mouton de Gruyter, Berlin, New York.
- Bittner, M. and Hale, K. (1995). Remarks on Definiteness in Warlpiri. In Bach, E., Jelinek, E., Kratzer, A., and Partee, B. B. H., Eds., *Quantification in Natural Languages*, pages 81–105. Springer Netherlands, Dordrecht.
- Bozhanov, B. and Derzhanski, I. (2013). Rosetta Stone Linguistic Problems. *Proceedings of the Fourth Workshop on Teaching Natural Language Processing*, pages 1–8.
- Comrie, B. (2013). Numeral Bases. In Dryer, M. S. and Haspelmath, M., Eds., *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. <http://wals.info/chapter/131>.
- Derzhanski, I. (2007). Mathematics in Linguistic Problems. In Dimitrova, L. and Pavlov, R., Eds., *Mathematical and Computational Linguistics. Jubilee International Conference, 6 July 2007, Sofia*, pages 49–52.
- Derzhanski, I. (2009). *Linguistic Magic and Mystery*. Union of Bulgarian Mathematicians, Sofia.
- McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 109–142.
- Rätsep, H. (2003). Arvsõnade päritolust eesti keeles. *Oma Keel*, 2:11–18.
- SPSS Inc. (2013). *IBM SPSS Statistics V22.0.0 Documentation*. IBM SPSS Statistics Base 22. SPSS Inc., Chicago IL.
- Suihkonen, P. (2001). Suomen ja sen sukukielten lukusanoista. *Matematiikkalehti Solmu*, 2.

Parallel Web Display of Transcribed Spoken Bulgarian with Its Normalised Version and an Indexed List of Lemmas

Marina Dzhonova
Sofia University “Sv.
Kliment Ohridsky”
Faculty of Slavic Studies
mdjonova@gmail.com

Kjetil Rå Hauge
Oslo University
ILOS
k.r.hauge@ilos.uio.no

Yovka Tisheva
Sofia University “Sv.
Kliment Ohridsky”
Faculty of Slavic Studies
yovka.tisheva@abv.bg

Abstract

We present and discuss problems in creating a lemmatised index to transcriptions of Bulgarian speech, including the prerequisites for such an index, and why we consider an index preferable to a search engine for this particular kind of text.

1. Introduction

This article focuses on the possibilities for automatic tagging of corpus of oral communication in the modern Bulgarian language. What distinguishes the object of our article from the more well-known corpora of Bulgarian language is the nature of the texts included in it. This corpus is not composed of written texts, but includes data representative of oral communication in different communicative situations with the participation of speakers of Bulgarian of varying status. The texts in the corpus are transcriptions of audio or video recordings of oral communication. In this sense, the written texts in the corpus are secondary to the original speech acts. The uniqueness of corpora of this type is related both to the specifics of the linguistic factors involved (spoken language, literary pronunciation, etc.) and to the establishment of standards and conventions for recording and transcribing oral speech.

Oral speech is one of the forms through which the modern Bulgarian language is realized. It is also its most dynamic form, where new tendencies in the language are introduced and the validity of normative criteria are contested. For an all-encompassing description and study of the modern Bulgarian language it is necessary to know both its written form and the oral variant. This understanding is the basis of the BgSpeech initiative (Tisheva and Dzhonova 2011; Tisheva and Dzhonova 2014; Tisheva 2014; Hauge and Tisheva 2014; Hauge et al. 2016), which brings together Bulgarianists and Slavists with research interests in oral communication (see bgspeech.net for participants). The creation of a corpus that is representative for the contemporary state of Bulgarian oral speech is one of the long-term tasks of the team. Resources of this type represent the called-for parallel to corpora of written (standard) texts. The creation of a corpus of oral speech complements and enriches the knowledge about the modern Bulgarian language. The inclusion of data on oral communication broadens the representativeness of linguistic research.

Compliance with the literary norm is mandatory in all cases in which the written form for realization of the Bulgarian language is chosen. In oral communication the picture is different — norms of literary pronunciation, as well as grammatical and lexical norms become “more elastic”, and their application to a great extent dispensable in different speech situations. The complex of linguistic means that are at disposal in oral speech follows the basic features of the national (official) language because it is part of it. But its phonetics and grammar do not fully follow all the specifics of the written literary language, nor do they comply with any existing dialect norm. Some of the peculiarities that are noted in the transcribed oral speech in the corpus are elisions, ellipses, abbreviated forms or phrases, overlapping utterances, incomplete utterances, repetition of constituents/phrases, colloquial constructions, pragmatic markers and discourse markers. The transcriptions also give information

about the paralinguistic means used by the speakers (pauses, gestures, mimics, phonetic paralinguistic means such as laughter, etc.), as they are an integral part of oral communication. Along with linguistic information, a mandatory condition for the corpus to be representative is to include non-linguistic information (socio-demographic features of the participants in the communication, data about the recording itself).

The special features of speech also call for a specific approach to designing this type of corpus. While the first level of annotation in corpora of written language are lemmatization and part-of-speech analysis, in corpora of spoken language part of the syntactic and pragmatic annotation is carried out as an integral part of the transcription of the recordings. This is necessary in order to determine the boundaries between the individual utterances in the organization of the transcription into a dialogical form. Simultaneous speaking, pauses, overlapping as well as non-verbal information and the communicative status of the utterances are noted by the transcriber in the initial processing of the texts. The same applies to the metadata that accompany every transcription — information about speakers and recordings is also provided by transcription. Part-of-speech and clausal annotation become the next stage for a corpus of oral speech.

Practice around the world shows that oral communication data can be collected into separate, self-contained corpora of varying volume and degree of representativeness or included in representative national corpora as a sub-corpus under the main database of written texts. The resource under consideration here is not part of a larger corpus.

The resource we present here is organised in the form of a small parallel corpus. It presents two parallel (tabular) text records of the same audio source, where one represents the result of an editing/normalizing process into the standard norm and the other the original transcriptions with the deviations from the norm indicated. This processing aims to facilitate both the extraction of data on the grammar and pragmatics of oral speech as well as the further automatic processing of the resource. Most of the texts represent unofficial colloquial speech, and in addition there are two interviews and one media text.

A search engine interface is an ideal tool for a user who wants to find out whether a certain item is a part of a given rather large set of items, for instance whether “floccinaucinihilipilification” is a word in English (it is, and it means ‘the action or habit of estimating something as worthless’ according to oxforddictionaries.com). But when the set of items is on a small scale and/or contains items that are confined to certain geographical or social entities or are scarce or of recent emergence, a gaping empty search field is of little help for the user, especially if the user is new to the field. Such sets of items are the vocabularies of dialects, professional or social jargon, neologisms, allegro forms, and colloquialisms. A better tool for such sets will be an index, that is, a list of all occurring forms with an explanation and/or an indication of the form’s place in the text.

The texts available at bgspeech.net are sets of comparably short transcriptions of spoken Bulgarian. Transcriptions of spoken language tend to contain a large number of spellings that reflect the actual pronunciation and thus differ from the standard spelling. Researchers on the hunt for data about, for instance, the use of subjunctives of cause, would search for, among other things, *защото* ‘because’, but might not consider the option of searching for an allegro form such as *умом*. This means that it could be necessary, depending on the degree of non-standard spellings in the transcripts, to produce normalised versions of them.

As described in a poster at CLIB2014, a part of the transcriptions are normalised, in the sense that in addition to the version with phonetic transcription there is a parallel version with the same texts normalised to the standard orthography. Normalisation has been effected through replacement of known pairs of semi-phonetic and standard spelling (1,358 in all), spellcheck with MacEst, developed by the Department of Computational Linguistics at the Institute for Bulgarian Language (hereafter DCL/IBL), and additional visual checking.

There are several advantages connected with transcribing spoken language in standard orthographic form. In a report on a study of a Russian dialect, the authors explain why they forego transcriptions like “*[он сво́йой жы́з’н’е это хоц’у́ погувур’ит’ / жы́с’ мо́ја прошл́а н’е о́ц’ен’ ва́жно / жы́ла ф-так’у́жо го́ды т’ежбóльйо / д’ит’ей у м’ен’а б’ыло н’ет’еро / подн’алá ја д’ит’ей до войн’ы / фторбóй сын пог’ип на войн’е]*” in favour of “*Об своей жизни это хочу поговорить. Жизнь моя прошла не очень важно, жила в такие годы тяжелые. Детей у меня было пятеро, подняла я детей до войны, второй сын погиб на войне.*” Five reasons are given:

transcription into standard language can be done quickly; it effectively solves the problem of normalization and standardization (as phonetic transcription systems used in different dialect corpora do not always coincide even for the same language); it makes the use of standard automatic annotation tools possible; it makes the data easily readable by non-linguist users; and loss of phonetic data in transcription may be made up for by aligning the transcription with the original audio, so they conclude: “All this boils down to the principle that, to make standard taggers applicable to the texts, we make as much phonetic adaptation as possible, reasonable and practicable without losing lexically, morphologically and syntactically relevant information (Waldenfels et al., 2014).

Furthermore, as the volume of transcribed texts in our case is comparatively small, an index, providing a full list of lemmas and forms, could provide a better overview of the vocabulary of the text than what one could attain by typing search terms into a search engine. In our case we have to do with 23 transcripts, varying in volume from 477 to 2,425 tokens and with a total of a little over 5,000 unique tokens, and a number of lemmas considerably smaller than that.

These transcripts are presented in a two-column view, normalised transcript to the left and the original to the right, with highlighting (red type) of the deviations that have been corrected as shown in fig. 1.

BGSpeech	
difftext	
Редактиран текст/Edited text	Оригинал/Original
М : брат ми имаше едни риби такива / които много трагично [свършиха]	М : , брат ми имаше едни риби такива / които ногу тръгичну [свършихъ]
А : [(неясно)]	А : [(неясно)]
Е : [аз пък съм чувала за рибите] на иван / какви мутанти били . за твоите риби / мутанти / дето майка ти се стряскала [като види че са живи]	Е : [аз пък съм чувъла за рибите] на иван / какви мутанти били . за твоите риби / мутанти / дето майка ти са стряскъла [като види че са живи]
А : [(неясно)]	А : [(неясно)]
И : зелено (неясно) два пръста навътре и не се [виждат]	И : , зилену (неясно) два пръста навътре и ни са [виждът]

Fig. 1: Two-column display

For the new display we are adding for each transcript a column to the left with a clickable list of lemmas and their attested wordforms, where a click on a wordform will lead to its instance in the text. Furthermore, there will be a fourth column on the right with an alphabetised list of all the corrected forms, that is, all the highlighted forms in the column with the original transcript. Each form in the list is clickable and will lead the the form in its context in the original text – see figs. 2 and 3. In addition, there will be a separate document with a full alphabetical list of all highlighted forms with links to the documents in which they occur.

