Computational Linguistics in Bulgaria

'16

SECOND INTERNATIONAL CONFERENCE

COMPUTATIONAL LINGUISTICS IN BULGARIA

CLIB 2016

9 September **2016**

Sofia, Bulgaria

DEPARTMENT OF COMPUTATIONAL LINGUISTICS

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

The Second International Conference *Computational Linguistics in Bulgaria (CLIB 2016)* is organised within the Operation for Support for International Scientific Conferences Held in Bulgaria of the National Science Fund Grant № ДПМНФ 01/9 of 11 Aug 2016.

National Science Fund

CLIB 2016 is organised by:

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

PUBLICATION AND CATALOGUING INFORMATION

Title:	**Proceedings of the Second International Conference *Computational Linguistics in Bulgaria (CLIB 2016)***
ISSN:	**2367-5675 (online)**
Published and distributed by:	**The Institute for Bulgarian Language *Prof. Lyubomir Andreychin*, Bulgarian Academy of Sciences**
Editorial address:	**Institute for Bulgarian Language *Prof. Lyubomir Andreychin*, Bulgarian Academy of Sciences** **52 Shipchenski prohod blvd., bldg. 17** **Sofia 1113, Bulgaria** **+359 2/ 872 23 02**

Proceedings of the

Second International Conference

Computational Linguistics in Bulgaria

9 September 2016
Sofia, Bulgaria

PREFACE

We are excited to welcome you to the second edition of the International Conference *Computational Linguistics in Bulgaria* (CLIB 2016) in Sofia, Bulgaria!

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

We would like to thank the authors who trusted us and submitted their contributions to CLIB 2016. Their efforts and high-quality research are the chief factor that enabled us to create an interesting and solid scientific programme. We would also like to thank our industrial participants for sharing their insights, ideas and know-how with the research community.

We would like to express our sincere gratitude to the members of the Programme Committee, who accepted to join us and invested a lot of expertise to provide valuable feedback to the authors. Special thanks are due to Prof. Svetla Kœva, who is the person behind the whole CLIB concept. We hope that CLIB 2016 will be a useful and productive experience that we all will enjoy!

CLIB 2016 Organising Committee

PROGRAMME COMMITTEE

Svetla Koeva (Chair) – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Cvetana Krstev – University of Belgrade
Denis Maurel – François-Rabelais University of Tours
Dragomir Radev – University of Michigan, Department of Electrical Engineering and Computer Science
Duško Vitas – University of Belgrade
Éric Laporte – University of Paris-Est Marne-la-Vallée
Hristo Krushkov – Plovdiv University *Paisii Hilendarski*
Hristo Tanev – Joint Research Centre of the European Commission, Ispra, Italy
Ivan Derzhanski – Institute of Mathematics and Informatics, Bulgarian Academy of Sciences
Ivelina Nikolova – Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Jan Šnajder – University of Zagreb, TakeLab
Karel Oliva – Institute of the Czech Language, Academy of Sciences of the Czech Republic
Kjetil Rå Hauge – University of Oslo
Maciej Ogrodniczuk – Institute of Computer Science, Polish Academy of Sciences
Maciej Piasecki – Wrocław University of Technology
Mariana Damova – *Mozaika* Ltd., Bulgaria
Marko Tadić – University of Zagreb
Mila Dimitrova-Vulchanova – Norwegian University of Science and Technology
Nikolay Vazov – University of Oslo
Preslav Nakov – Qatar Computing Research Institute, Hamad bin Khalifa University
Radka Vlahova – Sofia University *St. Kliment Ohridski*
Radovan Garabík – *Ľudovít Štúr* Institute of Linguistics, Slovak Academy of Sciences
Ruslan Mitkov – University of Wolverhampton
Stoyan Mihov – Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Tania Avgustinova – Saarland University
Verginica Barbu Mititelu – Research Institute for Artificial Intelligence, Romanian Academy
Zornitsa Kozareva – Yahoo! Labs

ORGANISING COMMITTEE

Svetlozara Leseva – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Tsvetana Dimitrova – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Ivelina Stoyanova – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Maria Todorova – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Valentina Stefanova – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Borislav Rizov – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Dimitar Hristov – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Martin Yalamov – Institute for Bulgarian Language, Bulgarian Academy of Sciences
Ekaterina Tarpomanova – Sofia University *St. Kliment Ohridski*
Rositsa Dekova – Plovdiv University *Paisii Hilendarski*

INVITED TALK

Dr. Preslav Nakov
(Qatar Computing Research Institute, HBKU)

Exposing Paid Opinion Manipulation Trolls in News Community Forums

The practice of using opinion manipulation trolls has been reality since the rise of Internet and community forums. It has been shown that user opinions about products, companies and politics can be influenced by posts by other users in online forums and social networks. This makes it easy for companies and political parties to gain popularity by paying for "reputation management" to people or companies that write in discussion forums and social networks fake opinions from fake profiles.

During the 2013-2014 Bulgarian protests against the Oresharski cabinet, social networks and news community forums became the main "battle grounds" between supporters and opponents of the government. In that period, there was a very notable presence and activity of government supporters in Web forums. In series of leaked documents in the independent Bulgarian media Bivol, it was alleged that the ruling Socialist party was paying Internet trolls with EU Parliament money. Allegedly, these trolls were hired by a PR agency and were given specific instructions what to write.

A natural question is whether such trolls can be found and exposed automatically. This is a very hard task, as there is no enough data to train a classifier; yet, it is possible to obtain some test data, as these trolls are sometimes caught and widely exposed (e.g., by Bivol). Yet, one still needs training data. We solve the problem by assuming that a user who is called a troll by several different people is likely to be one, and one who has never been called a troll is unlikely to be such. We compare the profiles of (i) paid trolls vs. (ii) "mentioned" trolls vs. (iii) non-trolls, and we further show that a classifier trained to distinguish (ii) from (iii) does quite well also at telling apart (i) from (iii).

KEYNOTE TALK

Prof. Dragomir Radev
(Department of Electrical Engineering and Computer Science, University of Michigan)

Natural Language Processing for Collective Discourse

Natural Language Processing (NLP) has become very popular in recent years thanks to new technologies like IBM's Watson, Apple's Siri, Google Translate, and Yahoo's text summarization system. One of the fundamental challenges in NLP is to automatically recognize similar words and sentences. I will talk about research done in the Computational Linguistics And Information Retrieval lab (CLAIR) on graph-based methods for similarity recognition and its applications to NLP tasks. These projects are related to Collective Discourse (text collections produced by large numbers of users) and its inherent properties such as centrality and diversity. In the first project we team up with the New Yorker magazine. Each week a captionless cartoon is published in the magazine and thousands of readers try to come up with funny captions for it. In our work, we try to uncover the topics of the jokes in the submitted captions. The second project is about analysing a corpus of word clues used in New York Times crossword puzzles. We compare different clustering methods for word sense disambiguation using these crossword clues. The third project is about the automatic generation of citation-based summaries of research articles. These summaries describe what readers of the papers find most important in the cited papers. If there is time, I will also briefly mention some applications to bioinformatics, political science, and social network analysis.

Table of Contents

How to Differentiate the Closely Related Standard Languages?

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

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

Ljubomir Popović
University of Belgrade
Faculty of Philology
foljupo@gmail.com

Andjelka Zečević
University of Belgrade
Faculty of Mathematics
andjelkaz@matf.bg.ac.rs

Abstract

In this paper the adequacy of the SETimes corpus as a basis for the comparison of closely related languages that are used in countries that emerged after the breakup of Yugoslavia is discussed by comparing it with other corpora. It is shown that the phenomena observed in this corpus and used to illustrate differences most specifically between Serbian and Croatian are consistent neither with their standards nor with other sources. Thus, results obtained on the basis of the SETimes corpus are corpus-biased and have to be reconsidered. This proves that the size of a corpus and its composition used in a linguistic research are crucial for assessing the obtained results.

1. Introduction

On the website *Southeast European Times*[1] the same news were published in English and in the languages of the Balkans, thus its content naturally imposed as a possible source for the creation of a parallel corpus of Balkan languages (Tyers and Alperen, 2010). A narrower version of the contents of this website served to list and illustrate examples of differences that exist between Serbian, Croatian and Bosniak language (Bekavac et al., 2008). Tiedemann and Ljubešić (Tiedemann and Ljubešić, 2012) used the material from this website[2] as a training set for the machine learning methods used for the procedure proposed for the differentiation of these three languages. Starting from this material other experiments were carried out as well such as, for example, the analysis of the possibility of transferring method of morphological processing from Croatian to Serbian (Agić et al., 2013) or experiments in the field of machine translation (Popović and Ljubešić, 2014). What should be emphasized here is that, in accordance with the afore-mentioned works, it can be concluded that the content of the website SETimes is a relevant source for resolving the issue of relationship between Serbian and Croatian.

Such resources, as well as experiments on them, are really useful and desirable as they complement the panorama of resources and methods for less-resourced languages, which include Serbian, Croatian and Bosniak. Thus, for example, it is very useful to have a reliable and objective method to identify in which of *today's official standard languages* a particular text was written. In doing so, we should not forget that these languages have long been regarded as one (Serbo-Croatian) language and that the texts on one of them are to the greatest extent understandable to readers coming from the territory of other languages that derived from Serbo-Croatian.

The question of differentiating these languages is a difficult task as they largely coincide, forming the so-called Neo-Shtokavian standard language diasystem (Popović, 2004). Therefore, the corpora must consistently reflect the differences that characterize these standards. If this is not the case, the results will — regardless of the quality of the applied methods and extent of resources — provide a misleading image of each language, as well as their mutual relationship.

[1] https://web.archive.org/web/*/http://www.setimes.com/. The website was shut down in April 2015. See also https://en.wikipedia.org/wiki/Southeast_European_Times

[2] http://nlp.ffzg.hr/resources/corpora/setimes/

In the light of the above observation, the aim of this paper is to examine the extent to which Serbian and Croatian corpora, made up of material from the website SETimes, are reliable in this respect, taking into consideration the other corpora of Serbian and Croatian languages, as well as the applicable official standards of these two languages. This paper will briefly indicate characteristic differences that some authors have noted in this corpus (Section 2.). Within Section 4., we will examine the relevance of these differences in comparison to other available corpora of Serbian and Croatian languages and compare their frequencies with data obtained from other corpora of these languages. Within Section 5., we will demonstrate that the SET-corpus differs from all the other corpora, which calls into question the validity of the results, while we will give an example of a simple criterion that could reliably identify the Croatian texts in Section 6..

Bearing in mind that we will often refer to SETimes corpora within the paper, we will indicate the Serbian part of this corpus with ST-sr and Croatian part with ST-hr.

2. The differences that were put forward

Based on the analysis of ST-corpus, the above-mentioned authors put forward a number of differences that exist between Serbian, Croatian and Bosniak. This paper will deal primarily with the differences between Serbian and Croatian, and where necessary, we will also include Bosniak examples.

2.1. Ekavian/Ijekavian

It is stated both in (Bekavac et al., 2008) and (Tiedemann and Ljubešić, 2012) that the Ijekavian pronunciation is characteristic of the Croatian (and Bosniak) language and that Ekavian is typical for the Serbian language,[3] and for this assertion they find confirmation in the ST-corpora. This is entirely wrong. *Namely, the Serbian language uses both Ekavian and Ijekavian pronunciation, at the level of standards, as well as in common usage*, thus the corpus of Serbian language that does not include an adequate sample of Ijekavian texts is not representative of the Serbian language. This kind of error causes erroneous results on the level of classification of languages as shown in (Zečević and Vujičić-Stanković, 2013). For example, by leaving out the texts written on Ijekavian Serbian from the corpus, Bosniak is made more similar to Croatian, and is set in an unjustified counter-distinctive relationship towards the Serbian language. It should be noted that *Službeni list BIH*, the official gazette of Bosnia and Herzegovina is published in all three languages, while Serbian version is always in Ijekavian pronunciation.[4]

Let us mention that the number of lexemes that are different in these two pronunciations is finite and that they can be mapped one-to-one, from one pronunciation to another. Some differences exist in the way forms are derived [5] but these are differences at the morphological level, not at the level of pronunciation.

2.2. Future tense

One of the two forms of the future tense when the enclitic form of the verb *ht(j)eti* 'to want' comes after the main verb is indicated in (Bekavac et al., 2008) as a difference at the morphological level with examples:[6]

(1) HR: *posjetit ću*
SR: *posetiće*
BA: *ću posetiti*
(EN: to visit)

The same distinction is also emphasized in (Tiedemann and Ljubešić, 2012) in both forms of the future tense (enclitic before and after the main verb), noting that within the Serbian language this repre-

[3]The basic facts about the relationship that exists between Ekavian and Ijekavian pronunciation in Serbian language, as well as the complex relationships of Croatian dialects can be found in META-NET White Paper Series (Vitas et al., 2012), (Tadić et al., 2012).

[4]`http://www.sluzbenilist.ba/`

[5]For example, the derived forms in Croatian would be *vjerojatnost* 'probability' and *predsjedatelj* 'chairman', while in Serbian corresponding forms would be *v(j)erovatnoća* and *preds(j)edavajući*.

[6]In examples BA stands for Bosniak, HR for Croatian and SR for Serbian.

sents the synthetic form of future tense in contrast to the analytical form in Croatian and Bosniak with an example:

(2) HR and BA: *vidjet ću* and *ću vidjeti*
SR: *videću* and *ću videti*
(EN: I will see)

Let us note that the form of the future tense in these examples *comes from differences in orthography, and not in languages*:[7] in the Serbian language, this form of the future tense is written as pronounced, while in the Croatian language it shows its morphological composition.

2.3. Foreign names

It is underlined both in (Bekavac et al., 2008) and (Tiedemann and Ljubešić, 2012) that the difference in the writing of foreign proper names exists: while they are transliterated in Serbian language, they are usually not in Croatian. Let us mention that this difference that also stems from different orthography norms is indeed of importance, as shown in (Krstev et al., 2013), as named entity recognition systems developed for Serbian can not be applied with equal success to Croatian and vice versa.

2.4. Lexical differences

2.4.1. One point of view

Lexical differences between the three languages are noted in (Bekavac et al., 2008: p. 36) and a series of examples are cited, such as:[8]

(3) HR:*glede* SR:*u pogledu* BS:*u vezi*
(EN:on/of/about/regarding)

(4) HR:*s|sa* SR:*s* BS:*s|sa*
(EN:with)

Lexical differences are the main criterion for distinguishing Serbian and Croatian, but only a limited number of lexemes is indicative. Besides, they need to be real differences. E.g. the preposition *s|sa* 'with' has both forms in Serbian language as well, thus the motive for the exclusion of the form *sa* is not clear.

2.4.2. Another point of view

Some lexical differences are incorporated in the method used in (Tiedemann and Ljubešić, 2012), which proposes a list of 25 Bosniak, Croatian and Serbian words representing the strongest discriminators amongst these languages. However, within this list of discriminators the equivalent lexemes are *not* presented, *nor* their translation into English language. The list itself consists of grammatical forms of words, hence, in Bosniak the words *izvještajima*, *izvještaja* 'report' appear as discriminators, and in: *posete*, *posetio*, *poseti* 'to visit', instead of the lemmas *izvještaj* or *posetiti*.[9] Most of the differences that exist between the Bosniak and Serbian come down to the difference between Ekavian and Ijekavian pronunciation (e.g. Ekavian *izveštaj*, Ijekavian *izvještaj*) which, with respect to Section 2.1., cannot be considered discriminative difference.

If the discriminators are replaced with the corresponding lemmas, then these words lose the discriminatory function in each of the languages. Taking into account that the word order in Serbian and Croatian is free, it is possible, in general, to rephrase the sentence in which the discriminator appears into the sentence in which another form of the same word is used that does not have the discriminatory function.

Discriminators of the Croatian language consist primarily of Croatisms, such as *tjedan* 'week', *tvrtka* 'company', *ravnatelj* 'director', *gospodarstvo* 'economy' or the names of months of the year

[7] In (Silić and Pranjković, 2005: p. 9) it is emphasized that in the Croatian form of the future tense in the example (1), the letter *t* from the base of the main verb is not pronounced, i.e. that in the pronunciation the base and enclitic are pronounced as one unit, hence, as in Serbian language.

[8] *u vezi* can be a prepositional construction, but not necessarily, thus, it is not always in opposition to *glede* and *u pogledu*.

[9] By reducing forms to lemmas, 13 "discriminators" remain for the Bosnian and 20 for the Serbian language.

(of which 10 out of 12 are recorded). Let us note that the Croatian Frequency Dictionary (FRK) (Moguš et al., 1999), does not register occurrence of some discriminators (*glede* 'regarding', *izvješće* 'report', *priopćenje* 'statement'), and that the 25th discriminator for the Croatian language is the instrumental form of the singular noun *konac* '(a) thread; (b) end': *koncu*, which is a common noun for all three languages.

The arbitrariness of discriminators is shown on the example of the 21st discriminator for Serbian language: that is the word *ren* 'horseradish' (written in lowercase). The word appears even 724 times (or 0,018% of the total number of words), however, within the corpus, it *always* represents a transcribed name of the politician *Rehn* (in Serbian *Ren*). Not even this word is discriminator if corpora is searched by lemmas, and not by isolated forms, considering that the form of its vocative: *rene* appears in Croatian, which actually represents proper name *Rene* written without an accent (in names *René van der Linden*, *René Magritte*, etc.).