Индекс/Index	Редактиран/Edited	Оригинал/Original	List of non-standard forms/pronunciation
<p>Click on any wordform after the colon to see it in its context</p> <p>: а: а а а а а а абе: абе абсолютен: абсолютно аз: аз ти те тя то те ти аз то тя то ти аз то ти ти то то аз те аквариум: аквариума аквариума аквариума аквариума ама: ама ама ама ама-ха: ха ами: ами баща: баща баща бе: бе бе бе без: без било: била бия: били блато: блато блъсна: блъсне блъсне брат: брат брат бяло-зелено-червен: зелено зелено зелено в: в в вече: вече вече вече взема: вземе</p>	<p>To return to the index click the word highlighted from the search</p> <p>М: брат ми имаше едни риби такива които много типично свършиха неясно</p> <p>А: аз пък съм чувала за рибите на иван какви мутанти били за твойте риби мутанти дето майка ти се стряскала като види че са живи</p> <p>А: неясно</p> <p>И: зелено неясно два пръста навътре и не се виждат</p> <p>М: абсолютно същите едни сомове имаше там и не знам какво те половината измряха обаче сомове живееха като пичове накрая ги бяха зарязали и ги дали на катя там да да ги гледа обаче тя си сменила квартирата накрая нямало вече жената какво да направи и ги занесла с аквариума неясно до коша</p>	<p>Words that are normalised in the column to the left are marked in red</p> <p>М: , брат ми имаше едни риби такива / които ногу тръгичну [свършихъ]</p> <p>А: [(неясно)]</p> <p>Е: [аз пък съм чувъла за рибите] на иван / какви мутанти били . за твойте риби / мутанти / дето майка ти са стряскъла [като види че са живи]</p> <p>А: [(неясно)]</p> <p>И: , зилену (неясно) два пръста навътре и ни са [виждът]</p> <p>М: [абсолютну същите] едни сомове имаше там и ни знам какво те пулвинъта измряха обаче сомувите живееха като пичове . нъкрая ги бяха зърязъли и ги дали на катъа там да - да ги гледъ / обаче тя си сменила квъртиръта . нъкрая нямълу вече жината</p>	<p>Click to see the form in context in the column to the left</p> <p>--- абсолютно аквариумчиту аквариумчиту аквариумъ ама и блату блъсни блъсни бъща вземи виждът вуда вудата вудата вудата вудата вудата вудатъ вудураслу гадну гадну гледъ глеъм гудина гулеми дода другъта другътъ дъно ената жината зилену зилену зилену зъдръж зънесла зърязъли изпъкнъли изрудил изсипвъл изтрия изхвърля имът ино инъта инъта инъче казвъм казвътъ каръл катъа квъртиръта киру киру къту къту къту лъснът мии миими минъ миришелу мириши мръснатъ напраи нещу ни ногу ногу ногу нъкрая нъкрая нъкрая нъли нъли нъли нъли нъли нъли нъля някву някву някъуй нямълу паднъл пийтъй пичуве повечи прибири</p>

Fig. 2: Index of occurring lemmas with clickable word forms and links to the normalised text

Индекс/Index	Редактиран/Edited	Оригинал/Original	List of non-standard forms/pronunciation
<p>Click on any wordform after the colon to see it in its context</p> <p>: а: а а а абсолютен: абсолютно аз: ти ние вие ние ние ние аз ние то те аз те те ти баща: баща беднотия: беднотия било: било бия: би би били би благодаря: благодаря бомба: бомби бягам: бягахме бягахме бягахме в: в в в в в в важен: важно важно важно век: век вече: вече вече вече вече вече: вече вечер: вечер вечерен: вечерно взема: вземаха виждам: вижда викам: викаха вир-вода: вода военен: военните войн: война</p>	<p>To return to the index click the word highlighted from the search</p> <p>П: А В дните около 15 септември голям интерес в мен предизвика това какво е било някога преди може би шейсет или седемдесет години и това какво представлява днешното образование Затова се срещнах с един може би а типичен представител на шопския край Какво представлява за теб и какво е тогавашното образование с какво си спомняш ти за него</p> <p>К: Д м премигва с очите поема си дъх Едно време децата ние специално ходехме с цървули кой каквото имаше това обличаше беднотия нищо А сега вие сте задоволени с много неща имате вече компютри показва с кимване към компютъра а ние едно просто радио и</p>	<p>Words that are normalised in the column to the left are marked in red</p> <p>П . А : В дните около 15 септември / голям интерес в мен предизвика това какво е било н'акога преди може би шейсет или седемдесет години и това како представл'ава днешното образование . Затова се срещнах с един . може би а / ипичен претставител на шопския край . Какво представлява за теп / и какво е тогавашното образование / с какво си спомн'аш ти за него ?</p> <p>К Д : / м / (премигва с очите , поема си дъх) . Едно време / децата / ние специ'ално ходехме с цървули / кой каквото имаше / т'ва убличаше / беднотия / нищо . А сега вие сте задоволени с много неща / имате вече кумпютри . / (показва с кимване към компютъра) . а</p>	<p>Click to see the form in context in the column to the left</p> <p>а / ипичен апсолютно б'агахме бегахме благодар'а бонби в / икаха вав вия воените възраст въоще горе-доле гудини гудини гудини даржаха даржеше децтво етака затамнявахме зем / аха зем'ата зем'ата испити испра / ти к'во каде каде каде кажем караж ку'ато кумпютри многи мойта н / >емахме н / екакси н'акога н'амаше н'амаше нана / долу напи / сали напоследака нау / чили ния ния ния ния носихмв обработва : ме освет / иха очилищтво очилище поми / ри помним помним потслони представл'ава прежи / вехме прежив'авахме претставител претставлява провал / им пулето радийо савсем савсем савсем сам</p>

Fig. 3: List of non-standard forms with clickable links to original transcription

2. Method

Our basic tool for lemmatisation is the Bulgarian morphological dictionary used in the production of the declination/conjugation patterns in Popov et al., 1998 and Popov et al., 2003, provided for us by Kiril Simov. The textual format of the dictionary was massaged into a more compact form using Applescript, and searches were made with database speed in the text editor BBEdit. An alternative would have been to use the lemmatiser provided by the Department of Computational Linguistics at the Bulgarian Academy of Sciences (<http://dcl.bas.bg/dclservices/index.php>), but lack of time for establishing a script for analysing the web pages sent in return from the lemmatiser made us go for a simpler solution. A custom-made script then traverses the normalised part of the HTML file, looking every wordform's lemma up in the morphological dictionary, producing a new document with each lemma connected with every one of its wordforms. A second script then produces HTML code for the new column from that document. The HTML and CSS coding patterns are borrowed from code by David J. Birnbaum and David Galloway at *The annotated Afanas'ev library* (<http://aal.obdurodon.org/about.php>).

3. Issues

3.1. Multiword Units

A considerable problem is posed by multiword units of the type *еду-кой/еду-чий си, кой/чий да е/било, който/чийто и да е/било*, all meaning 'whoever/whoseever'. Without special markup in the text to be lemmatised, each part of the unit will be lemmatised according to its single-form homograph. This problem has been addressed in a doctoral dissertation at IBL/BAS (Stoyanova, 2012), but there are still remaining problems — in IBL/BAS' lemmatiser, *било* in *когото и да било* is not recognised as a part of a multiword unit, and neither as a form of *съм* 'to be', but as a form of *бия* 'to beat':

```
<text>
<item><P> X <P> X</item>
<item><S> X <S> X</item>
<item>НЯМА Vs НЯМА VBIAr3s</item>
<item>ДА С ДА С</item>
<item>ГОВОРЯ Vs ГОВОРЯ VLITr1s</item>
<item>С R С R</item>
<item>КОГОТО Ps КОЙТО PROasm</item>
<item>И С И С</item>
<item>ДА Т ДА Т</item>
<item>БИЛО Vs БИЯ VLITxsno</item>
<item></S> X </S> X</item>
<item/>
</text>
```

In our case, we are slightly better off than the IBL/BAS lemmatiser, because we know exactly which texts we are going to lemmatise and can groom them to our requirements in advance, and not only that, we can also adjust the morphological dictionary, where each of these multi-word units is represented as one lemma. So we do the following, expressed in pseudocode:

```
for each lemma in morphological dictionary
  if the lemma is multiword
    for each wordform in the lemma's set of wordforms
      search for the wordform in texts to be lemmatised
      replace " " with "_" in text
      replace " " with "_" in the lemma and wordforms in morphological dictionary y
    end
  end
end
end
```

We now end up with a situation where all multiword items have been converted to singleword items, both in the morphological dictionary and in the text to be lemmatised, and all that remains is to go on with the job of lemmatising and by all means remember to convert all underscores back into spaces before we publish it.

3.2. Lemmatisation Errors

There is a definite need for disambiguation of homographs in the lemmatisation process — is *говори* a form of the verb *говоря* ‘to speak’ or of the noun *говор* ‘speech; dialect’? An educated guess for the right answer can be made by checking the immediate context of the word: if the preceding word form is an adjective in the plural (as in *западните говори* ‘the western dialects’), there is a considerable chance that the form belongs to the lemma *говор*, while if it is followed by a preposition (*говори за* ‘speaks of’, *говори с* ‘speaks with’), there is a similar chance that it is a form of the verb *говоря* (Simov et al., 2013).

Our lemmatisation was “quick-and-dirty” — we let the script accept the first hit for any given word form, expecting to do a clean-up job afterwards for cases like *прави* in *май беше където ти прави прическата* being classed as a of the adjective *прав* ‘right, correct’ instead of the verb *правя* ‘to do’; or *иска* in *каза нали че иска да е при мене* as an articulated form of the masculine noun *иск* ‘claim, action’ rather than as a form of the verb *искам* ‘to want’.

IBL/BAS’ lemmatiser, mentioned above, will do a better job with these, relating both *прави* and *иска* to their proper lemmas (although mislabelling the particle *май* as a noun, but that was a tricky one, with no left context):

```
<text>
<item><P> X <P> X</item>
<item><S> X <S> X</item>
<item>май Ns майNCMNsom</item>
<item>бешеVs съMVLIINd3s</item>
<item>където D където D</item>
<item>ти Ps аз PHi2s</item>
<item>прави Vs правя VLITe3s</item>
<item>прическата Ns прическа NCFsdf</item>
<item></S> X </S> X</item>
<item/>
</text>
```

```
<text>
<item><P> X <P> X</item>
<item><S> X <S> X</item>
<item>каза Vs кажа VLPTe2s</item>
<item>нали T нали T</item>
<item>че C че C</item>
<item>иска Vs искам VLITe3s</item>
<item>да C да C</item>
<item>еVs съMVLINr3s</item>
<item>при R приR</item>
<item>менеPs аз PHytl1s</item>
<item></S> X </S> X</item>
<item/>
</text>
```

However, the DCL lemmatiser did not excel in all cases. While our method (or lack of it) assigned the plural noun form *движения* ‘movements’ to the verb *движа* ‘to move’, the DCL lemmatiser, even with (or perhaps misled by) two plural adjectival forms in the left context, proposed the adjective *движен* ‘moved’:

```

<text>
<item><P> X <P> X</item>
<item><S> X <S> X</item>
<item>ВИЖ Vs ВИДЯ VLPTI2s</item>
<item>че С че С</item>
<item>такива Рр такъв PDAp</item>
<item>елементарни Ap елементарен Apo</item>
<item>движения Нр движенAqsmo</item>
<item></S> X </S> X</item>
</item/>
</text>

```

Both approaches failed miserably with the 1st person verb form *отчета* ‘to account for’ in *нали бях си насъбрала пари да и ги отчета*. Bypassing, or in our case, not even reaching to the same-stemmed noun *отчет* ‘account’, they suggested *отче* ‘father (in the religious sense)’:

```

<text>
<item><P> X <P> X</item>
<item><S> X <S> X</item>
<item>нали Т нали Т</item>
<item>бях Vs съмVLINe1s</item>
<item>си Р себе PFHzt</item>
<item>насъбрала Vs насъбера VLPTxsfo</item>
<item>пари Нр пара NCFpof</item>
<item>да С да С</item>
<item>иС и С</item>
<item>ги Рр аз PHza3p</item>
<item>отчета Нр отче NCNpon</item>
<item></S> X </S> X</item>
</item/>
</text>

```

4. Conclusion

The use of the method described in this test case shows what problems may arise when trying to use presently available programs for the automatic processing of Bulgarian text data. Normalisation of the word forms in the text is a necessity, as the available programs and morphological dictionaries include only data from the written language, and the remaining conversational syntactic structure may restrict the automatic annotation of the text. It is also obvious that a degree of manual assistance will be necessary in any case. The lessons learned so far will be applied to the tagging of the other speech data we have at our disposal and will hopefully facilitate user access to our data.

References

- Hauge et al. 2016: Hauge, K., Tisheva, Y., Džonova, M. (2016). BgSpeech i predstavjaneto na ustnata rech v Balgarskiya natsionalen korpus. *Problemi na ustnata komunikaciya*. T. 10. Chast 2. Veliko Tarnovo: UI „Sv. sv. Kiril i Metodij”, 175–186.
- Hauge, Tisheva, 2014: Hauge, K. Ro, Tisheva, Y. (2014). Paralelen korpus s dannii za balgarskata razgovorna rech – struktura i prilozhenie. *Ezikovi resursi i tehnologii za balgarski ezik*. Sofija: Akademichno izdatelstvo „Prof. Marin Drinov”, 142–153.
- Popov et al., 1998: Popov, D., Simov, K., Vidinska, S. (1998). *Rečnik za pravogovor, pravopis, punktuacija*. Sofija: Atlantis.
- Popov et al., 2003: Popov, D., Simov, K., Vidinska, S. (2003). *Pravopisen rečnik na bālgarskija ezik*. Sofija: Nauka i izkustvo.

- Simov et al., 2013: Simov, K., Ivanova, G., Mateva, M., Osenova, P. (2013). Integration of dependency parsers for Bulgarian. *The Twelfth Workshop on Treebanks and Linguistic Theories*, 145–156. Sofia.
- Stoyanova, 2012: Stoyanova, I. (2012). *Avtomaticchno razpoznavane i tagirane na sastavni leksikalni edinitsi v balgarskiya ezik*. Disertaciya za prisazhdane na obrazovatelna i nauchnata stepen „doktor”. Sektsiya po kompyutarna lingvistika, Institut za balgarski ezik, Balgarska akademiya na naukite.
- Tisheva, Dzhonova, 2011: Tisheva, Y., Dzhonova, M. (2011). Korpus s ustna balgarska rech – struktura i spetsifika. *Balgarski ezik* 3, 34–53.
- Tisheva, Dzhonova, 2014: Tisheva, Y., Dzhonova, M. (2014). Balgaristikata – mezhdu fishovete i multimedijnite korpusi. *Balgarski ezik, literatura i e-obuchenie*. Plovdiv: „Rakursi” OOD, 24–35.
- Tisheva, 2014: Tisheva, Y. (2014). Ezikovi bazi danni, korpusi i elektronni resursi za balgarskata ustna rech. *Littera et Lingua* 11, 1–2.
<http://slav.uni-sofia.bg/naum/lilijournal/2014/11/1-2/ytisheva>
- Waldenfels et al., 2014: Waldenfels, R. von, Daniel, M., Dobrushina, N. (2014). Why Standard Orthography? Building the Ustyia River Basin Corpus, an Online Corpus of a Russian Dialect.
<http://www.dialog-21.ru/digests/dialog2014/materials/pdf/WaldenfelsR.pdf>.

Integrating crowdsourcing in language learning

Georgi Dzhumayov
Institute for Bulgarian Language
“Prof. Lyubomir Andreychin”
Sofia, Bulgaria
dzhumayov.georgi@gmail.com

Abstract

This article aims to illustrate the use of *crowdsourcing* in an educational context. The practical part illustrates and provides the results of an online test conducted among 12th grade high school students from Bulgaria in order to gain new knowledge, find out common characteristics among the tenses and revise for their upcoming exams. They along with some interesting and inspiring teaching ideas could be used in an educational environment to provide easier, quicker and more interactive acquisition of a language. The experiment has been conducted by means of *Google forms* and sets the beginning of the establishment of an annotated corpus of right and wrong uses of the Bulgarian and English tenses too.

1. Objectives of the current research

The current survey conducted by means of *Google forms* aims to research 12th graders' awareness of English and Bulgarian tenses using *crowdsourcing*. They study Bulgarian as their mother tongue and English as a foreign language in *English Language School Plovdiv*. The main purpose is to inspire them to revise for their upcoming state exams and gain some more new knowledge while practising the tenses online by means of their smartphones.

The research also strives to illustrate the forms of all the tenses in a clear and logical way and find out the common characteristics between their uses and their suitability for various situations. Thus it would be easier for a speaker of one of the two languages in question to acquire the next one with fewer efforts. Moreover, students would be able to practise tense uses and names and also provide examples in order to revise for their exams in Bulgaria. Thus, through *crowdsourcing* with mobile devices we strive to compile an annotated corpus with tasks and examples illustrated in the corresponding theoretical part of that article.

2. Tenses – theoretical background

The following theoretical part endeavours to compare the tense systems of English and Bulgarian. It is constructed to make the reader aware of all the indicative tense forms of the languages in question and to pinpoint some of the similarities between their common uses and characteristics.

3. Bulgarian tenses ¹

The Bulgarian verb expresses an action as a process in time. Therefore, the verb form shows when the action is done. When the time of the verbal actions is grammatically stated, we discuss them from their position with respect to the moment of speaking.

¹ For more detailed description of the 9-member category of tense in Bulgarian, see Kutsarov (2007) and Nitsolova (2008).

In modern linguistics, it is assumed that Bulgarian possesses a category of tense with the following 9 members:

1. Present // Сегашно (пише)

Its forms show an action taking place at the moment of speaking, things that are always true or that happen on a regular basis.

2. Aorist // Минало свършено (писа)

Its forms show an action preceding the moment of speaking. It expresses an action that happened at a specific time in the past.

3. Imperfect // Минало несвършено (пишеше)

Its forms show an action taking place at the same time of a past orientation moment. It is mainly used to express a temporary situation that existed at or around a particular time in the past; frequently repeated or permanent past actions; and when we describe a setting while telling a story.

4. Perfect // Минало неопределено (писал е)

The forms express a result simultaneous with the moment of speaking. It expresses a past action but the precise moment when it happened is not clear. It is not known or important because what matters is the result of the action

5. Pluperfect // Минало предварително (беше писал)

Its forms show a result concurrent with a past orientation moment. It expresses an action which happened before another past action.

6. Future // Бъдеще (ще пише)

Its forms express an action taking place after the moment of speaking. It is used to show future actions.

7. Future in the Past // Бъдеще в миналото (щеше да пише)

The forms express an action taking place after a past orientation moment. It shows an action which was to be completed in the past and was future regarding another past action.

8. Future Perfect // Бъдеще предварително (ще е писал)

The forms show a result of an action after the moment of speaking. It expresses an action which is to take place in the future before another future action.

9. Future Perfect in the Past // Бъдеще предварително в миналото (щеше да е писал)

Its forms express a result after a past orientation moment. It shows a past action which is future regarding another past action which itself is prior to another action.

4. English tenses ²

English verbs show the time of the action or state that is being described. English tenses can be quite tricky for anyone who is on the point of commencing to learn the language – firstly because of their number and secondly because of their complexity in the terms of positioning and expressing actions and their relationships.

In 2015, in his works based on the tenses in English and Bulgarian in terms of machine translation, Todor Lazarov uses the following table to illustrate the English tenses using four main signs – anteriority, posteriority, perfection and duration.

	anteriority	posteriority	perfection	duration
Present simple	-	-	-	-
Present continuous	-	-	-	+
Present perfect	-	-	+	-
Present perfect continuous	-	-	+	+
Future simple	-	+	-	-
Future continuous	-	+	-	+
Future perfect	-	+	+	-
Future perfect continuous	-	+	+	+
Past simple	+	-	-	-

² For reference on English tenses, see <http://spot.colorado.edu/~michaeli/MichaelistenseHEL.pdf>

Past continuous	+	-	-	+
Past perfect	+	-	+	-
Past perfect continuous	+	-	+	+
Future simple in the past	+	+	-	-
Future continuous in the past	+	+	-	+
Future perfect in the past	+	+	+	-
Future Perfect continuous in the past	+	+	+	+

Table 1: Division of English tenses

1. Present Simple // Сегашно просто (work/s)

is used for permanent states, repeated actions, daily routines; for laws of nature, general truths, sports commentaries, reviews, timetables and directions.

2. Present Continuous // Сегашно продължително (am/is/are working)

is used for actions taking place at the moment of speaking; for actions going on around now; changing or developing situations.

3. Present Perfect // Сегашно перфектно (have/has worked)

is used for actions which started in the past and continue up to the present; for actions which have recently finished and whose result is obvious in the present; for actions which happened at an unstated time in the past.

4. Present Perfect Continuous // Сегашно перфектно продължително (have/has been working)

emphasizes on the duration of an action which started in the past and continues up to the present.

5. Past Simple // Минало просто (worked)

expresses an action which happened at a definite time in the past which is either stated or implied; for actions which happened immediately one after the other; for finished states or habits.

6. Past Continuous // Минало продължително (was/were working)

is used to express an action which was in progress at a stated past moment; two or more simultaneous past actions; describe the setting and the atmosphere when we tell a story.

7. Past Perfect // Минало перфектно (had worked)

expresses an action which happened before another past action or before a stated past time; an action which finished in the past and whose result was obvious in the past.

8. Past Perfect Continuous // Минало перфектно продължително (had been working)

puts emphasis on the duration of an action which started and finished in the past before another past action or stated moment.

9. Future Simple // Бъдеще просто (will work)

is used in predictions, on-the-spot decisions, promises, requests and threats.

10. Future Continuous // Бъдеще продължително (will be working)

expresses an action which will be in progress at a stated future time; an action which will definitely happen as a result of a routine or arrangement; for polite questions.

11. Future Perfect // Бъдеще перфектно (will have worked)

is used for an action which will be finished before a stated future time.

12. Future Perfect Continuous // Бъдеще перфектно продължително (will have been working)

emphasizes the duration of an action up to a certain time in the future.

13. Future Simple in the Past // Бъдеще просто в миналото (would work)

expresses an idea that in the past you thought something would happen in the future and in the construction of a II conditional sentence.

14. Future Continuous in the Past // Бъдеще продължително в миналото (would be working)

is used to show an idea that in the past an action was predicted or planned in a certain period in the future, regardless the fact that idea was not proved true.

15. Future Perfect in the Past // Бъдеще перфектно в миналото (would have worked)

expresses an action that would have been completed at a past time and in a III conditional sentence.

16. Future Perfect Continuous in the Past // Бъдеще перфектно продължително в миналото (would have been working)

expresses an action or an imaginary situation that would have been happening in the past but it is rarely used nowadays.

Considering the theoretical and the practical part of the current article, we can conclude that there are some useful similarities between the tense systems of the two languages in question—Bulgarian and English.

1. The Bulgarian Present // Сегашно has common features with the English Present Simple and Continuous.
2. The Bulgarian Aorist // Минало свършено is similar to the English Past Simple.
3. The Bulgarian Imperfect // Минало несвършено shares common features with the English Past Simple and Continuous.
4. The Bulgarian Perfect // Минало неопределено coincides with the uses of the English Present Perfect.
5. The Bulgarian Pluperfect // Минало предварително coincides with the uses of the English Past Perfect.
6. The Bulgarian Future // Бъдеще has common characteristics with the English Future Simple.
7. The Bulgarian Future in the Past // Бъдеще в миналото shares the characteristics of the English Future Simple in the Past.
8. The Bulgarian Future Perfect // Бъдеще предварително shares common features with the English Future Perfect.
9. The Bulgarian Future Perfect in the Past // Бъдеще предварително в миналото coincides with the uses of the English Future Perfect in the Past.

5. Definition of crowdsourcing

Various authors define *crowdsourcing* in a different way. Here we will provide some definitions of the term:

- According to the *Oxford Learner's Dictionary* the term *crowdsourcing* can be defined as “the activity of getting information or help for a project or a task from a large number of people, typically using the Internet”.
- Howe (2006) interprets *crowdsourcing* takes place any time a company makes a choice to employ the crowd to perform labour that could alternatively be performed by an assigned group of employees or contractors, even if the company is just now putting up a shingle. In other words, *crowdsourcing* need not require an active shift from current employees (or again, contractors) to the crowd; it can start with the crowd.
- *Crowdsourcing* may draw on the wisdom of the crowd, which can be smarter, more effective, and more reliable than the best individuals in that crowd (Surowiecki, 2005).

The main idea for the use of *crowdsourcing* is to make a group of people work on a task while sharing knowledge, skills and experience in order to resolve an issue or conduct an experience. In addition, a lot of people are eager to contribute to the invention of something new or to the solution of a pressing problem due to personal beliefs, social and intellectual motives. Hence, *crowdsourcing* would be the most convenient and easily accessible way of working together as this would lead to many people being satisfied from their mutual interaction and communication.

6. Crowdsourcing in education

By itself, *crowdsourcing* is unlikely to deliver the best educational experience, but it is a natural framework for learning (Weld et al, 2012). In fact, *crowdsourcing* gives educators a unique advantage while working in a new environment to develop skills which their students will definitely need in the future and will make them put what they have learned in theory into practice.

Although it is generally assumed that teachers and educators are keen to develop and use new techniques and approaches, the status-quo is not easily altered. Concerning the current educational system, *crowdsourcing* may come in handy and be applied to provide contemporary and accessible education in an online environment both for teachers and students. In addition, *crowdsourcing* techniques help us reach a significant number of people at very low costs.

Boshnakova (2015) claims that *crowdsourcing* can be used for:

1. gathering texts appropriate for a definite age group or based on a given topic;
2. discussion and preparation of a programme with the participation of practitioners and other specialists;
3. seeking out ideas for the development of educational policies;
4. sharing experience by people working in the field of pedagogy;
5. creating a multicultural learning environment;
6. evaluating students' work by other students or by qualified specialists.

7. Description of the experiment

For the purpose of that article high school students (all of them 12th graders) from Plovdiv studying at *English Language School Plovdiv* have been asked to answer several questions based on the tense systems of the English and Bulgarian languages. Here we have to mention that the students' mother tongue is Bulgarian and English is their first foreign language. There are no students lagging behind in their development or students having problems with processing the required school material. The total number of respondents having taken part in the survey up to the publication of the current article is thirty-three. The questions have been piled up in order to establish the beginning of an annotated corpus. It has been constructed by means of *Google Forms* where all the students had to distinguish five Bulgarian and five English verbs and the corresponding tenses they are used in. In addition to this, the students were supposed to type sentences in order to give examples of the correct use of various English tenses. The answers generated by the students in the last mentioned section can be used for the future development of a corpus consisting of correct and inappropriate usage of the tenses in Bulgarian and English.