2.5. Complements of modal verbs

As for the differences in the syntactic level, the above-mentioned works emphasize the differences in terms of complements of modal verbs: the construction *modal verb + infinitive* is more common in Croatian language, while in Serbian the construction *modal verb + da* 'to' + *present* is more frequent. In (Tiedemann and Ljubešić, 2012) this difference is illustrated with the following example:

(5) SR: *hoću da radim*
HR: *hoću raditi*
(EN: I wish to work)

2.6. *s:da* 'with:to'

As the difference at the syntactic level it is indicated in (Bekavac et al., 2008) that the preposition *sa* 'with' in Croatian and Bosniak is in use where in Serbian *da*-construction is used, which is illustrated by the following example:[10]

(6) BS: *će prestati* ***s*** *korištenjem*
HR: *će prestati* ***s*** *uporabom*
SR: *će prestati* ***da*** *koriste*
(EN: to stop using)

With phase verbs (such as *početi* 'to start' or *nastaviti* 'to continue') two types of complements can be used in Serbian and Croatian — the verbal and the prepositional construction. For example,

(7) SR: *prestao je da piše*
HR: *prestao je s pisanjem*
(EN: to stop writing)

This is not a question of syntactic difference, but it is rather a case of an interesting example of promoting individual *choice of stylistic option* (which is a question of individual style of translator) into cross-language difference. Hence, it is entirely possible for a Serbian author to write *prestati s korišćenjem*, as well as for a Croatian writer to use *prestati da koristi/koristiti*.

3. Formal shortcomings in SETimes-corpus

The corpus of texts from the website *SETimes* has formal deficiencies. First of all, translations into Serbian, Croatian and Bosniak in the respective corpora were not signed, thus the number of translators who participated in the translation process remained unknown, we do not even know if they were native speakers of Serbian, Croatian and Bosniak, nor whether the translators were required to follow specific guidelines as to ensure differentiation of languages through translations. Note in this regard was also given in (Tiedemann and Ljubešić, 2012: p. 2631) indicating that the observed differences are *not* "actual differences in language use or language norm".

[10] Let us note that within the example (6) the form of the future tense (underlined) is the same in all three languages, which is opposite to the difference indicated in (Bekavac et al., 2008) and cited in the example (1).

Neither ST-sr nor ST-hr were compactly encoded in Latin Extended-A, but instead contain characters from other code pages such as, for example, Greek and Cyrillic glyph *A*. Only the Cyrillic character *j* (Ǌ) occurs in ST-sr 1288 times, and in ST-hr 1231 times. As these characters represent separators of words when processing the corpora, their appearance changes the distribution of frequencies even with high-frequency words.

Signatures of pictures were not removed from the corpora: sequence [Getty Images] or, in transcribed form, [Geti Imidžis], appears 2809 times in ST-hr, and 2452 times in ST-sr.

Sequences identifying correspondents were not removed from the corpora, thus the sequence with the structure:

<proper name> + *for Southeast European Times from* + <toponym> — <date>

covers nearly 1% of tokens in each of the corpus.

Determining differences, based on the corpora of *SETimes*, indicates, primarily, that the differences are difficult to determine. Some of the observed distinctions are in fact orthographic or stylistic differences, rather than differences between languages, and some of the distinctions stem from unrepresentativeness of the corpus. The quantification of the observed differences was not given in the above-mentioned descriptions, hence we cannot determine their statistical relevance.

4. Suggested differences from the point of view of other corpora

4.1. The used corpora

	ST-sr	**ST-hr**	**L-sr**	**L-hr**	**H-ek**	**H-hr**	**H-msc**
Tokens	8,945,968	9,040,646	2,676,546	2,639,495	705,819	550,341	684,219
Words	3,940,296	3,891,179	1,157,857	1,146,467	304,324	238,797	298,683

Table 1: The size of used corpora.

In order to examine the relationship of languages presented in ST-corpora according to the official standards and usages of language, we compared the frequency distribution of these differences for the Serbian and Croatian languages on the ST-corpora presented in Section 2. with the corresponding distributions in other sources for Serbian and Croatian. For comparison, we used the so-called Henning's corpus[11] of literary works of writers who wrote at the time of Serbo-Croatian language, which we divided into three sub-corpuses: H-ek — works with Ekavian pronunciation, H-hr — works by Croatian authors with Ijekavian pronunciation and H-msc – works of non-Croatian authors with Ijekavian pronunciation.[12] We also used the corpus of literary works that have been translated (mainly) from English to Serbian (L-sr) and Croatian (L-hr).[13] These translations were created independently and mostly after the disintegration of Yugoslavia, translated by prominent literary translators, and published several times in high circulation. Dimensions of these corpora, including both ST-sr and ST-hr, are presented in Table 1.

	ST-sr	**ST-hr**	**L-sr**	**L-hr**	**H-ek**	**H-hr**	**H-msc**
Form (a) with insertions	0.564	0.552	0.336	0.407	0.205	0.272	0.239
Form (b)	0.212	0.186	0.141	0.101	0.161	0.025	0.155

Table 2: Distribution of two forms of the future tense in different corpora.

[11] This corpus of the early '90s is integrated into the web page `http://www.borut.com/library/index.htm` (May 7, 2015). It should be noted that some authors represented in this corpus are listed in the required reading for Croatian schools even today.

[12] Classification into corpora H-hr and H-msc was done according to criteria presented in 6.

[13] The corpus is described in (Vitas, 2014).

In addition to these corpora, we compared some differences with the Corpus of Contemporary Serbian language (SrpKor),[14] the Croatian National Corpus (HNK) from 2003,[15] with the data from the Croatian Frequency Dictionary (FRK) and the corpora that was used (Tiedemann and Ljubešić, 2012) for evaluation (PO, VL, DA).[16]

4.2. Future tense

The future tense is formed in two ways: either (a) as in the example (2) from the present tense of the verb *ht(j)eti* and the infinitive of the verb or (b) as in the example (1) by adding the enclitic of the verb *ht(j)eti* onto the form of the verb, either as univerbal (Serbian version) (Stanojčić and Popović, 2014) or non-univerbal form (Croatian version) (Silić and Pranjković, 2005). In the case (a) strings of words can be inserted between the enclitic and infinitive.

Simple lexical patterns allow modelling these forms of the future tense by using appropriate morphological dictionaries, thus obtaining the information about its relative frequency in the mentioned corpora presented in Table 2.

These data contradict the assertion that the form (b) of the future tense is more common in Croatian than the form (a), as indicated in (Bekavac et al., 2008). On the other hand, in (Tiedemann and Ljubešić, 2012) the difference in the form (b) is considered to be the main morphological difference; however, its frequency is very low.

4.3. Lexical differences

	ST-sr	**ST-hr**	**L-sr**	**L-hr**	**H-ek**	**H-hr**	**H-msc**	**SrpKor**	**HNK**
SR: *u pogledu*	0.047	0.002	0.003	0	0	0	0	0.006	0.002
BA: *u vezi*	0.021	0.004	0.006	0.003	0	0.006	0.002	0.016	0.005
HR: *glede*	0	0.058	0	0.003	0	0	0	0	?

Table 3: Frequencies of prepositions *u pogledu*, *u vezi* and *glede* in different corpora.

From the sample of the lexical difference in the example (3) we obtained the frequency of their use presented in Table 3.[17] Hence, outside of the *SETimes-corpus*, the dominant form is *u vezi* 'in connection, in relation'. The "Bosniak" form *u vezi* 'regarding, in terms of' is used more often in Serbian than the "Serbian" *u pogledu*, whereas the form *glede*, which is mentioned a strict discriminator in (Tiedemann and Ljubešić, 2012), is rather rare in other Croatian corpora. Moreover, preposition *glede* has not been recorded in FRK.

	ST-sr	**ST-hr**	**L-sr**	**L-hr**	**H-ek**	**H-hr**	**H-msc**
HR-25:	0	0.869	0	0.054	0	0.015	0.005
SR-ek-25:	0.825	0	0.096	0.001	0.98	0	0
SR-ijek-25:	0.216	0.27	0.001	0.064	0	0.087	0,044

Table 4: Frequencies of 25 discriminators.

Let us look at the distribution of the afore-mentioned 25 strongest discriminators. In addition to the frequency of discriminators for Croatian (HR-25), in Table 4 we also list the frequency of discriminators for Serbian both in their Ekavian (SR-ek) and Ijekavian (SR-ijek) form. What is interesting is that Ijekavian forms of Serbian discriminators have a significant number of occurrences in all Croatian corpora, which confirms the noticed deficiency of SETimes corpus in Section 2.1..

[14] `http://www.korpus.matf.bg.ac.rs/korpus/`

[15] According to `https://web.archive.org/web/20030207180909/http://www.hnk.ffzg.hr/korpus.htm` from March 30, 2003, the Croatian National Corpus contained 9,156,446 words. This web page provides a list of bigrams with a frequency above 100.

[16] Designation PO is for the Serbian daily *Politika*, VL for the Croatian *Večernji list* and DA for the Bosnian *Dnevni Avaz*.

[17] The frequency *glede* is not available in the specified source for the HNK, and it does not appear within the list of FRK.

4.4. The relationship *s:sa* 'with'

Preposition *s/sa* 'with' is listed in Subsections 2.1. (Example 4) and 2.6. (Example 6). The distribution of the forms *s* and *sa* is presented in Table 5.

The participation of the forms *s* and *sa* in ST-sr indicates a serious difference in relation to other corpora. Moreover, there are 1,868 occurrences of the preposition *s* in ST-sr, 86% in the expression *s obzirom* 'with respect to', as opposed to only 647 appearances of this expression in the ST-hr. A number of occurrences of the preposition *s* corresponds to expressions *s vremena na vreme* 'from time to time' (16), *s leva* 'from left' (45) and *s desna* 'from right' (45), therefore over 90% of the occurrences of this preposition is related to only four multi-word expressions. Within the ST-hr, more than 95% of appearances of the preposition *sa* is subject to the rule described in (Barić et al., 2003), that the next word after the form *sa* must begin with some of the following letters *s*, *š*, *z*, *ž*. This rule is consistently applied in other Croatian corpora except where the next word begins with the consonant cluster (e.g. *sa mnom* 'with me', *sa psom* 'with a dog', *sa dna* 'from the bottom', etc.). Ijekavian non-Croatian corpora (H-msc, DA) already deviate from this rule, while in contemporary Serbian Ekavian copora the limitations in terms of the use of the preposition *s/sa* are less strict, as indicated in (Piper and Klajn, 2014).

s/sa	**ST-sr**	**ST-hr**	**L-sr**	**L-hr**	**H-ek**	**H-hr**	**H-msc**	**SrpKor**	**FRK**	**PO**	**VL**	**DA**
f(s)	0.047	0.71	0.245	0.587	0.371	0.608	0.6	0.148	0.562	0.176	0.61	0.436
rank	367	11	39	15	23	13	17	40	20?	49	12	16
f(sa)	0.857	0.185	0.536	0.155	0.488	0.207	0.188	0.639	0.2	0.652	0.151	0.293
rank	9	32	15	56	15	42	49	10	40?	11	48	29
f(s)/f(sa)	0.055	3.83	0.46	3.79	0.76	2.94	3.18	0.23	2.84	0.27	4.03	1.49

Table 5: Frequencies and ranks of the preposition *s/sa* in different corpora

4.5. The conjunction *da* 'to'

	ST-sr	**ST-hr**	**L-sr**	**L-hr**	**H-ek**	**H-hr**	**H-msc**	**SrpKor**	**FRK**	**PO**	**VL**	**DA**
f	2.955	0.65	3.1	1.91	2.551	1.933	2.74	2.67	1.50	3.1	1.795	2.527
rank	4	12	3	5	4	5	3	4	4	4	4	4

Table 6: Frequency and rank of the conjunction *da* in different corpora

The conjunction *da* 'to' is the subject of the differences described in Sections 2.5. and 2.6.. It is extremely frequent and common for the entire Shtokavian area. The Table 6 indicates its relative frequency and ranking. For the sake of comparison, the data were added from the Corpus of Contemporary Serbian language (SrpKor), according to (Utvić, 2014), and the Croatian Corpus (HrvKor), then from the Croatian Frequency Dictionary (Moguš et al., 1999), as well as from control corpora used in (Tiedemann and Ljubešić, 2012). Also, the conjunction *da* has the rank 5 in the study (Škiljan, 1980).

The drop of the conjunction *da* to the 12th place in ST-hr compared to other corpora illustrates the serious anomaly in its use within this corpus. This is even more visible in the Table 7 that lists the ranking of the most frequent bigrams with *da* in corpora from Table 6.

	ST-sr	**ST-hr**	**L-sr**	**L-hr**	**H-ek**	**H-hr**	**H-msc**	**SrpKor**	**FRK**	**PO**	**VL**	**DA**
da se	1	**20**	1	2	1	1	1	1	2	1	2	2
da je	4	**12**	3	5	4	5	3	4	4	4	4	4
i da	21	**262**	11	26	8	13	14	3	32	3	11	10
da će	4	**38**	20	17	24	20	41	5	13	4	10	9
je da	6	**86**	12	35	13	53	35	7	18	8	9	3
da su	14	**69**	16	18	25	26	38	8	9	6	6	7
da bi	8	**95**	21	32	20	40	61	15	17	16	25	27

Table 7: Ranks of seven most frequent bigrams in different corpora

phenomenon	C_1	C_2	χ^2 value	p-value
Future tense: form (a) with insertions	All-hr	ST-hr	576.7842	< 0.001
Future tense: form (b)	All-hr	ST-hr	667.9133	< 0.001
mod da P	All-hr	ST-hr	547.3367	< 0.001
mod inf	All-hr	ST-hr	39.8918	< 0.001
mod da P	All-sr	ST-sr	11340.44	< 0.001
mod inf	All-sr	ST-sr	762.9576	< 0.001
HR-25	All-hr	ST-hr	10636.51	< 0.001
s/sa	All-hr	All-sr	8.8605	< 0.01

Table 8: Comparision of frequencies of observed phenomena; *mod da P* stands for the modal verb followed by a conjunction *da* and a verb in the present tense, while *mod inf* stands for the modal verb followed by an infinitive.

5. Concluding analysis

As experts in corpus linguistics state, a comparison of corpora and a corpora similarity assessment is a complex and multi-dimensional task (Kilgarriff, 2001). The analyses we performed here follow general principles as summarized in (Baroni and Evert, 2008) and tend to examine the distribution differences of phenomena of interest among pairs of Serbian and Croatian corpora.

In order to calculate the distributions we worked with two large corpora. The first one groups together all Croatian corpora (L-hr, H-hr, HNK, and VL, further denoted as All-hr) while the second one encompasses all available Serbian corpora (L-sr, H-ek, SrpKor, and Pol, further denoted as All-sr). The cumulative frequencies of all phenomena of interest with the respect to these corpora are compared to the frequencies from ST-hr and ST-sr corpora. The comparison is based on a χ^2 distribution test with one degree of freedom (Agresti, 2002) and computed with software package R. Table 8 presents obtained results. We did some additional exploration of confidence intervals not presented in the table to double check the significance of obtained results as the large samples may lead to highly significant p-value for minimal and irrelevant differences (Baroni and Evert, 2008).

The obtained results are statistically significant with 0.001 significance level or level 0.01 (*s/sa* example with p-value=0.002914) and therefore can confirm the deviation among ST-corpora and other corpora when the distribution of listed phenomenon comes into a question.

6. The real discriminatory differences — an example

The distribution of frequencies in the corpus composed of material from the website SETtimes indicates serious anomalies, as shown in Sections 4. and 5., thus making it unsuitable for any kind of comparison between the Serbian and Croatian standard language. Bearing in mind the relationship between Serbian and Croatian norms, it is necessary to find stable and sufficiently frequent **linguistic** differences on the basis of which it will be possible to make an **objective** identification of the language even on the level of short texts.

	ST-sr	ST-hr	L-sr	L-hr	H-ek	H-hr	H-msc	SrpKor	FRK	PO	VL	DA
f(T)	0	0.034	0	0.18	0	0.197	0	0	0.128	0	0.084	0.002
f(K)	0.044	0.007	0.215	0.055	0.280	0.045	0.279	0.133	0.078	0.150	0.017	0.163

Table 9: Distribution of pronouns *tko* and *ko* and their derivatives in different corpora

The interrogative pronoun *who* provides one linguistic criterion that distinguishes the Croatian standard from all other Neo-Shtokavian standards. This difference is not a matter of individual lexeme, but it rather relates to the system of pronominal words. Croatian standard encodes, both in written as well as in oral standard, an older *form of the nominative* of this pronoun *tko*, unlike other languages where its form is *ko*. As such, this pronoun is cited in both the Croatian Orthography (Jozić et al., 2013), as well

as in the Dictionary of Croatian (Anić, 1998) and Croatian grammars (Silić and Pranjković, 2005) and (Barić et al., 2003). Prefixes and suffixes are added to the form of the nominative of this pronoun to give indefinites and negatives, hence they can all be presented within the following expression:[18]

(T) **tko**|gdje**tko**|pogdje**tko**|i**tko**|
koje**tko**|ne**tko**|pone**tko**|ni**tko**|
sva**tko**|malo**tko**|što**tko**|**tko**god

opposite to the equivalent forms used in other languages emerged from the former Serbo-Croatian language:

(K) **ko**|gd(j)e**ko**|pogd(j)e**ko**|i**ko**|
koje**ko**|ne**ko**|pone**ko**|ni**ko**|
sva**ko**|malo**ko**|**ko**god

The distribution of these expressions in the observed corpora is given in Table 9. The frequency of the expression (K) in the Croatian corpora comes from the fact that the following forms are observed: *neko* and *svako* as adjective pronouns, proper name *Niko*, conjuction *kao* in the form *ko*, but not the nominal pronoun *ko*. From this stems the fact that the appearance of the words from the expression (T) in a particular text with a frequency greater than a threshold, e.g. 0.01%, absolutely identifies it as the text in Croatian language.

7. Conclusion

The described shortcomings of the corpora composed of texts from the website SETimes lead to the conclusion that this corpus does not represent adequately neither the Serbian nor the Croatian standard language. Results obtained by exploitation of this corpus, therefore, cannot be accepted as relevant to neither of two languages. It is necessary to develop a parallel corpus of Serbian and Croatian that would better represent both in size and its content the standards of the two languages as well as their usage. From such a corpus it would be possible to determine with more confidence the real differences between two languages.