Apart from having fun and experimenting with new for the target group educational tools and resources, all the students managed to revise some basic knowledge for the state exams that they had to sit in May 2018.³

Feel free to fill in our questionnaire and take part in our survey. Thus we will be able to use crowdsourcing to the fullest and enrich the aforementioned corpus with more examples.

8. Analysis of the results

It is obvious from the answers given by the students that they are much more aware of the names of the English tenses. However, they have problems with differentiating some Bulgarian tenses. In the first five questions students have to find out which English tense is used in the example provided. The first sentence is 1) *By the end of next month, she will have been teaching for 20 years.* Here 93.5% of all respondents give the correct answer – Future perfect continuous. In the second example 2) *He bought that house in 1996.* the highest percentage is reached – 100% claim that Past simple is used in the sentence. In the third place we have 3) *When Monica came home, John had already prepared the dinner.* This question has been answered correctly by 87.1% of all respondents. Here the percentage is slightly lower probably because the sentence is complex and comprises two verbal forms. Question number four in the current study is 4) *They had been looking for a car for 6 months before they found*

³ The following link gives direct access to the questions answered by the respondents. (<https://docs.google.com/forms/d/1RGqh5oA4ILmnaI5VvlCJ2Mk24WUmixcZXpQgUtPHbe0/edit>).

one. In this example two tenses are used – Past perfect continuous and Past simple but 90.3% of the respondents give correct answers. Fifth comes 5) *George has had the dog for 5 years*. This is the tense which causes most problems among the students according to the current study – only 83.9% claim that Present perfect has been used in the example provided. Probably, the reason for the mistakes here is that Present perfect is used to illustrate a past action whose result is visible in the present or which action is still in progress in the present. Almost 13% of all respondents say that Past perfect is the tense used in the fifth example.

The second part of the questionnaire also consists of five tasks each of which begins with *Write down a sentence using* and a different tense. It proves to be the most difficult among the three. Here the teacher can check various areas – whether the students know the names of the tenses, whether they can use them in an appropriate context. Spelling can be checked too. Firstly, the respondents had to type an example using the present continuous. Some of the correct examples typed by the students are shown below: 1) *I'm writing a book, called 'Marcia'*. 2) *I can't talk, I am driving right now*. 3) *I am waiting for the doctor at the moment*. 4) *Tom is reading a book while his mother is cleaning the room*. 5) *I am doing my homework now*. 6) *I'm seeing my friend this afternoon*. 7) *I am playing tennis now, phone me later*. Every example listed above shows correct use of tense and awareness of its name. However, there are some students that have provided wrong answers, probably thinking Present continuous is the same as Present perfect continuous – 8) *I have been working for them for 7 years now*. 9) *I have been studying english for many years*.

Secondly, the students had to type an example using the Present perfect, the tense whose results were lowest in the first part of the questionnaire. Some of the correct examples are: 1) *I have had this notebook for seven years*. 2) *I have never been to Paris*. 3) *Peter isn't coming tomorrow because he hasn't finished his project yet*. 4) *I have had many obstacles this school year and so far I have overcome all of them*. In this section we have only one inappropriate answer 5) *I have been living in Plovdiv for 4 years*, which contradicts the task requiring the use of present perfect in the exemplary sentences.

Thirdly, the respondents had to type an example using the Past perfect. Some of the correct examples typed by the students are shown below: 1) *I had picked out my suit a long time before the ball, however it arrived only 2 days before the event*. 2) *When we arrived at the meeting, the boss had already begun the presentation*. 3) *By the time we arrived, the film had already finished*. Many students forget that the Past perfect is used to express an action which happened before another past action or before a stated past time or an action which finished in the past and whose result was obvious in the past and provide an inappropriate answer, e.g. 4) *I had been in Naples four times*.

Fourthly, the students had to type an example using the Future perfect. Some of the correct examples are: 1) *I will have finished my university education by 2030*. 2) *I will have finished the science fiction novel by the end of the year*. 3) *By the end of the day, I will have met that boy*. 4) *I will have left Bulgaria by the time he comes back*. In this section we can encounter some interesting mistakes, too. For instance in this answer 5) *I will have my diploma legalized and translated into German*, the respondent uses the Future simple with the construction *have something done*, which is used to show that somebody else does something for the speaker.

Fifthly, the respondents had to type an example using the Past continuous. Some of the correct examples given by the students are: 1) *I was playing basketball when I fell and broke my leg*. 2) *My mother wasn't sleeping when I was talking with my friend about my birthday party*. 3) *I was washing the dishes while she was cooking the turkey*. 4) *I was going to attend the ball but I changed my mind the very last minute*. 5) *I was looking at the train as it passed under my balcony in the foggy afternoon*. These sentences show that the students can recognize the name of the tense and can use it in an adequate context. One interesting mistake that we encountered while analyzing the results of the survey is 6) *I was speaking loudly when he get angry*. In this example the respondent does not use the correct tense in the second part of the sentence which means he/she is neither aware of the tense agreement rules nor the Subject-Verb agreement. Moreover, we have one student who has used the Past perfect continuous instead of the Past continuous – 7) *I had been doing the housework all day long*.

The third and last part of the online test is based on the Bulgarian tense system. Just like the task illustrated in the first English part, here again the respondents dispose of five sentences and they have to guess the name of each Bulgarian tense. The first sentence is 1) *Той е ял супа. (He has eaten soup.)*

Almost 97% of the answers are correct which means that the students are very well aware of the Bulgarian Perfect // Минало неопределено. Only 3% of all the students say that the Imperfect // Минало несвършено has been used in the exemplary sentence. The second sentence is 2) *До юли той щеше да е писал на посланика. (By July he would have written to the ambassador.)* In this example the Future Perfect in the Past // Бъдеще предварително в миналото is used and almost 84% of the respondents give a correct answer. 6.5% claim that the Future Perfect // Бъдеще предварително is used and almost 10% give their answer to the Pluperfect // Минало предварително. Sentence number three is the one that causes most problems to the students. 3) *Вчера по това време учех. (I was studying at that time yesterday.)* They hesitate between two answers the Aorist // Минало свършено and the Imperfect // Минало несвършено but luckily 48.4% vote for the correct tense. One respondent votes for the Present // Сегашно and one for Perfect // Минало неопределено. Fourth comes 4) *Ще звънна по някое време. (I will call sometime.)* Here the situation is clear and the majority of the respondents give correct answers voting for the Future // Бъдеще. Sentence number five 5) *Някой беше драскал по вратата. (Somebody had been scratching the door.)* proves tricky as well. Almost 39% of the student claim that the Imperfect // Минало несвършено is used and unfortunately, they are wrong. 58% of the answers provided go for the correct choice – the Pluperfect // Минало предварително.⁴

9. Conclusions

The practical part of the current study uses one contemporary and extremely useful method of the computational linguistics in the sphere of the Bulgarian educational system – *crowdsourcing*. The results of an online test have been analysed. The questionnaire was conducted among 12th graders from a high school in Plovdiv, Bulgaria and was based on the tense system of the English and Bulgarian languages. By means of *Google forms* we have set the beginning of a corpus which could be used by linguists, translators and foreign language experts to construct new methodologies for teaching the two languages in question.

The current research was also useful for the respondents. While taking part in the investigation, they managed to revise the tenses in their mother language– Bulgarian, as well as the tenses in English– their first foreign language which would come in handy for their state exams at the end of the school year.

Concerning our future work, we strive to combine more questionnaires consisting of tasks assessing students' linguistic knowledge. In addition, we would try to enrich our corpus with a bigger collection of right and wrong students' examples which illustrate the use of tenses.

⁴ For tabular view of the results and the corresponding answers of the conducted research visit the following link. https://docs.google.com/spreadsheets/d/1F1KGSza9VumT_pv4e6oRkPF01MgfyDcqqfnJTWOBq2c/edit#gid=1970174785

References:

- Boshnakova, D. (2015). *Obrazovanie ot talpata za talpata*. Departament "Masovi komunikatsii". New Bulgarian University.
http://ebox.nbu.bg/mascom16/view_lesson.php?id=6#_edn10
(visited on 03.01.2018)
- Howe, J (2006). *Crowdsourcing. Why the power of the crowd is driving the future of business*.
http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html
(visited on 04.02.2018)
- Kutsarov, I. (1997). *Lektsii po balgarska morfologia*. Plovdiv.
- Kutsarov, I. (2007). *Teoretichna gramatika na balgarskia ezik. Morfologia*. Univ. izd. "Paisiy Hilendarsk". Plovdiv.
- Lazarov, T. (2015). *Osobenosti na glagolnite sistemi i nachinite za izrazyavane na vremeto v balgarski i angliyski. Semantichen transfer pri prevod na glagolnite formi ot balgarski na angliyski – Littera et lingua – Elektronno spisanie za humanitaristika*, 12, 1 – 2.
(visited on 04.01.2018)
- Lenhart, A. (2015). *Teens, social media, & technology overview 2015*. Pew Research Center: Internet, Science & Tech.
<http://www.pewinternet.org/2015/04/09/teens-social-media-technology-2015/>
(visited on 16.11.2017)
- Michaelis, L. (2006). *Time and Tense. The Handbook of English Linguistics*. Oxford: Blackwell.
- Nitsolova, R. (2008). *Balgarska gramatika. Morfologia*. Sofia.
- Surowiecki, J. (2005). *The wisdom of crowds*. New York: Random House.
- Teach Thought. (2012). *10 Ideas for Using Technology to Teach Writing*.
<https://www.teachthought.com/literacy/10-ideas-for-using-technology-to-teach-writing/>
(visited on 14.11.2017)
- Tsvetanova, K. (2017). *Kak da se namali tezhesta na ranitsata na uchenika?*
<https://www.dnes.bg/obshtestvo/2017/10/27/kak-da-se-namali-tezhesta-na-ranicata-na-uchenika-vylchev-predlaga.357506>
(visited on 16.11.2017)
- Weld et al. (2012). Weld, Daniel S., Eytan Adar, Lydia Chilton, Raphael Hoffmann, Eric Horvitz, Mitchell Koch, James Landay, Christopher H. Lin, and Mausam, *Personalized Online Education – A Crowdsourcing Challenge*,
<https://homes.cs.washington.edu/~weld/papers/weld-hcomp12.pdf>
(visited on 05.01.2018)

Dictionaries:

- <https://www.merriam-webster.com/>
(visited on 10.03.2018)
- <https://www.oxfordlearnersdictionaries.com/>
(visited on 10.01.2018)
- <https://dictionary.cambridge.org/>
(visited on 10.03.2018)

Bulgarian-English Parallel Corpus for the Purposes of Creating Statistical Translation Model of the Verb Forms. General Conception, Structure, Resources and Annotation.

Todor Lazarov

Department of Computational Linguistics,

IBL-BAS

todorlazarov91@abv.bg

Abstract

This paper describes the process of creating a Bulgarian-English parallel corpus for the purposes of constructing a statistical translation model for verb forms in both languages. We briefly introduce the scientific problem behind the corpus, its main purpose, general conception, linguistic resources and annotation conception. In more details we describe the collection of language data for the purposes of creating the corpus, the preparatory processing of the gathered data, the annotation rules based on the characteristics of the gathered data and the chosen software. We discuss the current work on the training model and the future work on this linguistic resource and the aims of the scientific project.

1. Introduction and brief background on the subject

The current work on the Bulgarian- English parallel corpus for the purposes of constructing a statistical translation model for verb forms in both languages continues previous works on this subject. As it has been previously stated, translating verb forms is very difficult even for human translation – even though the verb systems of both English and Bulgarian share numerous common characteristics, they differ in the manner in which they express the relations between events and points on the temporal axis, the action denoted by the verb and the information about these events. Nevertheless, as we speak about the opportunities of machine translation, both languages are resource rich, which makes theoretical and practical researches about different aspects of them reliable and the gathered data – practical for the purposes of natural language processing and machine translation.