Acknowledgment

This research was partly supported by the Serbian Ministry of Education and Science under the grant 178006.

References

Agić, Ž., Ljubešić, N., and Merkler, D. (2013). Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 48–57, Sofia, Bulgaria, August. Association for Computational Linguistics.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley and Sons.

Anić, V. (1998). *Rječnik hrvatskoga jezika [Dictionary of Croatian]*. Novi Liber.

Barić, E., Lončarević, M., Malić, D., Plavešić, S., Peti, M., Zečević, V., and Zinka, M. (2003). *Hrvatska gramatika [Croatian Grammar]*. Školska knjiga.

Baroni, M. and Evert, S. (2008). Statistical Methods for Corpus Exploitation. *Corpus Linguistics. An International Handbook, Mouton de Gruyter, Berlin*, pages 777–803.

Bekavac, B., Seljan, S., and Simeon, I. (2008). Corpus-based Comparison of Contemporary Croatian, Serbian and Bosnian. In *Formal Approaches to South Slavic and Balkan Languages FASSBL*, pages 33–39. Hrvatska znanstvena bibliografija i MZOS-Svibor.

[18] The form *ponetko* was confirmed in the Croatian corpus `http://riznica.ihjj.hr/`, but not in the Ortography.

Jozić, Ž., Bartolec, G. B., Hudeček, L., Lewis, K., Mihaljević, M., Ramadanović, E., Birtić, M., Budja, J., Kovačević, B., Ivanković, I. M., Milković, A., Miloš, I., Stojanov, T., and Despot, K. Š. (2013). *Hrvatski pravopis [Croatian Orthography]*. Institut za hrvatski jezik i jezikoslovlje.

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Krstev, C., Zečević, A., Vitas, D., and Kyriacopoulou, T. (2013). NERosetta—an Insight into Named Entity Tagging. In Vetulani, Z. and Uszkoreit, H., Eds., *Proceedings of 6th Language & Technology Conference*, pages 168–172.

Moguš, M., Bratanić, M., and Tadić, M. (1999). *Hrvatski čestotni rječnik [Croatian Frequency Dictionary]*. Školska knjiga.

Piper, P. and Klajn, I. (2014). *Normativna gramatika srpskog jezika [Normative Grammar of Serbian]*. Matica srpska.

Popović, M. and Ljubešić, N. (2014). Exploring Cross-Language Statistical Machine Translation for Closely Related South Slavic Languages. In *LT4CloseLang 2014, EMNLP 2014*, pages 76–84.

Popović, L. (2004). From Standard Serbian Through Serbo-Croatian to Standard Serbian. In Bugarski, R. and Hawkesworth, C., Eds., *Language in the Former Yugoslav Lands*, pages 25–40. Slavica Pub.

Rehm, G. and Uszkoreit, H., Eds. (2012). *META-NET White Paper Series*. Springer. Available online at `http://www.meta-net.eu/whitepapers`.

Silić, J. and Pranjković, I. (2005). *Gramatika hrvatskoga jezika [Croatian Language Grammar]*. Školska knjiga.

Stanojčić, Ž. and Popović, L. (2014). *Gramatika srpskog jezika [Serbian Language Grammar]*. Zavod za udžbenike i nastavna sredstva.

Tadić, M., Brozović-Rončević, D., and Kapetanović, A. (2012). *Hrvatski Jezik u Digitalnom Dobu – The Croatian Language in the Digital Age*. In Rehm and Uszkoreit (Rehm and Uszkoreit, 2012). Available online at `http://www.meta-net.eu/whitepapers`.

Tiedemann, J. and Ljubešić, N. (2012). Efficient Discrimination Between Closely Related Languages. In *COLING 2012*, pages 2619–2634.

Tyers, F. M. and Alperen, M. S. (2010). South-East European Times: A Parallel Corpus of Balkan Languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53.

Utvić, M. (2014). Liste učestanosti Korpusa savremenog srpskog jezika [Frequency lists of The Corpus of Contemporary Serbian]. *Naučni sastanak slavista u Vukove dane*, 43(3):241–262.

Vitas, D., Popović, L., Krstev, C., Obradović, I., Pavlović-Lažetić, G., and Stanojević, M. (2012). *Srpski jezik u digitalnom dobu – The Serbian Language in the Digital Age*. In Rehm and Uszkoreit (Rehm and Uszkoreit, 2012). Available online at `http://www.meta-net.eu/whitepapers`.

Vitas, D. (2014). O različitosti sličnog [On the Differences of Similar]. *Naučni sastanak slavista u Vukove dane*, 43(3):31–49.

Škiljan, D. (1980). *Lingvističko istraživanje Večernjeg lista [Linguistic Research of Večernji list]*. RO Vjesnik.

Zečević, A. and Vujičić-Stanković, S. (2013). The Mysterious Letter *J*. In *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants*, pages 40–44, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

‘While’ and ‘Until’ Clauses and Expletive Negation in a Corpus of Bulgarian and Ukrainian Parallel Texts

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

Olena Siruk
Institute of Philology, Taras Shevchenko National University of Kyiv; IMI—BAS
olebosi@gmail.com

Abstract

The combination of the meanings ‘while’ and ‘until’ in a single lexeme and the use of expletive negation with the latter meaning are widespread phenomena that are a rich source of research problems. In this paper we present a comparative bilingual Bulgarian and Ukrainian corpus-based study of several conjunctions that share these two meanings. We discuss the difference in the frequency of expletive negation in the two languages, the use of *až* ‘even, all the way’ in Ukrainian and the impact of the original language in translated texts.

1. Introduction

The combination of the meanings ‘while’ and ‘until’ in a single lexeme and the obligatory or optional use of expletive negation[1] with the latter meaning are widespread phenomena (found in the Slavic languages, Hindi, Hungarian, Italian, Ket, Persian, etc.) that are a rich source of research problems, due to the high level of crosslinguistic and diachronic variation and the complex interaction of a variety of criteria, which makes it hard to obtain unambiguous judgements.

This paper presents a comparative bilingual Bulgarian and Ukrainian corpus-based study of several conjunctions that share these two meanings, with focus on the use of expletive negation.

The working Bulgarian–Ukrainian parallel corpus is composed entirely of fiction, including both original Bulgarian and Ukrainian texts and translations from other languages. The current version, which contains 100 texts (mostly novels, but occasionally parts of large novels, as well as collections of short stories by the same author and, if translated, by the same translator), is made of ten sectors, each composed of texts with the same original language and measuring approximately 800,000 words on the Bulgarian and 700,000 words on the Ukrainian side. This amounts to an approximate total of 15 million words in the entire corpus. Two sectors contain translations from Russian and two from English (because of the larger amount of material available); the remaining original languages are Bulgarian, French, German, Italian, Polish and Ukrainian. All texts are aligned at sentence level.

2. The Experiment

The lexical items studied in this experiment were conjunctions with the meaning ‘while, until’, especially such as allow expletive negation when in the latter meaning. In Bulgarian these terms are *dokato*, *dokle*, *dokogato* and *do(r)de(to)*.[2] In Ukrainian they are *doky*, *dopoky*, *zaky*, *pokil’* and *poky*; in addition, the frequent combinations *až doky*, *až poky* ‘all the way until’ were treated as separate items, as was the particle *až* ‘even, all the way’ when it functions as a conjunction all by itself. All

[1] Also called pleonastic or paratactic negation, as opposed to semantic negation.

[2] We use the 1898 scientific transliteration system that is predominant in international linguistic publications on Cyrillic-written Slavic languages (Transliteration, 1898) for both Bulgarian and Ukrainian.

pairs of sentences or sentence fragments containing one of these items on one or both sides were located and counted.[3] A total of 8809 such pairs were found in the corpus, including:

- 3446 pairs of *bona fide* matches,
- 3873 occurrences of unmatched Bulgarian 'while/until' items; among them are 843 that feature adverbial participles on the Ukrainian side[4], 549 the temporal conjunction *koly* 'when' and 234 one of the compound conjunctions *v/u toj čas jak*, *tymčasom jak* and *todi jak* 'while, whereas',
- 1406 occurrences of unmatched Ukrainian 'while/until' items; among them 282 employ the conjunction *predi da* 'before' on the Bulgarian side and 202 have a verb of waiting in the matrix clause with an 'until' clause in Ukrainian and a *da*-clause in Bulgarian,
- and 84 pairs of sentences in which both sides contain one of the studied items, but the meaning is substantially different.

Among the Bulgarian items *dokato* dominates absolutely (6852 occurrences, or 92.56%). A distant second is *dodeto* (510 times, or 6.89%), which only appears in 29 of the 100 texts, with varying frequency, and only outnumbers *dokato* in the writing of one author (Bogomil Rainov); the only translations where *dodeto* has a tangible presence are JRR Tolkien's *Narn i Hîn Húrin*, translated by Lyubomir Nikolov, and Stendhal's *Red and Black*, translated by Atanas Dalchev.[5] Among the Ukrainian items the most common one is *poky* (3597 occurrences, or 72.87%), followed by *až poky* (587), *doky* (572), and *až doky* (103); *(až) doky* outnumbers *(až) poky* in only 7 texts out of 100.[6]

3. Semantics, Polarity and Aspectuality

The 'while/until' words in both languages combine several related meanings, which correlate, albeit not perfectly, with the aspectuality of the eventualities in both clauses and the polarity of the subordinate clause. Telicity, in turn, correlates to a high degree with Slavic aspect: imperfective and perfective verbs usually denote atelic and telic predicates, respectively.

		affirmative subordinate clause (Q)	
		atelic	**telic**
main clause (P)	**atelic**	P is happening *while* Q is happening	P is happening *until* Q happens
	telic	P happens *while* Q is happening	P happens *by the time* (*before*) Q happens

Table 1: The impact of aspect with an affirmative subordinate clause

A common exception to the correlation between aspect and telicity is a present tense form of an imperfective verb used for a telic event in vivid narration (1). Another is a perfective verb denoting its aftermath state (2):

(1) Bg: *[...] i otnovo vărvim, dodeto si davam smetka, če tova ne e nikakva ulica, a njakakăv pust meždu*selski *păt.*
Uk: *[...] i os' my znovu jdemo, poky ja usvidomljuju, ščo ce zovsim ne vulycja, a jakyjs' bezljudnyj sil's'kyj putivec'.*

[3] Except where the term has a discernibly different meaning, as is the case with Ukrainian *doky* 'until when? how long?' and *poky* 'for now, for the time being'. Contrariwise, the atemporal sense 'whereas' of Bulgarian *dokato* is hard to separate from the temporal one, so such occurrences were retained.

[4] The frequency of this translation correspondence is discussed by Martinova-Ivanova (2015).

[5] Notably, however, *dodeto* only appears twice in Nikolov's translation of *The Hobbit* by JRR Tolkien and not at all in Dalchev's translation of *The Gold Chain* by Alexander Grin, which demonstrates that such lexical preferences need not be a mark of the translator's personal style, but instead may reflect his approach to the genre of the particular text, along with the fact that *dodeto* has come to be perceived as somewhat archaic.

[6] The most pronounced preference for *(až) doky* is in Vasyl Zemliak's *Green Mills* (71 occurrences, as opposed to only 4 of *(až) poky*); in *The Swan Flock* by the same author, however, we find *(až) poky* 43 times and *(až) doky* 39 times.

'[…] and we're on our way again, till I realise that this is not a street at all, but some deserted country road.'[7]

(Bogomil Rainov, *The Day Doesn't Look Like the Morning*)

(2) Bg: *Ami toj šte izleze navăn, dokato se săbličaš…*
Uk: *Ta vin vyjde, poky ty rozdjahatymešsja…*
'Well, he'll go out [and stay outside] while you're undressing …'

(Pavel Vezhinov, *Libra*)

When the subordinate clause is negative, a third aspectual category comes into consideration, *viz.* consequent states of events, expressed in Bulgarian by perfect or pluperfect tense forms of perfective verbs but behaving as atelic predicates. In Ukrainian, which has replaced the aorist by the perfect, they assume the same form as telic predicates, although they can often be identified by the adverb *šče* 'still, yet'.

(3) Bg: *Iskam da ti obadja nešto, Džo, dokato ogănjat ne e ugasnal.*
Uk: *Ja xotiv by ščos' tobi skazaty, Džo, poky šče vohon' ne zhas.*
"Before the fire goes out [*in the translations:* while the fire hasn't gone out (yet)], Joe, I should like to tell you something."

(Charles Dickens, *Great Expectations*)

		negative subordinate clause (Q)		
		atelic	**telic**	**perfect**
main clause (P)	**atelic**	P is happening *while* Q is not happening	P is happening *until* Q happens (expletive negation)	P is happening *while* Q has not happened
	telic	P happens *while* Q is not happening		P happens *while* Q has not happened

Table 2: The impact of aspect with a negative subordinate clause

Negation in the subordinate clause is semantic if the predicate is atelic or perfect. With a telic predicate, as a rule, the negation is expletive. Exceptions, when the failure of a scheduled or recurrent event to happen is considered an event in its own right, are rare and potentially ambiguous.

(4) Bg: *Taka životăt prodălžavaše da si teče, dokato edin den kăm sredata na januari srebristoto konče i ezdačkata mu ne se javiha na ugovorenata srešta.*
Uk: *Tak vono use tryvalo doty, doky odnoho dnja v seredyni sičnja sribljasto-čala kobylka ta jiji veršnycja v domovlenyj čas ne z"javylysja.*
"So matters went on, till one day in the middle of January the silver-roan palfrey and its rider were missing [*in the translations:* did not show up] at the tryst."

(John Galsworthy, *The Forsyte Saga*)

The semantic kinship between '*while* Q is happening' and '*until* Q happens' on the one hand, and '*until* Q happens' and '*while* Q has not happened' on the other, provides a rationale for the optional use of the expletive negation (Barentsen, 1979).

The construction may further imply that P terminates either *no sooner* than Q or *no later* than Q. The former meaning may be emphasised by Bg *pone*, Uk *prynajmni*, *xoč(a)* 'at least' and Bg *čak*, Uk *až* 'all the way'; the latter, by Bg *samo*, Uk *lyše*, *til'ky* 'only'. Expletive negation also indicates

[7] English glosses are given in single quotes if they are ours, and in double quotes if they come from originals or published translations.

that P terminates *no sooner* than Q, but does so less strongly, and therefore can co-occur with the 'only' adverbs, although there no examples of this in our corpus.[8]

4. The Results

The corresponding constructions where a 'while/until' word is used in both languages are shown in Table 3. The rows correspond to the Bulgarian side and the columns to the Ukrainian one; ∂ stands for any 'while/until' word (except for Ukrainian *až*, which is represented and counted separately), → for an atelic predicate, ☼ for a telic one and ☼→ for a resultative form (in Bulgarian only). The cell in the lower right corner (except for the totals) sums up nine occasions on which a sentence is interrupted after a 'while/until' word.

	∂ →	***∂ ☼***	***až ∂ ☼***	***až ☼***	***až ∂ ne ☼***	***∂ ne ☼***	***∂ ne →***	***∂ …***	∑
∂ →	1285	18	4			1			1308
∂ ☼	47	588	228	35	83	318			1299
∂ ne ☼	1	36	76	9	133	449	1		705
∂ ne ☼→		1				76	4		81
∂ ne →						4	40		44
∂ …								9	9
∑	1333	643	308	44	216	848	45	9	3446

Table 3: Correspondences between 'while/until' structures

Sentences with 'while' clauses account for 40% of all. Bulgarian 'until' clauses correspond to Ukrainian 'while' clauses twice more often than the other way around.

Ihe table makes it evident that Bulgarian uses expletive negation in 'until' clauses more sparingly than Ukrainian does (705 versus 216+848=1064 times). Where the Ukrainian does not employ it, the conjunction is twice more likely to be *až doky/poky* than simply *doky/poky* if there is expletive negation in the Bulgarian, but far less likely otherwise, which confirms the notion that the functions of *až* and expletive negation are related.

Among the 588 pairs of sentences with 'until' clauses in which there is no expletive negation in either language and no *až* in Ukrainian, there are 95 sentence pairs which state that P manages to happen by the time Q does (5) and 55 in which P measures the time until Q happens (6).

(5) Bg: *A dokato se vărnete, šte si pogovorim za nešto seriozno s mis Džejn.*
Uk: *Poky vy povernetes', ja dam mis Džejn dejaki nastanovy.*
"I'll give Miss Jane a lecture till you come back."

(Charlotte Brontë, *Jane Eyre*)

(6) Bg: *Imame petnajset minuti, dokato Doktorăt prebroi negovite krăvni telca.*
Uk: *U nas je xvylyn p"jatnadcjat' času, poky Likar poličyt' joho erytrocyty.*
"We have fifteen minutes while [*in the translations:* until] the Doctor counts corpuscles."

(Stanisław Lem, *Eden*; English by Marc E. Heine)

[8] Examples from other sources: *Ti kaza, če edin loš šof'or e v bezopasnost samo dokato ne sreštne drug loš šof'or, nali?* "You said a bad driver was only safe until she met another bad driver?" (F. Scott Fitzgerald, The Great Gatsby, translated by Neli Dospevska); *Tezi razsăždenija vărvjat mnogo gladko i strojno, no samo dokato njakomu ne hrumne da zapita: „A otkăde vsăštnost se e vzel tozi Praatom?"* 'All this can be calculated very accurately and handsomely, but only until someone gets the idea of asking: "And just where did this Proto-Atom come from?"' (Stanisław Lem, *The Star Diaries*, translated by Lina Vasileva).