1.1. The difficulties of translating the verb forms from Bulgarian to English

In numerous previous papers on this subject it has been pointed that the main difficulties in the process of translating the verb forms from Bulgarian to English derive from the grammatical characteristics of these languages. The temporal systems of both languages share a common feature – they consist of different grammatical categories within the hyper-category of tense. Without discussing the grammatical peculiarities of Bulgarian and English, we will outline some of the main differences that contribute to qualitative and quantitative dissimilarities. The most tangible difference is the different number of tenses in the discussed languages: while the English tense system consists of 16 structurally dependable morphological tenses, the Bulgarian system has 9 morphological tenses and different lexical categories that can alter the meaning of the tense forms. Both Bulgarian and English have a category that expresses a completed action in relation to a referential point – the perfect tenses. An obvious difference is the presence of continuous tenses in English, which can express an action that is uncompleted related to the referential point, as opposed to Bulgarian where such tenses do not exist. Another tangible difference is that the Bulgarian verbs have lexical aspect, which is part of the semantics of the lexical unit and expresses the action as finished or unfinished related to the action`s

Keywords: verb form, corpus, annotation structure, statistical machine translation, translation model

own completion (Kucarov, 2007:551). Although many linguists would not hesitate to give a positive answer to the question about whether there is aspect in English and would point to the Progressive as example of an aspectual meaning, there are other linguists who would reject the idea of aspect in English at all. More on the differences between English and Bulgarian regarding the category of aspect can be found in Kabakciev (2000). Also word order in English is a decisive factor in distinguishing meaning when we have the same situation, the same participants, but only different position of the elements of the sentence which influences the meaning. Rendering this meaning in Bulgarian is not a problem; we choose the lexical verb and very often in Bulgarian we have to specify the type of action by adding affixes to the verb (Ivanova, 1968, Nedelcheva, 2012). These are only the main tangible differences of both languages' grammatical systems, but the main point is that the Bulgarian language has the possibility to grammaticalize different linguistic data through greater number of grammatical categories in around 2000 verb forms. The greater number of possible grammatical categories, therefore possible grammaticalized meaning, in Bulgarian contributes to high levels of ambiguity during translation, due to the fact that in English the possible grammatical categories are less and the grammaticalized information from Bulgarian as source language needs to be reduced or unevenly distributed between different grammatical categories in English as target language.

1.2. Overview of the proposed solutions and the current work on the problem

Nevertheless, as it has been pointed out before, the characteristics of grammaticalized information in Bulgarian and English verb forms share numerous similarities. While structurally the verb forms in Bulgarian and English can be studied as a specific type of grammatical collocations (Sinapova and Dochev, 1999), other studies (Vassileva, 2003) on Bulgarian and English temporal systems prove in a convincing way that the two languages are different in many respects. However, many linguists note parallels and similarities in the tense system, the categories of aspect and the temporal variations. That is why we have similar grammatical meaning in most of the verb forms that can be formally described and analysed. The current work on analysing and describing in what manner the grammatical information is transferred during translation can be divided in two main approaches, each based on two fundamental methods of machine translation.

1.1.1. Rule-based machine translation of the verb forms from Bulgarian to English

Previous researches on the matter have proposed that the similarities between Bulgarian and English are strong enough for constructing transfer-based rules for the grammatical categories and give a reliable linguistic explanation of how the grammatical information is transferred during translation. Different possible rule-based systems have been described for the purposes of constructing reliable transfer-based rules especially for Bulgarian-English translation. A common feature of the rule-based systems is that they consist of several structural layers that aim at deep formal linguistic comprehension of the language data (Iliev, 2014). Although the rule-based method in machine translation is reliable, as it depends on language models, which are constructed by people (and represent exterior linguistic competence), it is still the human perception of the linguistic phenomena. It is needless to point out that for the rule-based method we need large and accurate grammars and dictionaries, which must take into account all possible language variations. The possibilities of the rule-based method can offer an insightful comparativistic view of the linguistic processes that occur during translation, but they have limited application with regard to describing the complex process of translating the grammaticalized information of the verb forms.

1.1.2. Statistical machine translation and statistical translation models

Incorporating linguistic knowledge into statistical models is an everlasting topic in natural language processing. The last two decades of development in the field of NLP are considered to be the second flourishing of applying statistical methods in the field after the 1980's. Recently a number of machine translation efforts have focused on grammatical formalisms for performing source language analysis, transfer rule application and target language generation. It is worth mentioning several works, such as (Bond et. al, 2005) exploiting DELPH-IN1 infrastructure for developing HPSG grammars; (Riezler and Maxwell, 2006) using LFG grammar; working on a hybrid architecture consisting of an LFG grammar, an HPSG grammar, partial parsing; and using the Functional

Generative Description framework for language analysis on analytical and grammatical level. All the approaches rely on the advances in the development of deep grammar natural language parsing. The approaches share similar architecture and techniques to overcome the drawbacks of the deep processing in comparison to statistical shallow methods. Manually created word aligned bi- or multilingual corpora have proven to be useful resources in variety of tasks, e.g. for the development of automatic alignment tools, but also for lexicon extraction, word sense disambiguation, machine translation, annotation transfer, etc. However, one of the limitations of statistical machine translation is that it only translates words within the context of a few words before and after the translated word. For small sentences, it works pretty well. For longer ones, the translation quality can vary from very good to, in some cases, borderline nonsensical. It is almost always possible to see it has been machine-generated. Nowadays the statistical methods are incorporated into the neural machine translation approaches. Neural Machine Translation (NMT) is an end-to-end learning approach for automated translation, with the potential to overcome many of the weaknesses of conventional phrase-based translation systems. Unfortunately, NMT systems are known to be computationally expensive both in training and in translation inference. Also, most NMT systems have difficulty with rare words. These issues have hindered NMT's use in practical deployments and services, where both accuracy and speed are essential. In the late 2000s, a new machine learning technology called deep learning or deep neural networks, one that tries to mimic how the human brain works (at least partially), became a viable option for many hard to crack computer science problems thanks to advances both on the research side (how to build, train and run these large neural networks) and on the imputer side with the arrival of the extremely large scale computing power of the cloud. For the purposes of this article we will restrict ourselves from further discussion on the characteristics of the different MT systems and circumscribe a general description of statistical translation modelling. The statistical translation models consist of two general components:

- Language models: The goal of statistical language modelling is to build a statistical language model that can estimate the distribution of natural language as accurate as possible. A statistical language model is a probability distribution $P(s)$ over strings s that attempts to reflect how frequently a string s occurs as a sentence. Having a reliable language model is the first step towards building a statistical translation model.
- Translation models: The goal of statistical translation modelling is to represent the probability of a string in a target language to be the translation of a string in the source language. A string of a given language (e) is translated according to the probability distribution $p(e|f)$ that a string e in the target language is the translation of a string f in the source language.

Combining these two components a statistical translation model attempts to calculate the most likely translation of a string \hat{e} of the source language:

$$\hat{e} = \operatorname{argmax}_e P(f \vee e) P(e)$$

In this way the probability distribution $p(e|f)$ is calculated by combining the probabilities of the translation model for the two languages and the language model of the target language. A major benefit of this approach is that it allows the use a language model. This can be very useful in improving the fluency or grammaticality of the translation model's output.

As they calculate the statistical translation probabilities, statistical translation models directly depend on the quantity and quality of the available linguistic resources. The main principle of this approach is "more data is better data", thus a statistical model of certain language evaluates the probability of certain string of words to appear not by their grammatical correctness, but by the frequency of their usage in the available resources. That is why the first step towards statistical translation modelling is to have sufficient and dependable linguistic corpora with enough language data to ensure that the constructed models based on these resources are reliable and scientifically effective.

2. Resources for the creation of the Bulgarian – English parallel corpus for the purposes of constructing statistical translation model for verb forms

2.1. Requirements and selection of the suitable resources

The step that precedes the creation of the corpus itself is the collection and evaluation of reliable language resources that are suitable for the purposes of the corpus. As it has been stated before (Lazarov, 2016) there are several existing reliable corpora that can provide linguistic data for the purposes of our project. The fundamental requirements of our corpus determine the major characteristics that the available resources must have:

- the language resources must represent parallel Bulgarian-English sentence-aligned texts;
- in its meta-information it must be stated which language is the original language and which is the translation of the original;
- a verified layer of PoS-tags would be beneficial, but not necessary for our needs. The different types of language corpora contain different metadata. Some corpora do not contain information about the morphological characteristics of words, yet they are a valuable resource for monitoring and describing linguistic phenomena and their verification.

Having defined these requirements for the linguistic data that will be included in the corpus, we restrict our choice to the Bulgarian-English Sentence- and Clause-Aligned Corpus (BulEnAC). BulEnAC was created as a training and evaluation data set for automatic clause alignment in the task of exploring the effect of clause reordering on the performance of SMT. The BulEnAC is an excerpt from the Bulgarian-English Parallel Corpus – a part of the Bulgarian National Corpus (BulNC) of approximately 280.8 million tokens and 8.2 million sentences for Bulgarian and 283.1 million tokens and 8.9 million sentences for English. The Bulgarian-English Parallel Corpus has been processed at several levels: tokenization, sentence splitting, lemmatization. The BulEnAC consists of 366,865 tokens altogether. The Bulgarian texts comprise 176,397 tokens in 14,667 sentences, with average sentence length 12.02 words. The English part totals at 190,468 tokens and 15,718 sentences (12.11 words per sentence). The number of clauses in a sentence averages 1.67 for Bulgarian compared with 1.85 clauses per sentence for English. (Koeva et al, 2012).

Another resource that can provide reliable parallel language data is the Bilingual Library [<http://www.bglibrary.net/>], which although it does not provide PoS-tags or meta-information about the texts, includes a sufficient volume of Bulgarian-English parallel texts, which can be included in our corpus.

2.2. Assessment and relevance evaluation of the selected resources

We have to point out that both of the described resources do not meet the preset requirements for them. Although BulEnAC represents a reliable parallel PoS-tagged Bulgarian – English corpus with a sufficient volume, it does not contain information about the source and target language for each of the consisting sub-corpora. For the purposes of our project such information must be subjectively attached to each set of sentences, based on extra linguistic characteristics such as origin of the text, author, its source, etc. Contrasting to that, the Bilingual Library offers parallel texts with information about their source language, author, target language and translator, but it does not contain any linguistic information about the included texts or any alignment. Each of the resources' advantages and disadvantages were taken into account when the structure of the Bulgarian-English parallel corpus for the purposes of creating statistical translation model for verb forms was constructed. The corpus is constructed of small pieces of both resources, which were evaluated and selected after reviewing not only the linguistic information they can provide, but also the meta-linguistic. The approved resources span from particularly selected single sentences to entire coherent texts from different sources – such as news, short narratives, drama pieces and other literary works. The meta-information of the corpus includes data about the source (file name or URL) of each sub-partition of it, the source and the target language, the date of collection and the date of incorporation, the fact that a sub-partition is part of the BulEnAC provides information about whether it had a PoS-tag layer or not. The PoS-tag layer of

BulEnAC is used for correction and confirmation after both of the annotation layers of our corpus have been implemented.

3. Annotation of the corpus

3.1. The first annotation layer - principles and selected tools

During the phase of collection and evaluation of the appropriate resources for the corpus the problem about its annotation structure arose. The used linguistic material did not have equally distributed quality and quantity of annotation layers therefore the general annotation structure was constructed.

The linguistic data in the corpus has two layers of annotation. The first layer is the PoS-tags layer and it consists of PoS-tags of the words. For both languages the tool TreeTagger is used. The TreeTagger is a tool for annotating text with part-of-speech and lemma information. It was developed by Helmut Schmid (1995) in the TC project at the Institute for Computational Linguistics of the University of Stuttgart. Because the TreeTagger is adaptable to other languages if a lexicon and a manually tagged training corpus are available, it was chosen to be trained on the training data, obtained from the corpus after its initial manual second layer annotation.

The gathered linguistic data was first divided in small working files and, where needed, aligned sentence by sentence. Each aligned pair of sentences in Bulgarian and English receives a unique identifying number in order to be recognizable in the subsequent work. After the initial process of dividing and aligning the data, the TreeTagger is used to annotate both languages. For the tagsets used by the TreeTagger see: Santorini (1991) for English and Simov et al. (2004) for Bulgarian. After the process of annotation the data is manually checked and corrections are applied where needed. The annotated working files are separated for each language and meta-information is added to them. Each file receives an ID number in and it is saved as three column tsv (tab-separated values) file. The first column of each line of the file contains the word/token, the second column represents the lemma of the word and the third column is the prescribed PoS-tag. A blank line represents the sentence boundary.

3.2. The second annotation level - structure and tagset

For the second layer of annotation the tool WebAnno (Yimam et al, 2014) is used. WebAnno is a general purpose web-based annotation tool for a wide range of linguistic annotations including various layers of morphological, syntactical, and semantic annotations. Additionally, custom annotation layers can be defined, allowing WebAnno to be used also for non-linguistic annotation tasks. Different modes of annotation are supported, including a correction mode to review externally pre-annotated data, and an automation mode in which WebAnno learns and offers annotation suggestions. WebAnno accepts several file formats, but for the purposes of our project the CONLL file format was chosen. WebAnno uses a revised version of the CoNLL-X format. Annotations are encoded in plain text files (UTF-8, using only the LF character as line break, including an LF character at the end of file) with three types of lines: Word lines containing the annotation of a word/token in 6 fields separated by single tab characters; Blank lines marking sentence boundaries; and Comment lines. Sentences consist of one or more word lines, and word lines contain the following fields:

- ID number of the sentence – the prescribed unique number of the sentence
- ID number of the word/token – the length of the word/token marked by the initial character and the ending character.
- The word/token
- PoS tag – the first annotation layer
- The verbal tag – the second annotation layer
- Numerical relation between the two annotation layers – which elements of the first annotation layer are included in the second annotation layer.

For examples of the file format see Appendix A.

The second annotation layer is done manually through the WebAnno tool. The tagset of this layer consist of smaller number of possible tags than the first annotation layer. They can be staged over the first

layer. The WebAnno tool treats these entities as chunks, which means that a single tag can be prescribed to more than one entity from the first layer. The tagset of the second annotation layer is presented in Table 1.

Bulgarian		English	
Vaor	Verbal form in Aorist	Vprs	Verbal form in Present Simple
Vfutexact	Verbal form in futurum exactum	Vps	Verbal form in Past Simple
Vfutexpreat	Verbal form in futurum exactum praeteriti	Vfs	Verbal form in Future Simple
Vfutpraet	Verbal form in futurum praeteriti	Vprp	Verbal form in Present Perfect
Vfutur	Verbal form in Futurum	Vpp	Verbal form in Past Perfect
Vimperf	Verbal form in Imperfect	Vfp	Verbal form in Future Perfect
Vperfect	Verbal form in Perfect	Vprc	Verbal form in Present Continuous
Vplusqperf	Verbal form in plusquamperfect	Vpc	Verbal form in Past Continuous
Vpraesens	Verbal form in Praesens	Vfc	Verbal form in Future Continuous
		Vprpc	Verbal form in Present Perfect Continuous
		Vppc	Verbal form in Past Perfect Continuous
		Vfpc	Verbal form in Future Perfect Continuous
		Vfsp	Verbal form in Future Simple in the Past
		Vfpp	Verbal form in Future Perfect in the Past
		Vfcp	Verbal form in Future Continuous in the Past
		Vfpcp	Verbal form in Future Perfect Continuous in the Past

Table 1: Tagset of the second annotation layer

The targeted volume of the training data is 1,000 aligned sentences with the two layers of annotation. Since this layer of annotation is manually done by a single person, the pre-defined annotation conventions with an extended and elaborate tag-set will be made available and published later on after this stage of the annotation process is finished.

4. Current work and evaluation of the working process

The working process on the corpus can be divided in three major stages: collection and evaluation of the linguistic material; annotation of training data; and correction and evaluation of automatically annotated data. The current working flow is concentrated on the second stage. The manual annotation of the targeted volume of the training data appears to be the most time consuming stage of the working process and the most problematic. The assessment of the encountered problems and issues during the first two stages of our current work can be divided as follows:

- Pros:

1. The choice of both tools – the TreeTagger and WebAnno brought most of the positives to the working process. The fact that the TreeTagger provides already established set of annotation conventions provided the opportunity to reuse large sets of already annotated data and thus reduce the technical time needed for annotation and manual correction of the gathered linguistic data. Another contribution of the TreeTagger is that it can be trained on user predefined tagsets which allows alternating the used tagsets at any given point of the working process and creating new unique ones entirely for the purposes of this project.

The other main tool of the project - the WebAnno annotation tool also provides numerous opportunities for working with the collected data and versatile functions that meet the initial needs. One of the most beneficial features of the tool is that it offers import and export of the datasets in more than 12 different working file formats that are suitable for different purposes. The tool also provides different annotation levels which can be independent or logically bounded. In the case of the current work on the corpus the greatest advantage of the tool is its ability to perform a predictive annotation. The predictive annotation of the tool prescribes tags on the subsequent language data with certainty based on the previous occurrences of the tag. It can be tuned to be context-sensitive or subordinate constructed. This feature of the WebAnno annotation tool provides the opportunity to annotate the identical tokens in massive sets of data more efficiently and with fewer errors. It is also applicable for the process of manual correction since it offers the possibility to calculate inconsistencies between the automatically assigned tags.

2. The choice of language material also contributes to the efficiency of the working process and the achieved results. The fundamental requirement for the training data is to be representative. This means that the gathered data must demonstrate all of the studied language phenomena in a variety of contexts. The inner structure and the meta-information of BulEnAC provide the opportunity to select language data based on its targeted qualities – e.g. language pragmatics, source and target language, source of the text, year of publishing, etc. This feature of BulEnAC contributed to the greater variety of language material that is included in the training data.

- Cons.

1. The main problem faced during the manual work on preparing the training data is the insufficient variety of verb forms in both Bulgarian and English. Although the initially selected language resources were able to provide various language materials, they were not able to ensure the grammatical variety of the linguistic data. Previous studies (Lazarov, 2017) have shown that the distribution of tense forms in Bulgarian, as source language, is uneven and reliable statistical data can be obtained through large and representative corpora. The distribution of tense forms according to Lazarov (2017) is provided in Table 2. This work represents statistical data obtained from small corpus (of around 200 sentences) focusing on the frequency of occurrences of Bulgarian tense forms:

Tense form	Frequency of occurrences
Aorist	40,5%
Imperfect	20%
Praesens	19,5%
Futurum	10%
Perfect	4,5%
Futurum praeteriti	2%
Plusquamperfect	0,5%
Other verbal forms	3%

Table 2: Frequency of occurrences of tense forms for Bulgarian

Since this fact would affect the constructed statistical model, we have made several improvements of the initial working data. The initial conception of implementing whole texts in the corpus was dismissed and single not logically connected sentences were introduced to the initial working data.

After the preparation of the training data is completed the deficient tense forms will be artificially constructed and introduced to the training set. The success rate of this method will be assessed during the manual correction of the annotated data based on the model constructed by the training data.

5. Future work and research aims

After the training data is manually annotated, evaluated and completed with artificially constructed tense forms it will be used to train an annotation model on the TreeTagger. The PoS tags are intended to be the primary input data. The output data will be the tagset of the second annotation layer assigned to chunks of tokens from the input layer. The targeted volume of the corpus is 5,000 aligned pairs of sentences with the two layers of annotation. The current workflow aims at creating a corpus with frequency of tense form occurrences close to the presented in Table 2. As can be seen from the data presented in Table 2, the three most frequent tense forms represent more than 75% of the total occurrences. This fact will result in artificially constructed and translated tense forms for the purposes of creating scientifically representative corpus.

The targeted volume of 5,000 entries was determined after analyzing and preparing the suitable resources and after summarizing the available literature on the issue. On one side, although the major part of the preselected linguistic resources (BulEnAC) represent a perfect prerequisite to start the project at the stage of annotating the second layer, it is an automatically annotated resource and thus contains unresolved annotation issues, that have to be resolved beforehand. On the other side, the smaller part of the resources consists of linguistic data that can provide a reasonable diversity of temporal forms to amplify the data. The targeted volume was also determined after considering that most of the temporal forms practically have zero frequency in present-day Bulgarian (Kucarov, 2007). Aiming at collecting equal numbers of examples for all tenses would be labor-intensive and statistically inaccurate since the constructed corpus won't consist of adequate representative data. The targeted volume of the corpus aims at presenting enough translation variations of the Bulgarian temporal forms in English at a satisfactory level for future scientific researches based on the corpus data and the methodology for its construction.

The aims of this project and consequently the creation of the described corpus are to create a statistical translation model for verb forms, which will be based on reliable linguistic data. The statistical model will be able to provide an answer to the initial questions of this research: in what manner the grammatical information is transferred between Bulgarian and English; what type of grammatical information is transferred and what type is lost during the process of translation and why; how close are the verbal morphological categories of both languages and in what manner are they related; in what manner the combination of certain grammatical categories in Bulgarian influences the translation in English. Most of these questions already have elaborate theoretical explanations which will be empirically demonstrated.

The constructed corpus, the gathered scientific data and the constructed statistical language and translation models are envisioned to be freely available linguistic resources for various scientific purposes. The training models and the working files for the TreeTagger and WebAnno will be published together as part of the ready-to-use linguistic resource.

Acknowledgements

This article is part of project No. 72-00-40-221/10.05.2017: „Българо-английски граматични паралели с оглед на машинния превод. Обогаляване на статистически модел за превод от български на английски с лингвистична информация” (Bulgarian-English grammatical parallels with respect to machine translation. Enriching statistical translation mode from Bulgarian to English with linguistic information) sponsored by “Програма за подпомагане на млади учени и докторанти на БАН – 2017” (Support program for young scientists and PhD students at the Bulgarian Academy of Sciences – 2017).

References

- Bond, F., Oepen, S., Siegel, M., & Flickinger, D. (2005). Open source machine translation with DELPH-IN. *Open-Source Machine Translation Workshop at the 10th Machine translation Summit*, (pp. 15-22).
- Iliev, G. (2014). *Ezikovo motivirana optimizatsiya na mashiniya prevod*. Department of Computational Linguistics, IBL-BAS. Retrieved from http://roboread.com/doc/Iliev_G_Dissertation.pdf
- Ivanova, K. (1968). Varhu vzaimotnosheniyata na glagolnata prefiksaciya i kategoriyata prehodnost/neprehodnost v savremenniya balgarski knizhoven ezik. *Slavistitshen sbornik*, 156-157.
- Kabakciev, K. (2000). *Aspect in English: A 'Common-Sense' View of the Interplay Between Verbal and Nominal Referents*. Springer.
- Koeva, S., Rizov, B., Tarpomanova, E., Dimitrova, T., Dekova, R., Stoyanova, I., . . . Genov, A. (2012). Bulgarian-English Sentence- and Clause-Aligned Corpus. *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*.
- Kucarov, I. (2007). *Teoretitshna gramatika na balgarskiya ezik. Morfologiya*. Plovdiv: University press Paisii Hilendarski.
- Lazarov, T. (2016). Analysis of the Resources for Statistical Translation Model of the Verb Forms from Bulgarian to English. *Bulgarian language*, 96-102.
- Lazarov, T. (2017). Functional grammatical parallels with regards of translation of verb forms from Bulgarian to English. *Proceedings of the International Jubilee Conference of the Institute for Bulgarian Language (Sofia, 15 – 16 May 2017)*.
- Nedelcheva, S. (2012). Bulgarian Ingressive Verbs: The Case of Za- and Do-. *Godishnik of University of Shumen Konstantin Preslavsky*, pp. 72-89.
- Riezler, S., & Maxwell, J. T. (2006). Grammatical machine translation. *Human Language Technology Conference - North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL'06)*. New York; NY.
- Santorini, B. (1991). *Part-of-Speech Tagging Guidelines for the Penn Treebank Project*. University of Pennsylvania.
- Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin.
- Simov, K., Osenova, P., & Slavcheva, M. (2004). BulTreeBank Morphosyntactic Tagset. In *BulTreeBank Technical Report BTB-TR03*.
- Sinapova, L., & Dochev, D. (1999). Analyzing Bulgarian and English Collocations. *Problems of Engineering Cybernetics and Robotics*.
- Vassileva, A. (2003). Tense variations in Bulgarian narratives and their translational equivalents in English. Plovdiv University "Paisii Hilendarski". Retrieved from <http://georgesg.info/belb/doktoranti/vasileva/MA2.htm>
- Yimam, S., Castilho, E., & Gurevych, I. (2014). Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. *Proceedings of ACL-2014, demo session*. Baltimore, MD, USA.