This group also includes all sentences in which the matrix clause is a polar question; the 'no sooner' meaning conveyed by expletive negation and by Ukrainian *až* appears incompatible with interrogation (Derzhanski, 1999).

(7) Bg: *Bi li se săglasil da vzemeš čantata mi, dodeto minem prez mitnicata?*
Uk: *Čy ne pohodyšsja ty vzjaty moju sumku, poky my projdemo mytnycju?*
'Would you agree to take my bag until we go through customs?'
(Bogomil Rainov, *There Is Nothing Better than Bad Weather*)

If duration is stated, these devices are seldom used.

(8) Bg: *Toj tičal cjala nošt, dokato stignal do pešterata.*
Uk: *I vin bih usju nič, poky distavsja do pečery.*
"So he ran all the night till he came to the cave;"
(Rudyard Kipling, *The Jungle Book*)

(9) Bg: *A posle prodălži i tuk, ne čas i ne dva, a celi tri dni, dodeto ne naučiha i majčinoto mi mljako.*
Uk: *A potim tryvav i tut, ta ne hodynu i ne dvi, a cilyx try dni, až poky mene ne vypatraly do ostann'oji kryxty.*
'And then it [sc. the interrogation] continued here as well, not for an hour or two but for three whole days, until they got to know all my ins and outs.'
(Bogomil Rainov, *The Great Boredom*)

(10) Bg: *Tazi borba prodălži polovin minuta, dokato cveteto ne se predade i ne uvisna bezžizneno v răkata na Pavliš.*
Uk: *Cja borot'ba tryvala pivxvylyny, poky kvitka ne zdalasja i neruxomo povysla v ruci Pavlyša.*
'This struggle lasted for half a minute, until the flower gave up and hung limply in Pavlysh's hand.'
(Kir Bulychev, *Village*)

Furthermore, expletive negation is hardly used in Bulgarian with a verb of waiting as the matrix predicate; there are only six examples of this, all of them in translations from other Slavic languages. In Ukrainian it appears 18 times.

(11) Bg: *Šte se pribere otnovo v svoja Lubni i šte čaka mirno, dokato pronizitelnite trăbi na Gradiv ne go prizovat otnovo kăm podvizi…*
Uk: *Osjade u svojix Lubnax i čekatyme tyxo, až poky pronyzlyvi surmy Hradyvusa znov poklyčut' joho…*
"He will settle again in Lubni, and will wait quietly till the terrible trumpets call him to action again."
(Henrik Sienkiewicz, *With Fire and Sword*; English by Jeremiah Curtin)

(12) Bg: *[…] i tăj kato po mosta vărvjal goljam kervan natovareni muleta i kone, naložilo im se da počakat, dokato kervanăt se iztoči.*
Uk: *[…] ale po tomu mostu proxodyv same velykyj karavan nav"jučenyx muliv ta konej, i jim dovelos' počekaty, poky vsi vony na toj bik ne perexopljat'sja.*
"[…] and a caravan of pack-mules and sumpter-horses being in act to pass, it behoved them tarry till such time as these should be crossed over."
(Boccaccio, *The Decameron*; English by John Payne)

In many cases, however, expletive negation has little impact in sentences which state that P lasts until Q. Because of this, sentences that denote very similar situations often differ only in its presence or absence.

(13) Bg: *Svăržete me vednaga s Kiril Andreev, zvănete, dokato otgovori!*
Uk: *Z"jednajte mene zaraz z Kyrylom Andrejevym, dzvonit', doky ne vidpovist'.*
'Put me through to Kiril Andreev, ring until he answers!'
(Pavel Vezhinov, *Traces Remain*)

(14) Bg: *I gi izpălnjavaj, dokato toj ne te naznači za istinski gotvač.*
Uk: *I vykonuj, poky vin postavyt' tebe spravžnim axči.*
'And fulfil them [*sc.* the commissions of the administrator of the Sultan's kitchens] until he appoints you a regular cook.'
(Pavlo Zahrebelnyi, *Roksolana*)

Finally, in both languages there is an interrelation between the polarities of the matrix and the 'until' clause, so that negative matrix clauses show a strong preference for expletive negation.

		subordinate clause			Σ
		∂ ☼	***∂ ne*** ☼	***∂ ne*** ☼→	
main clause	**affirmative**	1267	409	72	1748
	negative	32	296	9	337
Σ		1299	705	81	2085

Table 4: The interplay of polarity in Bulgarian

		subordinate clause					Σ
		∂ ☼	***až ∂*** ☼	***až*** ☼	***až ∂ ne*** ☼	***∂ ne*** ☼	
main clause	**affirmative**	619	277	41	173	600	1710
	negative	24	30	4	43	248	349
Σ		643	307	45	216	848	2059

Table 5: The interplay of polarity in Ukrainian

5. Variation by Source

The frequent optionality of expletive negation makes it a likely mark of the author's or translator's personal style and an area of influence of the original language.

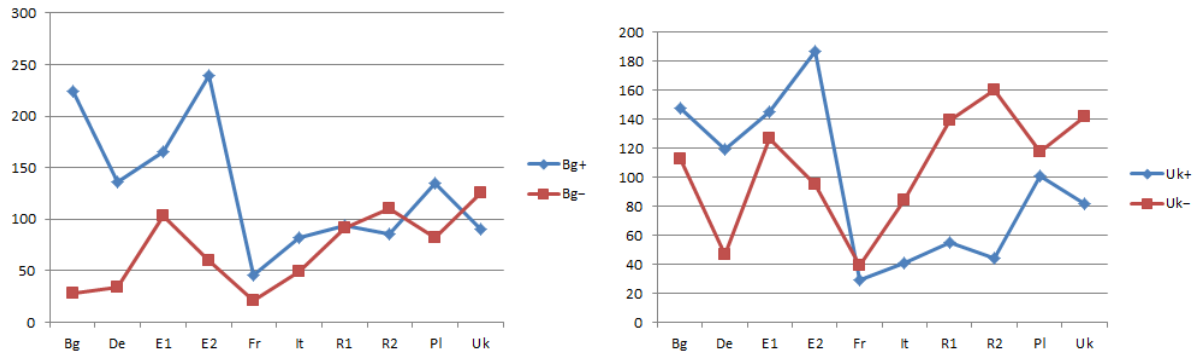

Figure 1: Quantities of affirmative and negative 'until' clauses by sector

Figure 1 shows the numbers of affirmative and negative 'until' clauses (without and with expletive negation) for each sector of the corpus in Bulgarian (on the left) and in Ukrainian (on the right). One can see that in the Bulgarian, German and both English sectors the affirmative 'until' clauses outnumber the negative ones in both languages, with the French sector coming close to this too. On the other hand, in the second Russian and the Ukrainian sector the negative 'until' clauses outnumber the affirmative ones in both languages, with the first Russian sector coming close.

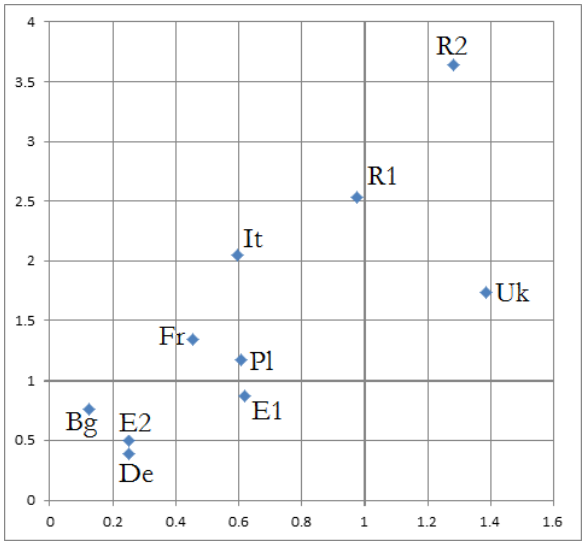

Figure 2: Ratio of negative to affirmative 'until' clauses by sector

Figure 2 shows the ratio of 'until' clauses with expletive negation to 'until' clauses without it for each sector of the corpus in Bulgarian (the *x* axis) and in Ukrainian (the *y* axis). It can be seen that on the Bulgarian side the lowest ratio of expletive negations is in the Bulgarian originals (which is to say that all translators use expletive negation more actively than the authors do) and the highest is in the translations from Ukrainian. On the Ukrainian side the lowest value is in the German sector. There is little doubt that the high frequency of expletive negation in 'until' clauses in Russian is the reason for which the Russian sectors of our corpus feature it in large quantities, but the distance between them, as well as between the two English sectors, proves that the individual authors and translators' choices also play a significant part. Interestingly, the translations from Italian and Polish, the other two languages which use expletive negation in 'until' clauses, assume a middle position in the picture.

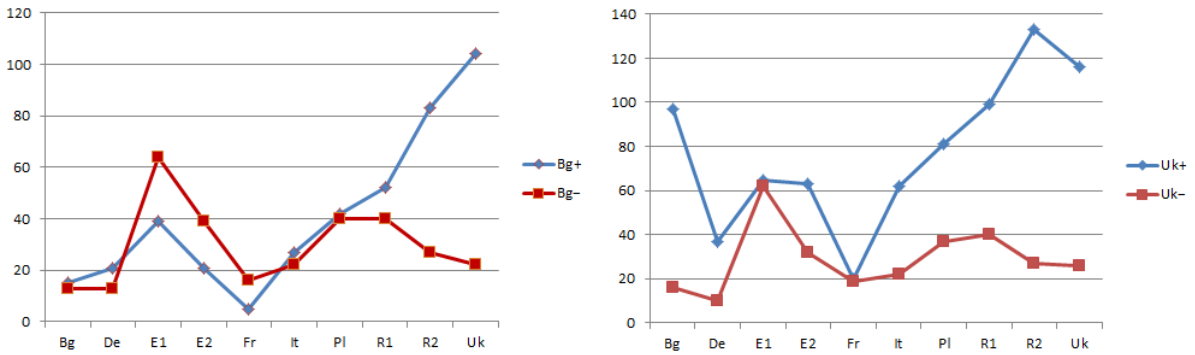

Figure 3: Quantities of affirmative and negative matrix clauses with a negative 'until' clause by sector

Figure 3 shows the numbers of affirmative and negative matrix clauses with a negative 'until' clause (with expletive negation) for each sector of the corpus in Bulgarian (on the left) and in Ukrainian (on the right). It is noteworthy that on the Bulgarian side in the French and both English sectors a negative 'until' clause is more often used with a negative than an affirmative matrix clause (reflecting the high frequency of 'P won't happen *until* Q' constructions in these languages), which is never the case on the Ukrainian side, though the French sector and one of the English come very close.

6. Conclusions

Although Bulgarian and Ukrainian are closely related and share the key phenomena of this study (verbal aspect, lexical merging of 'while' and 'until', expletive negation with the latter meaning), the comparison reveals considerable differences. Expletive negation is much more frequent in Ukrainian than in Bulgarian, where it is more used in translations than in original writing and is largely a mark of the author's style. In Ukrainian an important part is played by the particle and the conjunction *až*, which have no direct counterpart in Bulgarian (the particle is far more often used than Bulgarian *čak*, and the conjunction is not compatible with negation in the embedded clause). Finally, in translated text the grammar and usage patterns of the original language can have a significant impact on the translator's choices. A more detailed study on a larger corpus could endeavour to look for possible diachrony effects as well.

References

Barentsen, A. (1979). Nabljudenija nad funkcionirovaniem sojuza *poka*. In *Dutch Contributions to the Eighth International Congress of Slavists*. Lisse, 57–159.

Derzhanski, I. (1999). Monotonicity and Interrogation. In Corblin, F., Dobrovie-Sorin, C., and Marandin, J.-M. (eds.), *Empirical Issues in Formal Syntax and Semantics 2: Selected Papers from the Colloque de Syntaxe et Sémantique de Paris (CSSP 97)*. The Hague: Thesus, 277–292.

Martinova-Ivanova, P. (2015). Prevodăt na ukrainskite segašni deepričastija na bălgarski ezik. *Săpostavitelno ezikoznanie*, 2015(2):63–70.

Transliteration (1898).
`http://en.wikipedia.org/wiki/Scientific_transliteration_of_Cyrillic`.

Linguistic Data Retrievable from a Treebank

Verginica Barbu Mititelu
Research Institute for Artificial Intelligence "Mihai Drăgănescu"
Romanian Academy
vergi@racai.ro

Elena Irimia
Research Institute for Artificial Intelligence "Mihai Drăgănescu"
Romanian Academy
elena@racai.ro

Abstract

This paper describes the Romanian treebank annotated according to the Universal Dependency principles. We present the types of texts included in the treebank, their processing phases and the tools used for doing it, as well as the levels of annotation, with a focus on the syntactic level. We briefly present the syntactic formalism used, the principles followed and the set of relations.

The perspective we adopted is the linguist's who searches the treebank for information with relevance for the study of Romanian. (S)He can interpret the statistics based on the corpus and can also query the treebank for finding examples to support a theory, for testing hypothesis or for discovering new tendencies. We use here the passive constructions in Romanian as a case study for showing how statistical data help understanding this linguistic phenomenon. We also discuss the kinds of linguistic information retrievable and non-retrievable form the treebank, based on the annotation principles.

1. Introduction

Language resources are created both for the use of machines and for that of humans. Among the latter, several types of users can be recognised: linguistics or/and computer science researchers, teachers (of a native or foreign language), students (studying their native language or learning or studying a foreign one), or any speaker interested in various aspects of the language behaviour.

In this paper we focus on one language resource (a treebank) and show what kinds of linguistic information can be found. The language under focus here is Romanian, but the main lines of the presentation hold for any language having a treebank annotated in the same style.

In Section 2 we present the treebank: the types of texts to which the sentences in the treebank belong, processing steps, the levels of annotation, with a focus on the syntactic one: we briefly present the formalism used, the annotation principles, the inventory of relations used with emphasis on the language specific ones and exemplify with a sentence from the treebank. These data are meant as a background for understanding the rest of the paper. In Section 3 we show what kind of linguistic information can be found in the treebank, looking at passive constructions as a case study, whereas the information that cannot be found and the motivation for this are presented in Section 4. After that, we conclude the paper.

2. The Treebank

The resource which makes the topic of our paper is the Romanian treebank annotated according to the Universal Dependency (UD) guidelines[1]. A treebank is a collection of sentences annotated at the

[1] universaldependencies.org

syntactic level, i.e. syntactic relations among tokens in the sentence are marked and labelled according to their types.

2.1. The Corpus

The treebank, called RoRefTrees, contains 9522 sentences with an average length of 23 tokens. The sentences were selected from several text types: Romanian Wikipedia articles (Wiki), academic writing (Acad), newspaper articles (News), excerpts from different texts that are part of the bibliography of the Romanian Dictionary (Biblio), EMEA (Tiedemann, 2009) in Romanian, FrameNet (Baker et al., 1998) sentences translated into Romanian, the Romanian JRC-Acquis (JRC) (Steinberger, 2006), literature (Lit), medical texts (Medical). The distribution of sentences across text types is not equal, as seen in Table 1, where the Misc(ellanea) column represents a set of sentences from all the other text types (this set was firstly developed as the core of the treebank). Most sentences come from literary and legal texts. The least sentences are from medical texts, which were not among the texts we targeted at the beginning of our work, but added later on.

The tokens in the table below include both words and punctuation. The latter represents approximately 13% from the number of tokens (see Table 2). The longest sentences are in JRC and the shortest in the Biblio subcorpus (we ignored here the Misc subcorpus, given its mixed nature).

	Wiki	**Acad**	**News**	**Biblio**	**EMEA**	**Frame Net**	**JRC**	**Lit**	**Medical**	**Misc**	**TOTAL**
Sents	611	950	933	877	933	1092	1606	**1819**	277	424	**9522**
Tokens	14048	19991	23356	16876	19890	25654	**48295**	37308	7764	7959	**221141**
Length	23	21	25	19	21	23	**30**	21	28	19	**23**

Table 1: Distribution of text types in RoRefTrees.

2.2. Texts Processing and Annotation

The texts in the treebank are tokenised, lemmatised and annotated at the morphologic and syntactic levels. Tokenisation, lemmatisation and morphologic analysis were made with the TTL tool (Ion, 2007). Although TTL uses, for tokenization, a lexicon containing "words with space", we eliminated them in a post-processing phase to comply with the UD requirements: e.g., the compound preposition "de_la" (from) is split into "de" and "la". Words with hyphens, resulted from contractions, are treated by TTL as different tokens: e.g. *n-am spus* (not-have_I said "I haven't said") is tokenised as *n-*, *am* and *spus* (the hyphen marks the elision of the vowel in the adverb of negation *nu* ("not")).

2.3. The Syntax in the Treebank

The annotation level specific to treebanks is the syntactic one. For RoRefTrees, the syntactic formalism we adopted is dependency grammar: each sentence is analysed as a tree (i.e., a directed acyclic graph). Its nodes are the words and punctuation in the sentence, while the edges are relations established between two nodes. All relations are hierarchical. The higher node in a relation is the head and the lower one is its dependent. The only node that has no head in the tree is the root. Any head can have one or more dependents, or even none in the case of tree leaves.

Among the dependency grammars, we chose to work within the UD project, which aims at designing cross-linguistically consistently annotated treebanks for as many languages as possible, with the further aim of developing a parser that could run on sentences in any language.

The syntactic analysis of the sentences was made in an iterative bootstrapping way, starting from two previously available treebanks (Perez, 2014; Irimia and Barbu Mititelu, 2015), which were originally annotated following slightly different principles and sets of relations. The detailed comparison between them can be found in (Barbu Mititelu et al., 2016).