Appendices

```

1 #Text=012bg
2 1-1 0-4 фонд Ncmsi
3 1-2 5-7 ще Tx V̄futur[1] *->1-1
4 1-3 8-16 подкрепя Vpitf-r3s V̄futur[1] *->1-2
5 1-4 17-26 Балкански A-pi - -
6 1-5 27-36 кинодейци Ncmpi - -
7 1-6 37-39 . sent - -
8
9 #Text=012en
10 1-1 0-4 Fund NN
11 1-2 5-6 will MD V̄fs[1] *->1-1
12 1-3 7-14 support VV V̄fs[1] *->1-2
13 1-4 15-21 Balkan JJ - -
14 1-5 22-38 cinematographers NNS - -
15 1-6 39-40 . sent - -
16
17

```

Appendix A. Example of the file format.

Fingerprints in SMS Messages: Automatic Recognition of a Short Message Sender Using Gradient Boosting

Branislava Šandrih

Faculty of Philology, University of Belgrade
branislava.sandrih@fil.bg.ac.rs

Abstract

This paper considers the following question: Is it possible to tell who is the short message sender just by analyzing a typing style of the sender, and not the meaning of the content itself? If possible, how reliable would the judgment be? Are we leaving some kind of “fingerprint” when we text, and can we tell something about others based just on their typing style? For this purpose, a corpus of $\sim 5,500$ SMS messages was gathered from one person’s cell phone and two gradient boost classifiers were built: first one is trying to distinguish whether the message was sent by this exact person (cell phone owner) or by someone else; second one was trained to distinguish between messages sent by some public service (e.g. parking service, bank reports etc.) and messages sent by humans. The performance of the classifiers was evaluated in the 5-fold cross-validation setting, resulting in 73.6% and 99.3% overall accuracy for the first and the second classifier, respectively.

1. Introduction

It does not happen so rarely that we just see a message and know the sender, without even looking at the message header. Even though we miss signature, voice, mimics, sound and so many other components that written and oral communication contains, just by usage of emoticons, abbreviations, specific typos, grammar misses or specific use of punctuation — we can assume who are we communicating with. This is primarily true for the people with specific typing style. In the case of very short message, e.g. “Where are you?”, determination of the sender can become more difficult. The task is not easy at all even for humans, especially when we do not have any other information such as cell phone model of the sender, operative system the sender uses, location etc.

In this paper, Gradient Boost (Friedman, 2001; Hastie et al., 2009) model was trained in order to predict the message sender. This is done by using lexical and syntactic features. Extracted features are put on disposal as CSV file. The dataset, Python module for feature extraction and code for model training and evaluation are available at github.¹ Since the external validation dataset was not available, the performance estimation is done by using 5-fold cross validation (CV). Although no external language tools were used (such as dictionaries or taggers), the method is designed to achieve the best performance on Serbian text messages.

This paper is organized as follows. Related work done so far is listed and briefly described in section 2.. In Section 3. we describe underlying SMS dataset and in Section 4. we describe extracted features. In Section 5. the steps of creating classifiers and most relevant features used by these two models are described. Afterwards in Section 6. we display classification results of our models. Finally, we conclude paper and state future plans in Section 7.

¹Github repository, https://github.com/Branislava/sms_fingerprint

2. Related Work

Regarding the problem of author recognition, two most prominent research fields are Authorship Attribution (AA) and User Profiling. Most of the work done so far was related to the semantic analysis of the content (Pennebaker and King, 1999; Mairesse et al., 2007). Concerning AA, another approach in solving task of automatic recognition of the given text’s author is by observing *stylometric* cues. These stylometric features (Roffo et al., 2014: 33) include *lexical* (counts of words and characters in text) and *syntactic* (punctuation and emoticons) features. After extraction of these features, they are typically used with discriminative classifiers, so that each author represents one class. A survey about application of AA to Instant Messaging (IM) was conducted in (Stamatatos, 2009). In (Zheng et al., 2006) stylometric features were used with Support Vector Machine (SVM) and Artificial Neural Network (ANN) classifiers, while authors of (Abbasi et al., 2008) applied PCA projection for AA on corpora containing E-mails, IMs, feedback comments and even program code. Similar work on AA in IMs was also conducted in (Abbasi and Chen, 2008).

In (Orebaugh and Allnut, 2009) authors identify participants within IM conversation by observing sentence structure and usage of special characters, emoticons and abbreviations. Writing style of individuals is in focus in (Roffo et al., 2014). Authors analyze whether special interactional behavior, as the one present in the live communication, can emerge in chats. They also inspect if certain personality traits affect writing style. Authors conclude that some traits significantly affect chatting style and that some of them can be very effective with identifying a person among diverse individuals.

Similar research is conducted in (Eckersley, 2010) and (Laperdrix et al., 2016). These authors are more oriented at determining how trackable certain computer configuration is, based on Web browser version, the underlying operating system, the way emojis are displayed within Web browser, etc.²

3. The Dataset

A corpus of 5551 short messages structured as XML was collected from one person’s cell phone in a 4-years time period. Each message contains information about sender’s phone number, a date the message was sent, content of the message and other technical information. The corpus mostly consists of messages in Serbian, typed in both letters, Latin and Cyrillic, with some messages in English and German. The following two messages from the corpus are written in different letters, asking the same question in two different ways.³

```
<sms address="+381643057***" date="1424530897293" type="1" contact_name="Gri***"
readable_date="21.02.2015 4:01:37 PM" body="Disiiiiiiiiiiiiiiiiiiiiiiii :-)" />
```

```
<sms address="+381600854***" date="1511436828568" type="0" contact_name="MaJI***"
readable_date="23.11.2017. 12:33:48 PM" body="Где си?" />
```

Attribute *type* represents whether the message was sent (value 1) or received (value 0). DTD for this corpora and the total list of features and their values is available on github.⁴

4. Features and Fingerprinting

Stylometric features⁵ were extracted only from *body* attribute of `<sms>` elements and they can be divided into two categories: 1) lexical features and 2) syntactic features. This categorization is obtained from (Roffo et al., 2014: 33). Bag-of-Words features were not added to the final model as it turned out

²Am I Unique?, <https://amiunique.org/>

³Both messages contain “Where are you?” question, which is a common greeting line in Serbian. First message contains informal dialect-specific greeting, what can be observed by use of repeated letters and an emoticon. Second message is written in Cyrillic, that is normally less used in informal communication.

⁴DTD and extracted features
https://github.com/Branislava/sms_fingerprint/tree/master/dataset.
 For XML files with corresponding DTD, features can be extracted with Dataset class
https://github.com/Branislava/sms_fingerprint/blob/master/features_extraction/dataset.py

⁵Other authors use similar set of features, naming them “linguistic features”, e.g. in (Ebert, 2017: 55) and (Repar and Pollak, 2017).

they did not improve classification accuracy in this case. Along with dominant stylometric features, the final feature set was enriched with an additional set of common abbreviations and slang words.

4.1. Lexical Features

In this category, 11 features were extracted: number of characters, number of Cyrillic characters, diacritics count, number of umlauts, number of uppercase characters, number of lowercase characters, digits count, number of alphabet characters, number of occurrences of same consecutive characters,⁶ number of sentences starting with lowercase character⁷ and number of words starting with “ne”.⁸

These features were selected after careful analysis performed by human, as it seemed they could help with distinguishing message senders that make specific typos and grammatical mistakes, or the ones that write too long or very short messages. For example, the minority of senders write in uppercase or in Cyrillic only and ones that write in German (hence the umlauts count).

4.2. Syntactic Features

These features can be divided into two categories: 1) emoticons and 2) punctuation features.

4.2.1. Emoticons

Ninety-eight different emoticons were listed and classified into 9 groups. First group consists of emoticons that represent a smile (smiley), second one contains emoticons that have a happy face (happy) and similarly other groups are formed: sad, surprised, kiss, wink, tongue, skeptic, miscellaneous.⁹ In this specific dataset not all emoticons are present, and therefore the ones that are missing were discarded during preparation phase, keeping thirty-four emoticons. They are represented with corresponding regular expressions:

```

kiss :* :*{2,} :-* :-*{2,}
tongue :-p{2,} :p{2,} :-P{2,} :-P{2,}
sad :( :-( {2,} :-( :({2,} :-' ( :-' ( {2,}
smiley :-) ;) :) ({2,}: (:
wink ;-) ;) {2,} ;-){2,}
happy xD{2,} xD :D{2,} :-D{2,} :D
skeptic :/{2,} :/
surprised :o :-o
kiss =D =] 8-)
    
```

An absolute count of each emoticon appearance per message was added as a single feature. Afterwards, additional nine features were added as aggregated count of each emoticon type (e.g. total number of smiley emoticons, total count of all happy emoticons in a message etc.).

Emoticons have been useful in many research topics, such as sentiment analysis (Read, 2005; Škorić, 2017) or in short messages interpreting (Walther and D’Addario, 2001; Derks et al., 2007).

⁶Repeated characters, like in word “Disiiiiiiiiiiiiiii” make an impression of a person in a good mood.

⁷If one starts most sentences with lowercase characters, that is probably due to mobile phone operating system, what can be a partially identifying feature.

⁸Negation of verbs in Serbian is made by adding word “ne” before the verb, separately. It is a common mistake that people type this as one word, e.g. instead of correct negation “ne mogu”, one could write incorrectly “nemogu”.

⁹All emoticons with corresponding regular expressions

https://github.com/Branislava/sms_fingerprint/blob/master/features_extraction/emoji.py

4.2.2. Punctuation

For this dataset, nine punctuation-related features were considered to be important for sender dissemination: count of exclamation marks, count of question marks, count of dots, count of commas, total count of present punctuation, times when space followed punctuation, number of sentences separated by dot that does not precede space, count of . . (double dot) and count of ?? tokens.

These features are extracted with an idea that certain people always make similar typing mistakes. For example, some people tend to write “bad” punctuation, such as two dots (instead of one or three) or do not write spaces after punctuation, they “glue” the sentences together with a dot and no additional blank space, etc.

4.3. Combining Lexical and Punctuation Features

Sixteen more features were added as a ratio of already mentioned feature counts. These features are ratios of: number of exclamation marks/question marks/dots/commas/total punctuation/alphabetic characters/diacritics/umlauts/cyrillic/uppercase/lowercase/digits and message length, ratio of upper and lowercase characters and ratio of punctuation/cyrillic/digits and alphabetic characters. The list of all extracted text is available at github.

4.4. The Abbreviation Features

Abbreviations are very common in this specific dataset, and therefore a list of total 135 different abbreviations was made. Some of the abbreviations are: *ae* (hajde - “come on”), *dog* (dogovoreno - “deal”), *dop* (dopisivati - “chat”), *k* (ok - “ok”), *msm* (mislim - “I think”), *mzd* (možda - “perhaps”), *najrvr* (najverovatnije - “most probably”), *nmg* (ne mogu - “I cannot”), *nmvz* (nema veze - “nevermind”), *nnc* (nema na čemu - “you welcome”), *np* (nema problema - “no problem”), *npm* (nemam pojma - “I have no clue”), *nzm* (ne znam - “I don’t know”), *stv* (stvarno - “really”), *ustv* (u stvari - “actually”), *ves* (večeras - “tonight”), *zvrc* (zovi me - “call me”) etc.

5. Classification Model and Results

Two experiments were run, both with binary classification task. After several different classifiers evaluation, Gradient Boost model (Friedman, 2001; Hastie et al., 2009) turned out to achieve the best precision and accuracy in both cases. Detailed classifiers comparison is given in Section 6.

5.1. First Experiment: Specific Person vs. Others

The classification model was built to tell whether an unseen message was written by the native cell phone owner (positive class, label 0) or by someone else (negative class, label 1). Class labels are induced from *type* attribute of <sms> element explained in Section 3. There are 2,170 instances belonging to positive class and 3,381 instances belonging to negative class, making this dataset slightly unbalanced.

List of fifteen features that had the strongest influence on the model can be seen in Figure 1.¹⁰ The majority of the most influential features are lexical and punctuation features: ratio of uppercase characters and message length (significant for persons who write in uppercase), message length, ratio of upper and lowercase letters, presence of spaces after punctuation, usage of question marks and dots. The fact that these features showed up as most important was not a surprise, since it was expected that exactly these features are what makes person’s typing style distinguishable from other senders’.

5.2. Second Experiment: Human vs. Machine

Although the dataset is quite unbalanced in this case, the task is much easier than the previous. There are 918 instances belonging to positive class (label 0, messages sent from public services such as bank reports, parking services, mobile service providers etc.) and 4,633 messages sent from humans (label 1). Top 15 features that had the strongest influence were shown in Figure 2.

¹⁰Order of these most important features may vary during different cross-validation folds (depending on the message instances selected for the training set).