A first set of sentences (about 500) from these treebanks was manually annotated according to the principles and with the set of relations described below and, thus, a small parallel treebank was

created. A correspondence table for the annotations in these parallel treebanks was created and from it a set of structural transformations in the trees were automatically learned and applied, while the conversion of relations was made by a function. The results of the automatic mapping were manually and independently checked by three linguists and, after making the necessary corrections, the sentences were used to enlarge the parallel treebank and the mapping algorithm was retrained and afterwards applied to a new set of sentences. This procedure continued until all sentences from the two treebanks were mapped to the new annotation (see Barbu Mititelu et al., 2016 for the detailed description of this process).

2.4. Annotation Principles

The UD annotation principles are presented on the project website and we mention them here briefly. One central principle is the treatment of function words as dependents, not as heads (except for several clear cases). A flat structure (with the first occurring element as the head and all the others as its dependents) is preferred for coordination, multiword expressions, names, foreign, etc. Active and passive subjects and auxiliaries are marked distinctly. The clausal realisation of syntactic functions is marked distinctly from their lexical realisations.

2.5. The Set of Relations

The set of relations we used is the one in UD, which we augmented with a few language specific ones, motivated by linguistics phenomena in Romanian (see Barbu Mititelu et al., 2015 for motivations).

In UD there is a universal set of relations meant to be used for all languages. Language-specific relations are used for one or several languages displaying a certain phenomenon and are always subtypes of the universal set. In Figure 1 we put in normal font the universal relations. Their subtypes are marked by the presence of the arrow (↳). The language-specific relations used for several other languages in UD are **boldfaced**. They are used to mark the agent in passive constructions (`nmod:agent`), inherently reflexive verbs with a clitic pronoun (`expl:pv`), the reflexive clitic with a passive meaning (`expl:pass`), the clitic with impersonal value (`expl:impers`), the preconjunction (`cc:preconj`), and the noun with temporal value (`nmod:tmod`). The ***boldfaced and italic*** ones are (at least so far within UD) Romanian-specific: the obligatory prepositional object of a predicate (`nmod:pmod`), its clausal equivalent (`ccomp:pmod`), time adverbials (`advcl:tcl`), time adverbs (`advmod:tmod`), possessive dative (`expl:poss`).

Core dependents of clausal predicates			**Non-core dependents of clausal predicates**			**Special clausal dependents**		
Nominal dep	**Predicate dep**		**Nominal dep**	**Predicate dep**	**Modifier word**	**Nominal dep**	**Auxiliary**	**Other**
nsubj	csubj		nmod	advcl	advmod	vocative	aux	mark
nsubjpass	csubjpass		***↳nmod:pmod***	***↳advcl:tcl***	***↳advmod:tmod***	discourse	auxpass	punct
dobj	ccomp	xcomp	**↳nmod:tmod**		neg	expl	cop	
iobj	***↳ccomp:pmod***		**↳nmod:agent**			**↳expl:pv**		
						↳expl:pass		
						↳expl:impers		
						↳expl:poss		
Noun dependents			**Compounding and unanalyzed**			**Coordination**		
Nominal dep	**Predicate dep**	**Modifier word**	compound	mwe		conj	cc	punct
nummod	acl	amod	name	foreign	goeswith		**↳cc:preconj**	
appos		det						
nmod		neg						
Case-marking, prepositions, possessive			**Loose joining relations**			**Other**		
case			list	parataxis	remnant	**Sentence head**	**Unspecified dependency**	
			dislocated		reparandum	root	dep	

Figure 1: Syntactic relations used in RoRefTrees.

The relative frequency of all these relations in RoRefTrees is presented in Table 2. The most frequent relation is `nmod` (marking the nominal modifier of a word). Punctuation comes next and prepositions

(marked with the case relation) after it. Further discussions about the interpretation of data in this table can be found in section 3.1.

Relation	Rel. freq. (%)	Relation	Rel. freq. (%)	Relation	Rel. freq. (%)
nmod	14.6996	ccomp	1.02717	expl	0.24251
punct	13.0446	expl:pv	1.01966	goeswith	0.11675
case	12.2549	cop	0.87435	ccomp:pmod	0.0957
amod	6.56939	iobj	0.81823	remnant	0.06013
det	4.76257	nsubjpass	0.79418	advmod:tmod	0.05411
nsubj	4.63781	parataxis	0.78115	foreign	0.05111
ROOT	4.33166	auxpass	0.73556	expl:impers	0.0466
conj	4.02451	nmod:pmod	0.71501	list	0.04359
advmod	3.76847	neg	0.71	cc:preconj	0.03708
dobj	3.5941	name	0.65939	advcl:tcl	0.03658
mwe	3.04093	expl:pass	0.53814	compound	0.03658
cc	3.03893	appos	0.50106	csubjpass	0.02806
mark	2.89312	xcomp	0.46699	vocative	0.02756
acl	2.28032	nmod:tmod	0.38982	dep	0.00902
aux	2.27631	nmod:agent	0.38431	discourse	0.00802
advcl	1.48414	csubj	0.35776	reparandum	0.0005
nummod	1.34334	expl:poss	0.28811		

Table 2: The relative frequencies of the relations in RoRefTrees.

2.6. Example

A tree from RoRefTrees is presented in Figure 2. It renders the syntactic analysis of the sentence:

(1) (2) Textele acordului, anexelor, protocolului și Actului final se atașează la prezenta decizie.

(2) Texts-the agreement-of-the, annexes-of-the, protocol-of-the and Act-of-the final SE-Cl3SgAcc attach at present-the decision.

"(2) The texts of the agreement, of the annexes, of the protocol and of the Final act are attached to the present decision."

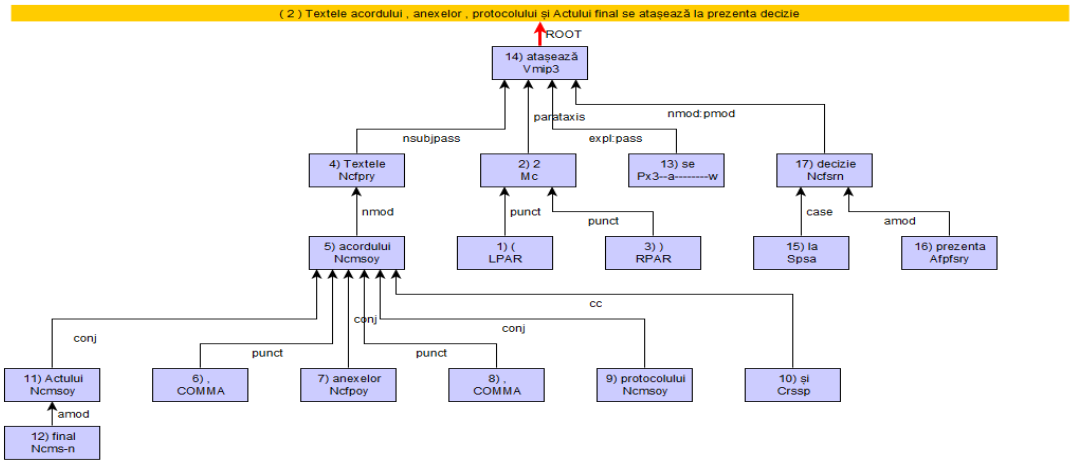

Figure 2: A tree from RoRefTrees.

This is a sentence with a verbal root (*atașează*), a reflexive clitic with passive value (*se*), a nominal subject of a passive verb marked as such (*Textele*). The subject has four coordinated nominal modifiers (*acordului*, *anexelor*, *protocolului*, *Actului*), out of which only the first is analysed as a dependent of the subject, while the others are analysed as conjuncts of it. The commas between the coordinated elements are also attached to the first conjunct, just like the coordinating conjunction (*și*). The preposition (*la*) is attached via the `case` relation to the noun it precedes. The number of the law article (2) (this sentence is from the JRC subcorpus) is attached as `parataxis` to the root of the tree. All punctuation is attached to the head via the relation `punct`: final punctuation to the root of the tree, the parentheses to the element they isolate from the rest of the sentence.

3. What data can a linguist find in the treebank?

A treebank can offer precious information to a linguist in two ways: statistically and by searching it. We will consider them in turn below.

3.1. Let the numbers talk!

In this section we focus on the linguistic relevance of the figures and per cents in the tables above and below. The relative frequencies of specific phenomena either with respect to the whole treebank or to subparts of it can offer information that is difficult to have access to without a treebank. They offer the linguists solid ground for quantitative statements that were difficult to make before the existence of corpora. We will use the passive construction in Romanian as a case study and in the rest of this subsection we will analyse the data pertinent to it found in RoRefTrees.

The passive voice has two possible realisations in Romanian:

- with auxiliary:

(2) Copilul este sărutat de părinte.
Child-the is kissed by parent.
"The child is kissed by his parent."

The passive auxiliary is *este* in this example, the third person singular of the verb *a fi* ("to be").

- with reflexive clitic:

(3) Contractul se va semna mâine de către reprezentanții celor două instituții.
Contract-the SE-Cl3SgAcc will sign tomorrow of towards representatives-the those-of two institutions.
"The contract will be signed tomorrow by the representatives of the two institutions."

The reflexive clitic with passive meaning is *se* in such constructions.

The relations identifying the passive in our treebank are: `auxpass` (the passive auxiliary), `refl:pass` (the reflexive clitic with a passive value), `nsubjpass` (the nominal subject of a passive verb), `csubjpass` (the clausal subject of a passive verb), `nmod:agent` (the nominal agent complement of the verb). A clausal realization of the agent complement is possible in Romanian, but it never occurred in our treebank. The first two relations are mandatory for a sentence to be interpreted as passive (but they cannot co-occur). The others are optional: the absence of the subject (either nominal or clausal) is possible given the fact that Romanian is a pro-drop language, whereas the absence of the agent nominal is "a vivid iconic manifestation of the most salient functional-pragmatic feature of the passive voice – agent suppression" (Givón, 2001: 126).

The data in Tables 2 and 3 shows several things related to passive. First, no relation specific to passive is too frequent in any text type. From this we can conclude that active voice is much more frequent than passive voice. Second, the distribution of the passive construction across the text types shows that the passive is the most frequent in the EMEA texts. According to linguistic literature on this topic (see Quirk et al., 1984: 166; Givón, 2001: 125; among others), informative texts favour passive constructions; the data in Table 4 shows the same tendency: EMEA sentences, i.e. scientific texts, have the highest relative frequency of passive structures, while Lit and Biblio, i.e. imaginative texts, have the lowest relative frequency of passive constructions. What is intriguing is that Wikipedia texts, which belong to the category informative rather than imaginative, show a lower frequency of passives than imaginative texts. A motivation for this will probably be found when a semantic analysis of the sentences from Wikipedia is made.

Third, the passive voice with auxiliary is much more frequent than the reflexive passive: the relative frequency of `auxpass` is higher than the relative frequency of `expl:pass`. In Table 3 we can see that this statement holds true for all text types in the treebank, with the only exception of the JRC sentences, in which the impersonal reflexive form prevails.

Fourth, the passive subjects are also most frequently realised in the EMEA sentences (0.0154) (see the line "passive subjects" in Table 4). However, the lexicalization of the subject in passive sentences happens most frequently in the Biblio subcorpus. The explanation resides in the fact that these sentences were selected by the dictionary editors to serve as examples of the usage of a lexical unit, so they must be characterized by semantic and syntactic completeness, coherence, cohesion, lack of ambiguity.

Fifth, the most frequent type of subject realisation is the nominal one (in more than 95% of the cases: see line "%nsubjpass" in Table 4) and its relative frequency is the highest in the Acad subcorpus. This correlates with the data in Tabel 2, which show higher relative frequencies for `nsubj` and `nsubjpass` than for `csubj` and `csubjpass`.

Sixth, the relative frequency of the realisation of agents in passive structures is below 50%, with the highest in Acad: 0.5085. However, one can see that in Wikipedia texts the relative frequency of the realisation of agent is 1.1641. This is informative of the fact that nominal agents occur in constructions that are not syntactically passive, but carry a passive meaning: for instance, the verbal nominalisation in this example:

(4) Sărutarea copilului de către părinte ....
Kissing-the child-the-of of towards parent ….
"The kissing of the child by his parent …."

The noun *sărutarea* ("the kissing") preserves the semantic arguments of the verb it is derived from: the agent and the patient. The former is realised in the same morpho-syntactic form as in the passive voice, namely with the compound preposition "de către" (*by*).

	Acad	**News**	**Biblio**	**EMEA**	**FrameNet**	**JRC**	**Lit**	**WIKI**
auxpass	0.0081	0.0106	0.0036	**0.0151**	0.0067	0.0068	0.0038	0.0038
expl:pass	0.0036	0.0063	0.0029	0.0082	0.0007	**0.0102**	0.0026	0.0009
nsubjpass	0.0083	0.0116	0.0052	**0.0147**	0.0045	0.0110	0.0031	0.0011
csubjpass	0.0001	0.0005	0.0001	**0.0007**	0.0002	0.0002	0.0001	0.0000
nmod:agent	**0.0060**	0.0053	0.0024	0.0025	0.0027	0.0043	0.0023	0.0056

Table 3: The relative frequency of relations connected to passive voice in RoRefTrees subcorpora.

	Acad	**News**	**Biblio**	**EMEA**	**FrameNet**	**JRC**	**Lit**	**WIKI**
passive structure	0.0117	0.0170	0.0065	**0.0233**	0.0074	0.0171	0.0065	0.0048
passive subjects	0.0084	0.0121	0.0053	**0.0154**	0.0047	0.0112	0.0032	0.0011
$\frac{passive\ subjects}{passive\ structures}$	**0.7136**	0.7121	**0.8153**	0.6609	0.6401	0.6553	0.4896	0.2239
% agent	**0.5085**	0.3131	0.3692	0.1079	0.3596	0.2499	0.3486	**1.1641**
% nsubjpass	**0.9880**	0.9574	0.9811	0.9542	0.9551	0.9814	0.9661	1

Table 4: Further relative frequencies connected to passive voice in RoRefTrees.

3.2. What types of searches can be made in the treebank?

Besides analysing the figures in the statistics drawn from the treebank, the linguist can also search for various structures and their instantiation in it. RoRefTrees are available for download on the UD website, with the content from the last release. The treebank can also be queried online using different tools: at http://bionlp-www.utu.fi/dep_search, using SETS querying system, described at http://bionlp.utu.fi/searchexpressions-new.html; at http://lindat.mff.cuni.cz/services/pmltq/#!/home, using PML Tree Query, described at https://ufal.mff.cuni.cz/pmltq/doc/pmltq_doc.html; at http://clarino.uib.no/iness/page?page-id=iness-main-page, with the INESS (Rosén et al., 2012) infrastructure, described at http://clarino.uib.no/iness/page?page-id=iness-documentation.

One can search a treebank for a multitude of linguistically relevant data. Their analysis reflects the grammatical theory that was used for annotation. We present below several examples of searches:

- the arguments of a certain verb: one can extract all core dependents of the respective verb, even with the aim of creating a valence dictionary of the verbs in the treebank; these core dependents are words linked to the respective verb by any of the relations `nsubj`, `nsubjpass`, `csubj`, `csubjpass`, `dobj`, `iobj`, `ccomp`, `ccomp:pmod`; besides them, one must also consider `nmod:pmod` and `nmod:agent` relations, although they are classified under non-core dependents in Figure 1;
- the parts of speech a certain syntactic function can be realised by: for example, what parts of speech the root of a clause can be; in RoRefTrees one will find verbs, interjections, nouns, adjectives and adverbs as roots. If Romanian traditional grammar has the notions of predicative interjections and adverbs, so these two parts of speech are no surprise among the results, then the adjective and nouns are unexpected roots in non-elliptical structures, but this is the result of the convention used for annotating the copula verb *a fi* (“to be”): a dependent on the adjective or noun, linked by the `cop` relation: in Figure 3 we present the analysis of the adjective *frumoasă* from the sentence in (5) as the root of the sentence.

(5) Fata este frumoasă.
Girl-the is beautiful.
“The girl is beautiful.”

Figure 3. A sentence with an adjectival root.

- the words realising a syntactic function for a certain word: one may want to identify the semantic restrictions on a certain argument of a verb; this can be done by analysing all the words filling that position in the argument structure of the respective verb in the treebank;
- the parts of speech between which a certain syntactic relation establishes: for example, `iobj`, which is found in our treebank as occurring between nouns, pronouns as dependents and verbs, adjectives or interjections as heads. The analysis can go even further: one can look at various morphologic characteristics of these parts of speech, such as case for nouns or pronouns;
- the word order (even in different types of sentences, such as declarative, interrogative, exclamatory, affirmative, negative); an interesting study for a language with relative free word order would be the position of the subject, when lexicalised: pre- or post-verbal position.
- etc.

4. What Cannot Be Found in RoRefTrees?

The conventions in the formalism adopted for creating the treebank have consequences in the type of information retrievable from the treebank. We discuss several disadvantages of the annotation here.

When designing the set of relations to be used in the syntactic annotation (within UD), both structure and function were considered. Some relations clearly reflect the way dependents function in the sentence: `dobj`, `iobj`, etc. Others reflect rather the morphologic components: see `nmod` and `advmod` relations: the former functionally corresponds to an adverbial when it attaches to a verb, adjective or an adverb, but when attaching to a noun, it corresponds to an attribute; the latter is an adverb or adverbial phrase that serves to modify the meaning of its head. There are others that combine both aspects: `nsubj`, `csubj`: they are used for the same syntactic position (a subject), but the former is used for nominals filling this position, while the latter for clauses.

Sometimes, the same relation is used to link both arguments and adjuncts to their heads: e.g. `advmod`. It is impossible to automatically distinguish between adverbs that are arguments, as in (6), and those that are adjuncts, as in (7), as the same relation (`advmod`) links them to their head.

(6) El se poartă frumos.
He Se-Cl3SgAcc behaves beautifully.
"He behaves himself."
(7) El cântă frumos.
He sings beautifully.
"He sings beautifully."