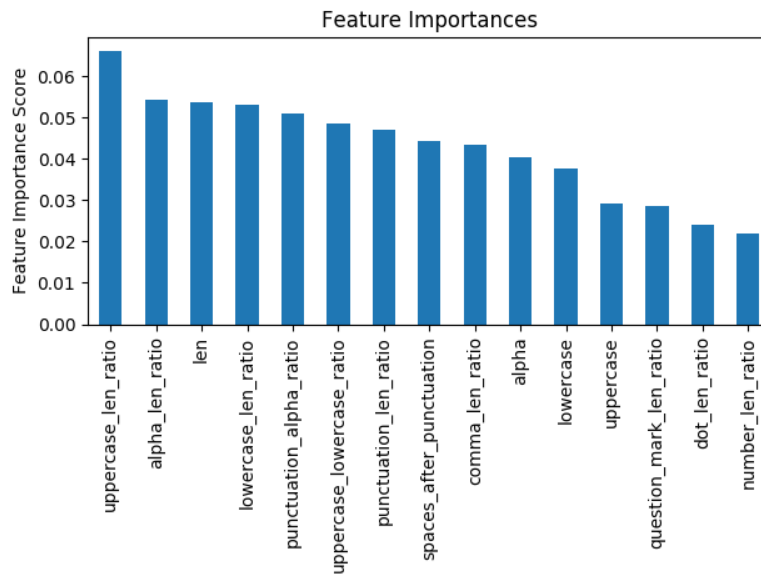


Figure 1: Most important features for the first model

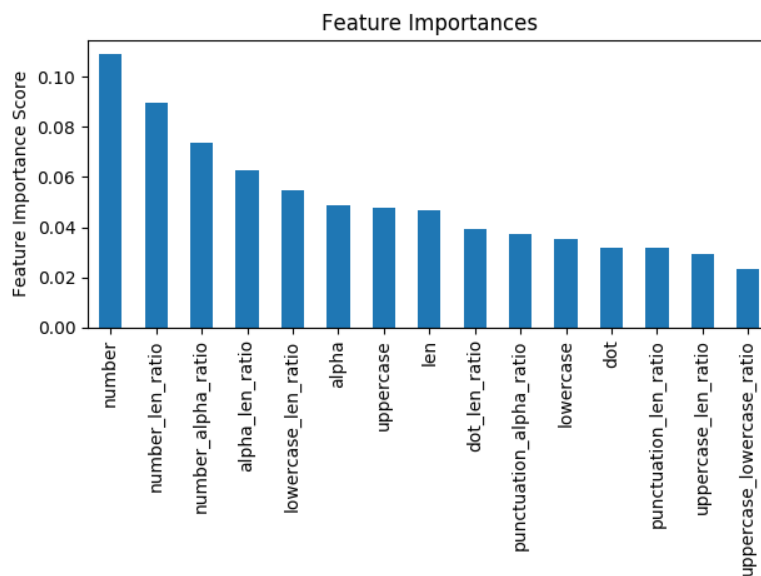


Figure 2: Most important features for the second model

Most of these features are related to presence of numbers, which is expected. These reports mainly consist of different digits that represent date and time when the report was sent, amount of money in a bank account, time when the parking card expires, etc. Similarly, these messages length is also somewhat specific, i.e. reports usually contain more tokens than regular humans' messages. Another common feature is number of the dot character used in comparison to other characters. Reports are usually longer and contain a few sentences, each concluded with a dot, which could not be guaranteed for informal messages. We can also notice that features have stronger influence (higher scores, *y*-axis) than in the previous experiment.

6. Results

We tested and compared the following algorithms implemented in *SciKit-Learn*, Machine Learning module for Python (Pedregosa et al., 2011):

SVM Support Vector Machine is a supervised machine learning algorithm that can be used for both classification and regression problems. Classification is performed by finding the hyper-plane that separates samples from different classes with the highest possible margin. In the case that samples are linearly separable, i.e. it is possible to find a hyper-plane that separates training samples good enough, SVM is linear. If samples are not linearly separable, a kernel function for the classifier should be selected. This means mapping all samples into other, higher-dimensional space, where the separating hyper-plane can be obtained. Beside kernel function, parameters of this classifier are: penalty parameter C , γ (ignored if kernel is not Radial Basis Function (RBF), default value *auto*), tol (tolerance for stopping criterion, default value is 0.001), $class_weight$ (if not given, all classes are supposed to have weight 1) and max_iter (maximum number of iterations; by default, this number is unlimited).¹¹

MLP Artificial Neural Network (ANN) is a machine learning framework that attempts to mimic the learning pattern of natural biological neural networks. Multi-layer Perceptron is a class of feed-forward ANNs that consists of at least three layers of nodes (input, hidden and output layer). It is a supervised learning algorithm that, given a set of features, can learn a non-linear function approximator for either classification or regression task. As for parameters, except regularization term $\alpha = 1$, we used default values: $hidden_layer_sizes = 100$, the rectified linear unit function (relu) for $activation$ parameter, stochastic gradient-based optimizer (adam) for $solver$, $tol = 0.0001$ as tolerance for optimization etc.

Gradient Boost Gradient boosting is a sequential technique that combines a set of weak learners, usually decision trees, and delivers improved prediction accuracy in an iterative fashion. Trees are added one at a time and a gradient descent procedure is used to minimize the loss when adding new trees. After calculating error or loss, the outcomes predicted correctly are given a lower weight and the ones miss-classified are weighted higher, until best instance weights are found. Before building the final classifier, grid search was performed in order to find optimal classifier parameters. At the end, model was tuned with next parameter values: $learning_rate = 0.1$, $n_estimators = 160$, $min_samples_split = 10$, $min_samples_leaf = 30$, $max_depth = 9$, $max_features = 11$, $subsample = 0.8$ and $random_state = 10$.

The performance of the classifiers was evaluated in the 5-fold CV setting using the following basic measures: accuracy, precision, recall and F-score. As a baseline, a classifier that always predicts the majority class in the dataset was used.

Detailed results for the 1st experiment are given in Table 1.

Classifier	Accuracy	Precision (+ class)	Recall (+ class)	F-score (+ class)	Precision (- class)	Recall (- class)	F-score (- class)
Baseline	0.609	0.000	0.000	0.000	0.609	1.000	0.757
Linear SVM (C=0.025)	0.714	0.643	0.612	0.619	0.763	0.779	0.768
Linear SVM (C=1)	0.715	0.641	0.635	0.631	0.769	0.766	0.764
RBF SVM	0.619	0.708	0.049	0.091	0.617	0.984	0.759
Neural Net	0.686	0.656	0.485	0.528	0.723	0.815	0.757
Gradient Boosting Classifier	0.736	0.673	0.641	0.653	0.777	0.796	0.785

Table 1: Classification results for the 1st experiment with different algorithms and parameter settings

For detailed results of the 2nd experiment see Table 2.

¹¹Support Vector Classifier class in *sklearn*
<http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>

Classifier	Accuracy	Precision (+ class)	Recall (+ class)	F-score (+ class)	Precision (- class)	Recall (- class)	F-score (- class)
Baseline	0.835	0.000	0.000	0.000	0.835	1.000	0.910
Linear SVM (C=0.025)	0.984	0.964	0.937	0.950	0.988	0.993	0.990
Linear SVM (C=1)	0.989	0.968	0.966	0.967	0.993	0.994	0.993
RBF SVM	0.947	1.000	0.679	0.805	0.940	1.000	0.969
Neural Net	0.982	0.939	0.953	0.946	0.991	0.987	0.989
Gradient Boosting Classifier	0.993	0.984	0.973	0.978	0.995	0.997	0.996

Table 2: Classification results for the 2nd experiment with different algorithms and parameter settings

7. Conclusion and Future Work

The method described in this paper is aimed at solving supervised classification task on short Serbian messages. In order to solve this task in supervised manner, it is important to have representative corpus of SMS data and metadata such as sender's name, phone number, etc. Due to privacy concerns, people are having trust issues and are not willing to share their SMS messages. As a consequence, evaluation of the method developed in this paper is not performed on other datasets with the same structure and content. Twitter data might seem as a good candidate (Twitter corpora are publicly available, there is the same character count threshold and there is plenty of it), but Twitter posts and SMS messages are not having the same purpose. SMS message is addressed for specific person and most often asks question or answers one. Twitter posts mostly contain opinions or comments, referring to other users or topics using hash tags. These hash tags are very common in tweets and can be a rich source of even more text features. Although the problem itself could be stated on any type of text that is interchanged between two or more sides (Facebook posts, tweets, E-mails, SMS messages, forum posts, Viber/WhatsApp messages etc.), it is expected that, due to difference in purpose of these different services, different approach should be applied for each.

Examining only emoticons, punctuation usage or abbreviations is not enough to identify a person. Even for a human, it would be impossible to tell difference between persons who are writing with perfect grammar and without emoticons. But with additional information like one used in (Laperdrix et al., 2016) and (Eckersley, 2010), this task might be simple. In the future work, it is intended to generalize the problem so Facebook and Twitter posts can be evaluated. This is primarily aimed at enriching model with new features, such as message semantics (word meanings, context, used language dialect and chat history), sender's gender, common phrases used by a sender and even information about the device from which the message is sent (e.g. the device model or underlying operating system).

At the time being, current results are implying that this kind of identification is possible, at least as one of the steps in the authorship attribution.

Acknowledgements

This research was supported by the Serbian Ministry of Education and Science under grant #178006.

References

- Abbasi, A. and Chen, H. (2008). Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace. *ACM TOIS*, 26(2):1–29.
- Abbasi, A., Chen, H., and Nunamaker, J. (2008). Stylometric Identification in Electronic Markets: Scalability and Robustness. *Journal of Management Information Systems (JMIS)*, 25(1):49–78.
- Derks, D., Bos, A. E., and Von Grumbkow, J. (2007). Emoticons and Social Interaction on the Internet: the Importance of Social Context. *Computers in human behavior*, 23(1):842–849.
- Ebert, S. (2017). *Artificial Neural Network Methods Applied to Sentiment Analysis*. Ph.D. thesis, Ludwig-Maximilians-Universität München.
- Eckersley, P. (2010). How Unique is your Web Browser? In *Proceedings of the 10th International Conference on Privacy Enhancing Technologies*, volume 6205, pages 1–18. Berlin, Heidelberg: Springer - Verlag.
- Friedman, J. H. (2001). Greedy Function Approximation: a Gradient Boosting Machine. *Annals of statistics*, pages 1189–1232.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). Springer, 2 edition.
- Laperdrix, P., Rudametkin, W., and Baudry, B. (2016). Beauty and the Beast: Diverting Modern Web Browsers to Build Unique Browser Fingerprints. In *37th IEEE Symposium on Security and Privacy (S&P 2016)*, San Jose, United States, pages 878–894. IEEE.
- Mairesse, F., Walker, M., Mehl, M., and Moore, R. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research (JAIR)*, 30:457–500.
- Orebaugh, A. and Allnutt, J. (2009). Classification of Instant Messaging Communications for Forensics Analysis. *Social Networks*, pages 22–28.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.
- Pennebaker, J. and King, L. (1999). Linguistic Styles: Language Use as an Individual Difference. *Journal of Personality and Social Psychology*, 77(6):1296–1312.
- Read, J. (2005). Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In *Proceedings of the ACL student research workshop*, pages 43–48. Association for Computational Linguistics.
- Repar, A. and Pollak, S. (2017). Good Examples for Terminology Databases in Translation Industry. In *eLex 2017: eLex 2017: The 5th biennial conference on electronic lexicography, Netherlands, 19-21 September 2017*, pages 651–661.
- Roffo, G., Giorgetta, C., Ferrario, R., and Cristani, M. (2014). Just the Way You Chat: Linking Personality, Style and Recognizability in Chats. In *International Workshop on Human Behavior Understanding*, pages 30–41. Springer.
- Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the Association for Information Science and Technology (JASIST)*, 30(3):538–556.
- Škorić, M. (2017). Classification of Terms on a Positive-negative Feelings Polarity Scale Based on Emoticons. *Infotheca: Journal for Digital Humanities*, 17(1):67–91.
- Walther, J. B. and D’Addario, K. P. (2001). The Impacts of Emoticons on Message Interpretation in Computer-Mediated Communication. *Social science computer review*, 19(3):324–347.
- Zheng, R., Li, J., Chen, H., and Huang, Z. (2006). A Framework for Authorship Identification of Online Messages: Writing-style Features and Classification Techniques. *Journal of the Association for Information Science and Technology (JASIST)*, 57(3):378–393.



Department of Computational Linguistics

<http://dcl.bas.bg/>

ISSN: 2367-5675