In Figure 1, one can notice that the clausal realisation of both the direct and indirect objects is linked to the head by the same relation, `ccomp`, which means that no distinction between the two positions can be made automatically. One way of disambiguating this relation is to look for a `dobj` or `iobj` of the same head: as there cannot be two `dobj` or `iobj` relations of the same head, the co-occurrence between a `dobj` and a `ccomp`, for instance, will help infer the fact that the subordinate clause fills the indirect object slot of the head argument structure. Otherwise, we cannot see another way for telling the values of the `ccomp` apart.

5. Conclusions

Nowadays, when language resources are being created and their size is in continuous increase, the researchers interested in the study of a language focus more on these resources, search them for known facts and new emerging tendencies. Besides merely reflecting various phenomena, corpora in general and treebanks in particular also inform about their frequency, which can mark either an increasing tendency or, on the contrary, rare phenomena.

We presented above the Romanian treebank annotated according to UD conventions and discussed about several information types a linguist can search for and find in it. Others remain covert and other solutions need to be found for spotting them in the treebank.

Acknowledgements

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS–UEFISCDI, project number PN-II-RU-TE-2014-4-1362.

References:

Baker, C. F., Fillmore, Ch. J., Lowe, J. B. (1998). The Berkley FramNet Project. *Proceedings of the 17th International Conference on Computational Linguistics* - Volume 1.

Barbu Mititelu, V., Irimia, E., Mărănduc, C. (2015). Universal and Language-specific Dependency Relations for Analysing Romanian. *Proceedings of the Third International Conference on Dependency Linguistics (DepLing2015)*, August 24-26, Uppsala, Sweden.

Barbu Mititelu, V., Ion, R., Simionescu, R., Irimia, E., Perez, C.-A. (2016). The Romanian Treebank Annotated According to Universal Dependencies. *Proceedings of HrTAL*, September 29 – October 1, Dubrovnik, Croatia.

Givón, T. (2001). *Syntax: An Introduction*. Vol. II. Amsterdam/Philadelphia: John Benjamins.

Ion, R. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian*, PhD thesis, Romanian Academy (in Romanian).

Irimia, E., Barbu Mititelu, V. (2015). Building a Romanian Dependency Treebank. *Corpus Linguistics 2015*, Lancaster, UK, 21-24 July 2015.

Nivre, J., Hall, J. Nilsson, J. (2006). MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*, May 24 – 26, 2006, Genoa, Italy.

Perez, C.-A. (2014). *Resurse lingvistice pentru prelucrarea limbajului natural*, PhD thesis, A.I. Cuza University of Iasi.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Rosén, V., De Smedt, K., Meurer, P., Dyvik, H. (2012). An Open Infrastructure for Advanced Treebanking. In: Hajič, J., De Smedt, K., Tadić, M., Branco, A. (eds.) *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, May 21-27, Istanbul, Turkey.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiș, D., Varga, D. (2006). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. May 24-26, Genoa, Italy.

Tiedemann, J. (2009). News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (eds.) *Recent Advances in Natural Language Processing* (vol V). Amsterdam/Philadelphia: John Benjamins.

Towards the Automatic Identification of Light Verb Constructions in Bulgarian

Ivelina Stoyanova, Svetlozara Leseva, Maria Todorova
Department of Computational Linguistics
Bulgarian Academy of Sciences
{iva, zarka, maria}@dcl.bas.bg

Abstract

This paper presents work in progress focused on developing a method for automatic identification of light verb constructions (LVCs) as a subclass of Bulgarian verbal MWEs. The method is based on machine learning and is trained on a set of LVCs extracted from the Bulgarian WordNet (BulNet) and the Bulgarian National Corpus (BulNC). The machine learning uses lexical, morphosyntactic, syntactic and semantic features of LVCs.

We trained and tested two separate classifiers using the Java package Weka and two learning decision tree algorithms – J48 and RandomTree. The evaluation of the method includes 10-fold cross-validation on the training data from BulNet ($F_1 = 0.766$ obtained by the J48 decision tree algorithm and $F_1 = 0.725$ by the RandomTree algorithm), as well as evaluation of the performance on new instances from the BulNC ($F_1 = 0.802$ by J48 and $F_1 = 0.607$ by the RandomTree algorithm). Preliminary filtering of the candidates gives a slight improvement ($F_1 = 0.802$ by J48 and $F_1 = 0.737$ by RandomTree).

1. Introduction

Multiword expressions (MWEs) have been estimated to represent a substantial portion of the lexical system of a language. For example, it has been reported that 41% of the literals of the Princeton WordNet 1.7 are MWEs (Sag et al., 2002). Other scholars propose that multiword expressions are quantitatively equivalent to simple words (Jackendoff, 1997) or even that the number of MWEs is much more prevalent than the number of single words (Melčuk, 1998). This makes the systematic description of MWEs a very important task which influences largely the performance of the applications in the field of information extraction, text summarisation, machine translation and other NLP areas.

The work presented here focuses on LVCs as a subclass of Bulgarian verbal MWEs with a view to their automatic recognition and annotation. LVCs consist of a verbal component and a complement that function as a semantic whole. Like nominal predicates, the LVCs' complement carries the predicative meaning of the MWE.

After overviewing the related work in the field (Section 2.), we discuss the specific properties of LVCs (Section 3.) with a focus on Bulgarian as a morphologically rich language with free word order. Section 4. presents a method for automatic identification of LVCs which relies on machine learning and uses as features various lexical, semantic, morphosyntactic, syntactic, statistical, and derivational properties of LVCs and their components. The evaluation of the method includes 10-fold cross-validation on training data compiled from the Bulgarian WordNet (BulNet) and the Bulgarian National Corpus (BulNC), as well as evaluation of the performance on new examples from the BulNC. We conclude the paper with a discussion of the results.

2. Related Work

Recent research into MWEs focuses on verbal and other MWEs and the description of their components and structure (Villavicencio et al., 2004; Gregoire, 2010; Francopoulo, 2013; Gralinski et al., 2010). Nonetheless, the challenges to the lexicographic description of verbal MWEs posed by morphologically rich languages have not been completely addressed yet. These include: rich inventories of synthetic and analytical verb forms with a complex word order, flexible word order of the components of the verbal MWEs, structural features, such as mandatory and optional components, the possibility of having discontinuous components with intervening external elements, etc.

The description of MWEs, such as the ones proposed by Nunberg et al. (1994), Sag et al. (2002), Baldwin et al. (2003), among others, deal with the restrictions imposed on the internal structure, syntactic behaviour and semantic properties of MWEs, which affect significantly their linguistic annotation and automatic processing. LVCs require special treatment in natural language processing as part of the group of verbal MWEs since, unlike free phrases, their meaning is not fully decomposable to the meanings of their components, and they are often not translated to other languages literally. LVCs also need to be distinguished from idioms since they have greater syntactic flexibility and are more semantically predictable as compared with idioms.

Mostly, research on LVCs is focused on either a limited number of light verbs, e.g. candidates containing the verbs *make* and *take* (Stevenson et al., 2004), or on certain syntactic subtypes, e.g. verb–noun (Fazly and Stevenson, 2007) or verb–preposition–noun combinations (de Cruys and Moirón, 2007).

A variety of methods for the identification of LVCs have been reported: semantic (de Cruys and Moirón, 2007), statistical (Gurrutxaga and Alegria, 2011), rule-based (Vincze et al., 2011), or hybrid methods (Tan et al., 2006). Some methods are focused on the alignment of LVCs in parallel corpora (Samardžić and Merlo, 2010).

Tu and Roth (2011) propose a supervised system that applies machine learning on a manually annotated corpus of positive and negative examples of LVCs with the six most frequent English light verbs: *do*, *get*, *give*, *have*, *make* and *take*. They train two systems – one on statistical and one on contextual features. Their findings show that both show similar results in general, however the one trained with contextual features is more accurate and robust with respect to identical surface structures that may or may not be LVCs, e.g. *have a look*.

Nagy et al. (2013) aim at a full coverage of LVCs and propose a two-step method by which they first identify potential LVC candidates in running texts and then use a machine learning-based classifier to select LVCs from the candidates. The selection of LVCs is based on feature templates like semantic or morphological features of LVCs in context. Their approach distinguishes cases where the phrase does not function as an LVC from true LVCs.

Linguistic knowledge is crucial to the work on LVC recognition and researchers have made use of various linguistic features – morphological, including derivational features, lexical, syntactic, semantic features. The studies on LVCs uniformly make use of the surface syntactic structure typical of LVCs, predominantly V NP and V PP constructions. The light verbs are usually specified in advance and are restricted to a well-defined small set (Stevenson et al., 2004; Tu and Roth, 2011). The opposite approach – accounting for a broader range of light verbs, as presented in the data – is proposed by Nagy et al. (2013). This latter approach is also adopted in our study.

Syntactic information is also employed, to make sure that a potential LVC represents a syntactic unit. Depending on the adopted syntactic framework and linguistic resource, researchers use combinations that represent particular dependency relations (Nagy et al., 2013; Chen et al., 2015) or combinations of constituents (Tu and Roth, 2011).

The works on LVCs typically employ particular restrictions on the semantics of the nominal component. Tu and Roth (2011) and Nagy et al. (2013) make use of the tendency of LVCs to correspond with derivationally related verbs, e.g. *pravya poseshtenie – poseshtavam* (*pay a visit – to visit*). Nagy et al. (2013) and Chen et al. (2015) also use the semantics of the nominal component by looking at its WordNet hypernyms, such as *activity* or *event*. Stevenson et al. (2004), who look for verbs instead of for event nouns (thus capturing those LVCs in which the nominal component is a deverbal noun coinciding

in form with the verb), use a selection of Levin's verb classes as the complements of light verbs. Levin's classes are also adopted by Tu and Roth (2011). Chen et al. (2015) employ diverse semantic information from different resources, including hypernyms from WordNet, the WordNet noun types (the type is the semantic primitive, such as *person, animal, artifact, change, state*, etc. that is assigned to each noun and verb synset), as well as richer semantic descriptions, such as WordNet senses, word senses from OntoNotes (Pradhan et al., 2007) and information from PropBank (Palmer et al., 2005).

Within the context of Slavic languages, work on LVCs has been reported for Russian (Mudraya et al.,), Serbian (Samardžić, 2008; Samardžić and Merlo, 2010), Croatian (Gradečak-Erdeljić and Brdar, 2012), Czech (Urešová et al., 2016) and Polish (Przybyszewski, 2015), mainly with a view to LVCs' annotation in treebanks and multilingual parallel corpora or to the theoretical description of their compositionality and semantics. With respect to Bulgarian, LVCs have been tackled within a proposal made by Koeva (2006) of a framework for morphosyntactic description of MWEs, which has subsequently been partially incorporated in the construction of a large dictionary of Bulgarian MWEs enriched with diverse morphological, syntactic and structural information (Koeva et al., 2016).

Recent work on LVCs has been undertaken for many languages, Slavic languages and Bulgarian in particular, within the PARSEME Shared Task on automatic detection of verbal Multi-Word Expressions[1]. The Shared Task aims at the automatic identification of verbal MWEs in running texts. This paper is part of the work of the Bulgarian team on the Shared Task.

3. Properties of LVCs

LVCs are usually treated as constructions involving complex predication that takes place between a verb and another predicative element (Butt, 2003; Goldberg, 2003; Jackendoff, 1974; Wittenberg et al., 2014), among others. The verb belongs to a relatively small set of verbs whose meaning is more or less abstract ('semantically bleached') and mainly express aspect, directionality or aktionsart of the predicate (Butt, 2003; Wittenberg et al., 2014). The semantic properties of light verbs correspond to the fact that they have high frequency, and, as we have observed in the data, generally exhibit high polysemy.

The structure of LVCs varies according to the light verb's formal complement. It is usually a noun phrase that corresponds to the direct object position of the verb – *vzemam reshenie* (*make a decision*) or a PP corresponding to an indirect object – *vlizam v kontakt* (*come into contact*). The light verb's complement may also be an adjective – *pravya lud* (*make crazy*) or an adverb *vzemam predvid* (*take into consideration*).

The LVC's complement is in fact the semantic predicate and contributes the major part of the complex meaning of the expression. As a rule, it is an abstract entity with eventive or similar semantics and is frequently expressed by a deverbal predicative noun, although as noted by Cinkovà and Kolářová (2005), deadjectival nouns such as *vazmozhnost* (*possibility*) also occur. It has been proposed that the complements' sense is non-figurative (Vincze et al., 2016).

As frequently noted in the literature, an important trait of LVCs, which follows from the eventive nature of the complement, is that the construction often has a corresponding synonymous single verb derivationally related to the eventive noun. The verb–noun relation is usually through suffixation *resha/V – reshenie/N* (*decide – decision*) or through zero derivation *dokladvam/V – doklad/N* (*a report – to report*). This trait has often been employed as an additional diagnostic for LVCs although its large coverage over the data is usually taken for granted. At least judging from our data for Bulgarian, we may conclude that it is not very large. We have found out that only 265 of the 621 LVCs found in the Bulgarian WordNet have a single verb counterpart. For instance, the substitution may not be possible due to lexical gaps, e.g. *slagam kray (na)* (*put an end to*) does not have a corresponding verb that is derivationally related to the noun *kray* (*end*). On the other hand, it does not automatically follow from the substitutability with a single verb that an MWE is an LVC. The 265 LVCs in BulNet that have a corresponding single verb make up half of the 541 BulNet MWEs with a single verb correspondence. Nevertheless, as derivation is relatively easy to identify, this feature may be successfully employed in conjunction with more decisive ones.

[1] http://typo.uni-konstanz.de/parseme/

Another characteristic of LVCs is the possibility to refer to the same event by using the nominal alone, e.g. *He had a walk in the park* vs. *His walk in the park* (Vincze et al., 2016). This trait was used as a diagnostic in the manual categorisation of MWEs in the Bulgarian WordNet as either being LVCs or non-LVCs, as well as in the inspection of LVC candidates extracted from the Bulgarian National Corpus which took place in the process of compiling the training and the test data.

The nominal complement in V NP LVCs tends to be able to take a plural and/or a definite form, e.g. *vzemam reshenie* (*make a decision*) may be found as *vzemam reshenieto* (sg. def.), *vzemam resheniya* (pl. indef.) *vzemam resheniyata* (pl. def.) although this is not always the case – e.g. *vzemam uchastie take part* does not allow free variation of the noun. Still, this tendency of LVCs may serve in addition to other diagnostics to distinguish LVCs from idioms with the same surface structure.

Another specific feature of the LVC complement is that it may be modified, e.g. *vzemam vazhno reshenie* (*make an important decision*), *vzemam deyno uchastie* (*take an active part*). This is another linguistic trait that distinguishes LVCs from some other types of MWEs, idioms in particular, which may allow only very limited modification (practically a lexical variant of the idiom), e.g. *vdigam letvata* (*raise the bar*) and its variant *vdigam letvata visoko* (*raise the bar high*).

The components of an LVC may also be separated by other elements, such as adjuncts of the entire LVC, e.g. *vzemam* ***barzo*** *reshenieto* (*make* ***quickly*** *the decision*), or elements that are external with respect to the LVC. Among the latter are the question particle *li* and pronominal clitics, e.g. *Vzeha* ***li*** *reshenieto?* (*Did they make the decision?*) and *Napraviha* ***mu*** *operatsiya.* (*They made* ***him*** *an operation*). More than one external element may be found *Napraviha* ***li mu*** *operatsiya?* (*Did they make* ***him*** *an operation?*) as well as longer sequences. This trait is shared with free phrases and many idioms, but we point it out as it needs to be taken into account when determining the search scope for an LVC in the corpus.

4. A Method for Automatic Identification of LVCs

We developed a method for automatic recognition of LVCs in running text based on observations made on the properties of light verbs and LVCs discussed by various authors (see Section 3.), as well as some specific features that we consider relevant for Bulgarian and other morphologically rich languages. The method is implemented in Java, using the Weka library for data mining (Hall et al., 2009).

4.1. Resources

For the purposes of automatic identification of LVCs we compiled a subcorpus of the Bulgarian National Corpus (BulNC)[2] (Koeva, 2014a), containing news (35,758 texts, amounting to 10,655,068 words) and fiction texts (443 texts, a total of 6,237,024 words). The corpus was annotated using the Bulgarian Language Processing Chain (Koeva and Genov, 2011), which is available as a web service using a RESTful API. The annotation includes sentence splitting, tokenisation, POS tagging and lemmatisation.

We also used another language resource, the Bulgarian Wordnet (BulNet) (Koeva, 2014b), from which we extracted a list of 2,239 verbal MWEs (MWE synonyms in verb synsets) containing at least a verb and a noun. We determined the internal syntactic structure of each MWE by analysing its components as a sequence of POS tags and obtained the following structural types: verb – direct object (V–NP), e.g. *vzemam dush* (*take a shower*) or verb – indirect object (V–PP), e.g. *vzemam pod vnimanie* (*take into consideration*). MWEs of other syntactic types, e.g. V–AdvP, V–AP, were not taken into account. The set of MWEs selected in this way constitutes the main part of the training data for the machine learning, after being manually divided into LVCs and non-LVCs.

Further, we used BulNet to extract words that can occur as part of LVCs. First, we extracted 74 highly ambiguous verbs (verbs with 15 or more senses in BulNet). These verbs were subsequently examined and non-light verbs were filtered out. The remaining 46 verbs were merged with a list of 81 verbs that were found as the heads of those MWEs in BulNet that were manually validated as LVCs. After the duplicate entries were removed, the compiled list totaled 105 verbs. Table 1 presents the distribution of light verbs with respect to the number of senses in BulNet and their frequency in the BulNC. Only a

[2] `http://search.dcl.bas.bg/`

# senses	# verbs	Frequency	# verbs
<5	13	<50	4
≥ 5	68	≥ 50	77
≥ 10	42	≥ 100	70
≥ 20	23	≥ 500	43
≥ 50	4	≥ 1000	31

Table 1: Distribution of light verbs according to: (a) number of senses in BulNet; (b) frequency in the BulNC.

small number of verbs have less than 5 senses (13 verbs) or low frequency of less than 50 occurrences (4 verbs), and no verb has both low frequency and a small number of senses.

Next, BulNet served us to extract semantic information about the components of the LVCs. All the verb and noun synsets in the Princeton WordNet (and respectively in BulNet) are each assigned a single semantic primitive out of a list of language-independent primitives that represent the unique beginners of the separate hierarchies in WordNet (Miller, 1998) (initially organised in separate lexicographer files). We consider 10 of the noun semantic primitives, such as *noun.act*, *noun.state*, *noun.cognition*, etc., as potentially expressing predicative meaning, while excluding the remaining 15 noun primitives, such as *noun.artefact*, *noun.person*, etc.[3] The set of potential semantic primitives of all the possible senses of a given noun were used as features in the machine learning.

4.2. Machine Learning Features

For the purposes of machine learning we defined a number of features capturing the essential linguistic traits of MWEs and LVCs in particular.

1. Lexical features

   We use the verb's lemma as a feature in the machine learning, relying on the fact that certain light verbs can potentially combine with certain (classes of) nouns, e.g. *poemam* {*risk, otgovornost*} (*assume* {*risk, responsibility*}), while other combinations are limited or impossible, e.g. **vzemam* {*risk, otgovornost*} (*take* {*risk, responsibility*}).

2. Semantic features

   The semantic features include the semantic primitives of the nouns which are extracted from BulNet. As noted above, we selected 10 (of the overall 25) of the noun semantic primitives which are relevant for predicative nouns: *noun.act, noun.cognition, noun.communication, noun.event, noun.feeling, noun.motive, noun.phenomenon, noun.process, noun.relation, noun.state*. For a given ambiguous noun, all the possible labels were extracted and represented as a set. In the cases where the different senses of a noun correspond to different labels, additional procedures were performed. If a noun is associated with a semantic primitive that is not typical for predicative nouns, the primitive (and the respective sense) was excluded from the noun's description. For instance, the noun *vapros* (*question, issue*) was found in BulNet with the following primitives: {noun.act, noun.communication, noun.cognition, noun.attribute, noun.event} and the sense having the primitive {noun.attribute} was excluded. However, a noun which predominantly appears in BulNet in non-predicative senses (more than half of the senses), is taken to be non-predicative and is consequently ignored as a possible nominal component within an LVC.

3. Statistical features

   The statistical features contain information about the frequency of potential LVCs and their components in the corpus, i.e. the log-frequency (logarithm of the observed absolute frequency to the base of 2) of: (a) the verb, (b) the noun, and (c) the LVC candidate. The logarithmic transformation linearises the distribution of frequencies and allows for simpler correlation analysis with other

[3]The list of primitives is available at `https://wordnet.princeton.edu/man/lexnames.5WN.html`

features. Based on the observed frequencies we also calculated the association measure (using Mutual Information, MI) of the LVC candidate in order to determine whether it is a collocation and, potentially, an MWE.

4. Morphosyntactic features

   The morphosyntactic features account for the fact that the nominal complement of many LVCs does not occur in a single fixed form, but may take both singular and plural and/or indefinite and definite forms. Of course, there are cases in which there are restrictions on the form of the nominal complement, e.g. *pravya vpechatlenie* (*make an impression*), in which the noun is used as part of the LVC only in the singular indefinite form. Moreover, in rare cases the noun may occur with two different senses in different LVCs with the same verb, where the only difference is the form of the noun, e.g. *vzemam myarka* (sg. indef.) (*take measures*, to measure dimensions) as opposed to *vzemam merki* (pl. indef.) (*take measures, actions*). We leave the detailed analysis and handling of these cases for the future.

   Variability in components is more likely for LVCs than for idioms, that is why we introduce a binary feature which takes **true** if the noun is found in more than one form in the corpus (singular and/or plural, indefinite and/or definite, count (for masculine nouns)) and **false** if the noun is invariable.

5. Syntactic features

   The syntactic features included in the machine learning account for the following properties of LVCs:

   (a) **LVCs allow different word order.** As the relatively free word order in Bulgarian makes it possible for the complement to precede the verb in various contexts, we took into account both word order variants. The feature takes the value *true* if more than one word order is registered in the corpus and *false* otherwise.

   (b) **Components may take modifiers.** As mentioned above, the LVC components may be separated either by modifiers and adjuncts of the LVC or by external elements. For the purposes of the current study, we limited the distance between the light verb and its noun complement (or the noun complement of the PP in V PP LVCs) to be up to two tokens. Possible modifiers of the noun were limited to adjectives preceding the noun. The feature takes the value *true* if an example with a modifier is found in the corpus and *false* otherwise.

   (c) **LVCs allow external elements to occur between their components**. External elements were identified by their POS in order to generalise the cases. The feature takes the value **true** when the POS tags of the elements (found at distance of at most two tokens) are other than 'adjective' or 'preposition', and the value **false** otherwise. Adjectives are considered as possible modifiers to the noun (see (b) above), while prepositions are likely to introduce a PP component of the vMWE. Another restriction currently adopted is that the tokens that may separate the components of an LVC cannot be punctuation marks or conjunctions since these usually mark phrase or clause borders.

6. Derivational features

   We defined a derivational feature that takes into consideration the strong tendency for predicative nouns to be of deverbal stems and therefore – to be derivationally related to a verb. The feature takes the value **true** if a derivational relation is found, and the value **false** otherwise.

   In order to establish a derivational relation, we looked for a common stem between a (potential) nominal component of an LVC and any verb. The common stem was estimated empirically using the output of a stemmer implemented for this and other related tasks. The stemmer matches words which share a substring whose length is at least 70% of each word's length and longer than 4 characters. For instance, in the LVC *nanasyam vreda* (*cause damage*) the noun *vreda* (*damage*) is marked as derivationally related to the verb *vredya* (*to damage*) by matching the stem *vred-*.

4.3. Compilation of the training and the test dataset

The main part of the training dataset consists of the 2,239 V–NP and V–PP MWEs which were classified into two categories – 'LVC' and 'non-LVC' (see Section 4.1.) using automatic procedures and manual post-editing. As a result, a total of 461 MWEs were identified as LVCs and the remaining – as other types of verbal MWEs. To overcome the low number of the LVCs in the training data and the lack of non-MWE instances, we extended the training set with additional data from the BulNC. To this end we extracted verb–noun pairs with frequency of at least 10 in the corpus, which were then manually categorised into 'LVC' (true) and 'non-LVC' (false) by two independent annotators. We took into account the instances in which the annotators agreed.

In determining whether an MWE from BulNet or a candidate extracted from the BulNC is in fact an LVC, the annotators took into consideration several linguistic factors: (a) whether the verb qualifies as a light verb (i.e. is on the list of light verbs we identified); (b) whether the noun denotes an event or a similar semantic type of entity (state, property, etc.); (c) whether the noun is used in a non-figurative meaning; (d) whether the noun alone may be used to denote the same event.

For instance, using these diagnostics we conclude that the candidate *nanasyam shteti* (*cause damage*), which complies with (a)–(d), is an LVC: the verb has an abstract causative meaning with which it combines with a variety of nouns; the noun denotes an event or a result of an event; it is used in its primary literal sense; the noun can be used alone to refer to the event, as in: *Shtetite ni ne byaha kompensirani.* (*Our damages were not recompensed.*) In contrast, consider the idiom *podavam raka* (*lend a hand*) where: the verb is not semantically bleached; the noun may denote an act, but only in a figurative sense, whereas its literal sense denotes a body part. Besides, the noun cannot be used alone to refer to the event.

As a result, the training dataset compiled from BulNet and the BulNC consists of 2,623 instances, 897 of which are LVCs and the remaining are either non-MWEs or other categories of MWEs (e.g., idioms).

The test dataset comprises 200 unique candidates with frequency of at least 10 extracted from the BulNC in the same way as the additional training instances and annotated by the annotators into LVCs and non-LVCs, with equal number of both categories.

4.4. Method outline

We trained and tested two classifiers on the feature set (Section 4.2.) and the training set (Section 4.3.) using two different learning algorithms based on decision trees – J48 and RandomTree (Hall et al., 2009). The method for LVC identification is performed in the following steps:

(1) Identify LVC candidates in the corpus – the occurrences of a verb and a noun in the corpus which have at most two tokens between them (except punctuation and conjunctions), taking into account the possibility for a free word order.

(2) Filter the LVC candidates – remove candidates with low frequency in the corpus as their statistical measures are unreliable.

(3) Analyse the LVC candidates based on the occurrences of the verb–noun pairs in the corpus in order to determine the variations in their form and word order, the possible modifiers and external elements separating the LVCs components.

(4) Apply the trained classifier to classify the LVC candidates – distinguish LVCs from other categories of phrases: (a) other types of decomposable MWEs – where the verb is a content verb, or non-decomposable MWEs – idioms; and (b) collocations which are not MWEs.

4.5. Evaluation

We performed two-step evaluation: cross-validation on the training set and evaluation on new test data. Table 2 shows the results from the 10-fold cross-validation on the training set.

Algorithm	Precision	Recall	F_1
J48	0.739	0.794	0.766
RandomTree	0.710	0.741	0.725

Table 2: Comparison of the 10-fold cross-validation using different algorithms (J48 and RandomTree).

	Main method			Main method & Filtering		
Algorithm	Precision	Recall	F_1	Precision	Recall	F_1
J48	0.776	0.830	0.802	0.794	0.810	0.802
RandomTree	0.482	0.820	0.607	0.684	0.800	0.737

Table 3: Results from the application of the method on the test dataset of LVC candidates.

Table 3 presents the results from the application of the method on the test dataset of 200 unique LVC candidates extracted from the BulNC. The evaluation is lemma-based and each candidate is counted once (and not with its frequency in the corpus). The table provides a comparison between the main method with two different decision tree algorithms (J48 and RandomTree) and the main method supplemented with filtering of LVC candidates. The filtering included: excluding candidates with low association measure below the threshold of 2.0 (which are unlikely to be MWEs); and excluding candidates with verbs that are not light verbs (which are not in the list of 105 verbs, see Section 4.1.) and/or nouns that do not belong to the predicative categories (as defined by the semantic primitives). Performing filtering prior to machine learning ensured that a large number of improbable LVC candidates were excluded before the application of the ML method which does not perform well for low frequency candidates due to their unreliable statistical measures.

5. Discussion

The results reported in this paper are comparable to the performance of similar methods for other languages, such as the one developed by Nagy et al. (2013), while outperforming others which do not take into account semantic features, such as the method reported by Vincze et al. (2011). This emphasises the importance of semantic features such as the semantic primitive of the noun. Experiments with reducing the group of predicative noun primitives to only *noun.act* and *noun.event* show that these are the most significant primitives and although the recall falls (0.790 with J48), the precision improves (0.782 with J48).

As a large proportion of the training data were extracted from BulNet (a lexical database), they do not cover all types of MWEs, and LVCs in particular, in terms of usage variety. One of the most important results at this stage is that we obtained a reliable set of Bulgarian LVCs extracted (semi-)automatically from different language resources, using linguistic heuristics. The list of light verbs we compiled is more comprehensive than the usually adopted lists and reflects the diversity and productivity of LVCs. Moreover, the training set was extended to include LVCs from the BulNC (from unrestricted texts), which significantly improved the results (compared to $F_1 = 0.494$ trained purely on instances from BulNet and using J48). This is expected since the data from the corpus reflect the usage of LVCs while BulNet also includes rare and untypical LVCs which have low frequency in the corpus and hence – yield unreliable statistical measures. The inclusion of more real-life examples is expected to improve further the performance of the method.

Although LVCs fall into a small and clear-cut set of syntactic structures, they also are syntactically flexible as they allow intervening elements, as well as various transformations such as passivisation, nominalisation, etc., which makes their discovery in unrestricted text much more challenging. The results reported in existing literature and in this paper show that although LVCs seem to be a relatively well-defined class, their semantic traits are not specific enough to distinguish them with high precision from free phrases, collocations and idioms. These facts point to the necessity to include more contextual and semantic features and to use the LVCs' traits in a more productive way in engineering the ML features.

References

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

Butt, M. (2003). The Light Verb Jungle. In *Harvard Working Paper in Linguistics*, volume 9, pages 1–49. John Benjamins.

Chen, W., Bonial, C., and Palmer, M. (2015). English Light Verb Construction Identification Using Lexical Knowledge. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2368–2374. AAAI Press.

Cinkovà, S. and Kolářová, V. (2005). Nouns as Components of Support Verb Constructions in the Prague Dependency Treebank. In Šimkovà, M., Ed., *Insight into Slovak and Czech Corpus Linguistics*, pages 113–139. Veda.

de Cruys, T. V. and Moirón, B. V. (2007). Semantics-based Multiword Expression Extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Prague, June 2007*, pages 25–32. ACL.

Fazly, A. and Stevenson, S. (2007). Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures. In *Proceedings of the Workshop on MWEs, Prague, Czech Republic, June, 2007*, pages 9–16. ACL.

Francopoulo, G. (2013). *Lexical Markup Framework*. John Wiley and Sons.

Goldberg, A. (2003). Words by Default: The Persian Complex Predicate Construction. In Francis, E. J. and Michaelis, L. A., Eds., *Mismatch: Form-Function Incongruity and the Architecture of Grammar*, volume 22, pages 117–146. Stanford: CSLI Publications.

Gradečak-Erdeljić, T. and Brdar, M. (2012). Constructional Meaning of Verbo–nominal Constructions in English and Croatian. *Suvremena lingvistika*, 38.

Gralinski, F., Savary, A., Czerepowicka, M., and Makowiecki, F. (2010). Computational Lexicography of Multi-Word Units. How Efficient Can It Be? In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications. Coling, August 2010*, pages 2–10.

Gregoire, N. (2010). DuELME: A Dutch Electronic Lexicon of Multiword Expressions. *Language Resources and Evaluation*, 44:23–39.

Gurrutxaga, A. and Alegria, I. (2011). Automatic Extraction of NV Expressions in Basque: Basic Issues on Cooccurrence Techniques. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), Portland, Oregon, USA*, pages 2–7. ACL.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. In *SIGKDD Explorations*, volume 11.

Jackendoff, R. (1974). A Deep Structure Projection Rule. *Linguistic Inquiry*, 5(4):481–505.

Jackendoff, R. (1997). The Architecture of the Language Faculty. *Computational Linguistics*, 24.

Koeva, S. and Genov, A. (2011). Bulgarian Language Processing Chain. In *Proceeding to The Integration of multilingual resources and tools in Web applications Workshop in conjunction with GSCL 2011*. University of Hamburg.

Koeva, S., Stoyanova, I., Todorova, M., and Leseva, S. (2016). Semi-automatic compilation of the dictionary of bulgarian multiword expressions. In *Proceedings of the Workshop on Lexicographic Resources for Human Language Technology (GLOBALEX 2016), Portorož, Slovenia, 24 May 2016*, pages 86–95.

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

Koeva, S. (2014a). The Bulgarian National Corpus in the context of World Theory and Practice (Balgarskiyat natsionalen korpus v konteksta na svetovnata teoriya i praktika). In Koeva, S., Ed., *Language Resources and Technologies for Bulgarian (Ezikovi resursi i tehnologii za balgarski)*, pages 29–52. Marin Drinov Academic Publishing House.

Koeva, S. (2014b). WordNet and BulNet (Wordnet i BulNet). In Koeva, S., Ed., *Language Resources and Technologies for Bulgarian (Ezikovi resursi i tehnologii za balgarski)*, pages 154–173. Marin Drinov Academic Publishing House.

Melčuk, I. (1998). Collocations and Lexical Functions. In Cowie, P., Ed., *Phraseology. Theory, Analysis, and Applications*, pages 23–53. Oxford: Clarendon Press.

Miller, G. (1998). Nouns in WordNet. In *WordNet: An Electronic Lexical Database*, pages 24–45. MIT Press.

Mudraya, O., Piao, S. S., Rayson, P., Sharoff, S., Babych, B., and L L.).

Nagy, I., Vincze, V., and Farkas, R. (2013). Full-coverage Identification of English Light Verb Constructions. In *Proceedings of the International Joint Conference on Natural Language Processing, Nagoya, Japan, 14-18 October 2013*, pages 329–337. University of Hamburg.

Nunberg, G., Sag, I., and Wasow, T. (1994). Idioms. In Everson, S., Ed., *Language*, pages 491–538. Cambridge University Press.

Palmer, M., Guildea, D., and Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2007). Ontonotes: a unified relational semantic representation. In *First IEEE International Conference on Semantic Computing (ICSC-07), Irvine, CA*, pages 517–526. IEEE.

Przybyszewski, S. (2015). Some Problems with the Description of Paradigms of Polish Verbal Multiword Units. In E. Gutierrez Rubio, M. Falkowska, E. K. M. S. W., Ed., *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)*, volume 18, pages 213–223.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, pages 1–15. Springer-Verlag.

Samardžić, T. and Merlo, P. (2010). Cross-lingual Variation of Light Verb Constructions: Using Parallel Corpora and Automatic Alignment for Linguistic Research. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, Uppsala, Sweden*, pages 52–60. ACL.

Samardžić, T. (2008). Light verb constructions In English Language and Literature Studies. In *Structures across Cultures*, pages 59–73. Faculty of Philology, Belgrade.

Stevenson, S., Fazly, A., and North, R. (2004). Statistical Measures of the Semi-Productivity of Light Verb Constructions. In *Proceedings of the Workshop on MWEs, Barcelona, Spain, July, 2004*, pages 1–8. ACL.

Tan, Y. F., Kan, M.-Y., and Cui, H. (2006). Extending Corpus-based Identification of Light Verb Constructions Using a Supervised Learning Framework. In *Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts, Trento, Italy*, pages 49–56. ACL.

Tu, Y. and Roth, D. (2011). Learning English Light Verb Constructions: Contextual or Statistical. In *Proceedings of MWE 2011, Portland, Oregon, USA*, pages 31–39. ACL.

Urešová, Z., Bejček, E., and Hajič, J. (2016). Inherently Pronominal Verbs in Czech: Description and Conversion Based on Treebank. pages 78–83.

Villavicencio, A., Copestake, A., Waldron, B., and Lambeau, F. (2004). The Lexical Encoding of MWEs. In Tanaka, T., A. Villavicencio, F. B., and Korhonen, A., Eds., *Proceedings of the ACL 2004 workshop on multiword expressions: Integrating processing. Barcelona, Spain*, pages 80–87.

Vincze, V., Nagy, I., and Berend, G. (2011). Detecting Noun Compounds and Light Verb Constructions: A Contrastive Study. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), Portland, Oregon, USA*, pages 116–121. ACL.

Vincze, V., Savary, A., Candito, M., and Ramisch, C. (2016). *Annotation Guidelines for the PARSEME Shared Task on Automatic Detection of Verbal Multiword Expressions. Version 5.0*. `http://typo.uni-konstanz.de/parseme/images/shared-task/guidelines/PARSEME-ST-annotation-guidelines-v5.pdf`.

Wittenberg, E., Jackendoff, R., Kuperberg, G., Paczynski, M., Snedeker, J., and Wiese, H. (2014). The Mental Representation and Processing of Light Verbs. In Bachrach, A., Roy, I., and Stockall, L., Eds., *Structuring the Argument. Multidisciplinary Research on Verb Argument Structure*, pages 61–80. John Benjamins.

HR4EU – Using Language Resources in Computer Aided Language Learning

Daša Farkaš
Faculty of Humanities and Social Sciences, Zagreb
dberovic@ffzg.hr

Matea Filko
Faculty of Humanities and Social Sciences, Zagreb
msrebaci@ffzg.hr

Marko Tadić
Faculty of Humanities and Social Sciences, Zagreb
mtadic@ffzg.hr

Abstract

In this paper we present the HR4EU – web portal for e-learning of Croatian language. The web portal offers a new method of computer aided language learning (CALL) by encouraging language learners to use different language resources available for Croatian: corpora, inflectional and derivational morphological lexicons, treebank, Wordnet, etc. Apart from the previously developed language resources, the new ones are created in order to further facilitate the learning of Croatian language.

We will focus on the usage of the treebank annotated at syntactic and semantic level in the CALL and describe the new HR4EU sub-corpus of the Croatian Dependency Treebank (HOBS). The HR4EU sub-corpus consists of approx. 550 sentences, which are manually annotated on syntactic and semantic role level according to the specifications used for the HOBS. The syntactic and the semantic structure of the sentence can be visualized as a dependency tree via the SynSem Visualizer. The visualization of the syntactic and the semantic structure of sentences will help users to produce syntactically and semantically correct sentences on their own.

1. Introduction

In this paper we present the HR4EU – web portal for e-learning of Croatian. The HR4EU is the first portal which offers Croatian language courses which are free-of-charge and developed by language professionals. Moreover, the HR4EU also integrates bidirectional interaction with some of the language resources for Croatian developed previously. For the purpose of this paper, we will focus on the interaction between the HR4EU and one of these language resources – the Croatian Dependency Treebank (HOBS) and show how language resources, developed primarily for NLP tasks, can be used as a valuable tool in the computer aided language learning.

The paper is structured as follows: in Chapter 2, we briefly present the HR4EU portal and its relation to Croatian language resources. In Chapter 3, we describe two layers of the Croatian Dependency Treebank: the syntactic and the semantic layer, as well as the SynSem Visualizer, a newly developed tool for visualization of dependency trees. Chapter 4 is dedicated to the HR4EU sub-corpus of the HOBS, which was developed to facilitate the understanding of Croatian syntax and semantic relations between a verb and its arguments to the HR4EU users. The paper ends with the concluding remarks.

2. HR4EU – web portal for e-learning of Croatian

Since Croatian is a language with relatively small number of speakers, its presence on the web is limited. A few e-learning sites which offer users the possibility to learn Croatian are expensive (e.g. E-learning

course of Croatian as a second and foreign language - HiT-1[1]), developed by non-native speakers of Croatian (e.g. Surface languages[2]) or present the learning material in static manner avoiding the usage of existing language technologies (e.g. Easy Croatian[3], Basic Croatian[4]). With the HR4EU portal we aim to bridge this gap and develop a modern e-learning system which integrates bidirectional interaction with previously developed language resources (LRs). This e-learning system is developed by linguists, which are also native speakers and have experience in building LRs. Moreover, the great efforts were made to make this portal visually attractive to users.

The HR4EU portal is divided into four sections:

a) **Courses**, where users can find three general courses: beginner, intermediate and advanced, as well as two specialized courses: Croatian for students and Croatian for business users. Courses are equipped with interactive lessons, quizzes, dictionary, grammar books, tasks for practicing writing skills, etc. For the purpose of courses at the HR4EU portal we have recorded more than 1.600 audio tracks and approx. 200 illustrations, in order to obtain their interactivity and multimodality.

b) **Language Resources**, the section which includes description of LRs for Croatian language and a short video for each LR that is used as an additional learning tool throughout the courses. Short video tutorials provide users with the introduction to the particular resource (cf. 2.1.)

c) **About Croatia**, providing the cultural context for learning Croatian via nine interactive maps presenting most important cities, events, famous Croats, landscapes, cultural heritage, gastronomy and ethnology, etc.

d) **Living in Croatia**, offering useful information to foreigners in Croatia, e.g. the list of important state institutions.

The first section, Courses, is developed in Moodle, an open source e-learning platform, which provides teachers or course developers with numerous tools and activities that can be used in e-learning course (e.g. interactive lessons, quizzes with multiple question types, dictionary, books, and assignments). However, the Moodle itself is a "robust"[5] system, which was restructured and modified both visually and functionally in order to become interactive, attractive and effective e-learning tool. Several new plugins and possibilities were introduced, e.g. HINT and NOTE buttons (Figure 1), which provide users with help when they answer the question incorrectly, or with the additional information about words or grammar used in question if they answer the question correctly.

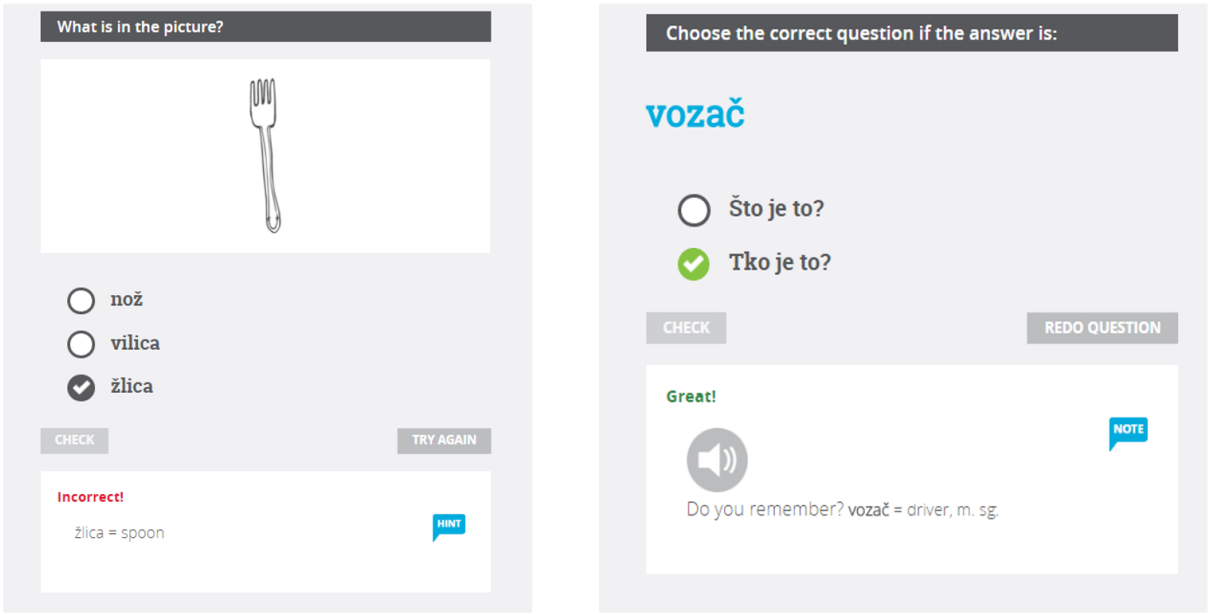

Figure 1: Hint and Note buttons

[1] http://www.unizg.hr/homepage/learn-croatian/e-learning-course-of-croatian/
[2] http://www.surfacelanguages.com/
[3] http://www.easy-croatian.com/
[4] http://basic-croatian.blogspot.hr/
[5] https://docs.moodle.org/31/en/About_Moodle

The other three sections are developed in WordPress, but with the identical visual theme as used in the Moodle part. The portal contains multimodal content: audio files, video tutorials, interactive maps, pictures, links to the various sites about Croatia, etc., since the quality approaches to the computer aided language learning have to use all the possibilities that are offered by multimodal technology. This is why the HR4EU portal introduces language learners with various language resources for Croatian and their usability when learning a new language.

2.1. Language resources at HR4EU

As stated before, apart from the interactive and multimodal content, the HR4EU portal introduces language learners to the usability of language resources in the (computer aided) language learning. Thus, the one part of the HR4EU portal is dedicated solely to Croatian language resources. There, the users can find out more about language resources which can be particularly useful for them and which are therefore used as a helping tool throughout the courses. Each resource, namely Croatian National Corpus[6] (216,8 million words; Tadić, 1996), Croatian Morphological Lexicon[7] (3,9 million word forms; Tadić and Fulgosi 2003, Tadić 2006), Croatian Wordnet[8] (23.122 synsets with 47.906 lexical units; Raffaelli et al. 2008; Oliver et al. 2015), CroDeriV[9] – a morphological database of Croatian verbs (14.491 verbs; Šojat et al. 2013), and Croatian Dependency Treebank[10] (4.000 sentences; Tadić 2007, Agić et al. 2014), is provided with a brief description, link to the respective search interface and a short video tutorial. Short video tutorials were made especially for the Croatian language learners, since the most of them have never seen or used Croatian LRs, or LRs in general, before.

The lessons and quizzes are designed to encourage the users to use LRs, e.g., to find the appropriate word form or lemma in Croatian Morphological Lexicon or to learn semantically related words via lexical hierarchies or synsets in Croatian WordNet or derivationally related words in verbal derivational database CroDeriV.

Moreover, this system is designed in a way that language learners can also be helpful in improvement of existing LRs. Learners' activity will be used to enhance and enlarge existing LRs by tracking their activity yielding empty results, and adding them to the respective resources. Furthermore, users' answers in *Practice your writing skills* tasks will be used to build the new LR, the corpus of Croatian as a second language. This corpus will be particularly useful to language teaching specialists, because it will offer a possibility to extract morphological and syntactic errors of users.

However, some of the existing LRs weren't helpful for language users in their primary shape, because the language material they contain is too complex for language learners which have just begun to learn Croatian. Nevertheless, they served as a model for building a new LR on syntactic and semantic level which can be helpful even to the learners at a beginner level. In the following chapters we thus present the syntactic and semantic layer of the Croatian Dependency Treebank and the application of this model to the corpus of sentences from the HR4EU courses.

3. Croatian Dependency Treebank – HOBS

The Croatian Dependency Treebank (Tadić 2007, Agić et al. 2014) is a corpus of approx. 4.500 sentences extracted from the Croatia Weekly 100kw, the newspaper sub-corpus of the Croatian National Corpus. The sentences are manually tagged according to the modified Prague Dependency Treebank specification for annotation at the analytical level.[11] The part of the HOBS (approx. 3.500 sentences) is also manually tagged with semantic roles, according to the specification developed for the Croatian semantic role labelling. The SynSem Visualizer enables the queries across this 3.500 sentences which are annotated both on the syntactic and semantic level. Here we will briefly describe the two abovementioned layers and the queries enabled by the SynSem Visualizer.

[6] hnk.ffzg.hr

[7] hml.ffzg.hr

[8] crown.ffzg.hr

[9] croderiv.ffzg.hr

[10] hobs.ffzg.hr

[11] https://ufal.mff.cuni.cz/pdt2.0/

3.1. HOBS – syntactic layer

There are two slightly different manually annotated versions of HOBS at syntactic level. The first version is annotated in complete accordance to the Prague Dependency Treebank annotation guidelines for annotating at the analytic level (cf. Appendix 1, footnote 6). The second version is annotated with the modified PDT specification (cf. Appendix 1), which is adjusted to the syntactic structures of Croatian language. This specifically pertains to the different annotation of dependent clauses, which has also improved the parsing results. This second version is freely available for search via SynSem Visualizer (cf. 3.4.) and further annotated with semantic role labels.

3.2. HOBS – semantic layer

Semantic role labelling is essential for many NLP tasks, especially when it comes to information extraction. It is a logical step immediately after the resources on the syntactic level have been built.[12] Semantic layer of HOBS presents first steps towards automatic semantic role labelling in Croatian.

In order to build a training set for the automatic semantic role labelling of Croatian texts, we first had to design a tagset for Croatian semantic role labelling. Since the manually tagged sentences will be used as a training set for the automatic semantic role labelling system, we had to be careful when it comes to our specification: the labels had to be verb-independent and of a limited number. The initial set of tags was revised during the manual annotation, i.e. the tags which proved to be very frequent and distinctive enough from the existing ones were added to the tagset. The final set consists of seventeen tags followed by the examples of sentences in which these tags should be used. (cf. Appendix 2 for the SRL specification for Croatian).

Tags can be divided in two groups: first group comprises verbal arguments, and second group adjuncts, mainly different types of adverbials and adverbial and attribute clauses. We have decided to include adjuncts into our SRL specification because they often give more specific and detailed information about the described event and can be very useful later in e.g. information extraction tasks.

3.3. SynSem Visualizer

The two above presented layers of the HOBS are encoded in the CONLL format. Although this format is useful to most of the professional linguists, it is not useful to the users of the HR4EU portal, and even to the non-computational linguists. This results in the lower usability and visibility of this language resource, so we decided to develop a visualizer which will enable the search and the hierarchical representation of the syntactic and semantic structure of Croatian sentences.

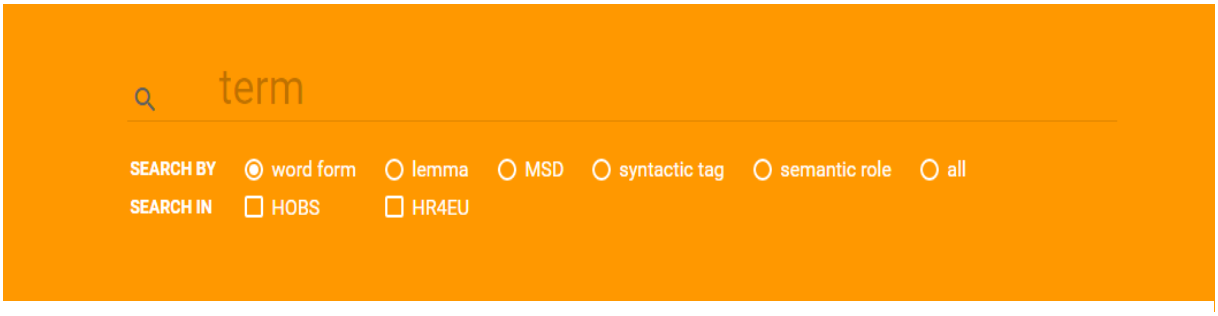

Figure 2: SynSem Visualizer search interface

The SynSem Visualizer enables the queries across the Croatian Dependency Treebank on the syntactic and semantic level. It is developed as a database-driven web application, and written in Django, a widely used Python framework. It enables the graphical representation of the hierarchical sentence

[12] There are several LRs for Croatian at syntactic level: above mentioned Croatian Dependency Treebank, SETimes.HR Treebank (Agić and Ljubešić 2014), Universal Dependencies Treebank (http://universaldependencies.org/) and a dependency parser (Agić 2012).

structure (cf. Figures 3, 4 for the representation of the hierarchical sentence structures via SynSem Visualizer).

The HOBS and the HR4EU corpora can be searched independently, or they can be both searched at the same time. The SynSem Visualizer enables search by word form, lemma, morphosyntactic tag, syntactic tag or semantic role (cf. Figure 2). Users are provided with the MSD, syntactic and semantic specifications on the website, so they can easily manage their searches.

4. HR4EU sub-corpus

Although the language resources are mainly used in NLP tasks, they can also be used in the computer aided language learning. This is why they are an essential part of courses at the HR4EU portal. However, language learners often don't have linguistic background, so some of the resources have to be adjusted to their needs, or even new resources, which can be used both in NLP and CALL have to be built. For the purpose of the HR4EU portal we have developed a sub-corpus of the HOBS which consists of approximately 550 syntactically and semantically annotated sentences used in the Croatian language courses available at the HR4EU web portal. These sentences are manually annotated on both syntactic and semantic level according to the model and specification used for HOBS and presented in previous chapter.

The syntactic structure of sentences in the HR4EU sub-corpus is not as complicated as the syntactic structure of the newspaper sentences contained in HOBS. It is adjusted to the beginner level users of Croatian.[13] However, even the syntax of the simple sentences of the morphologically rich language as Croatian can be challenging to the speakers of other, especially non-Slavic languages. Thus, the graphical representation of syntactic structure of sentences used in courses could improve the users' understanding of different grammatical relations. The example of the syntactic tree from the HR4EU corpus is presented in Figure 3.

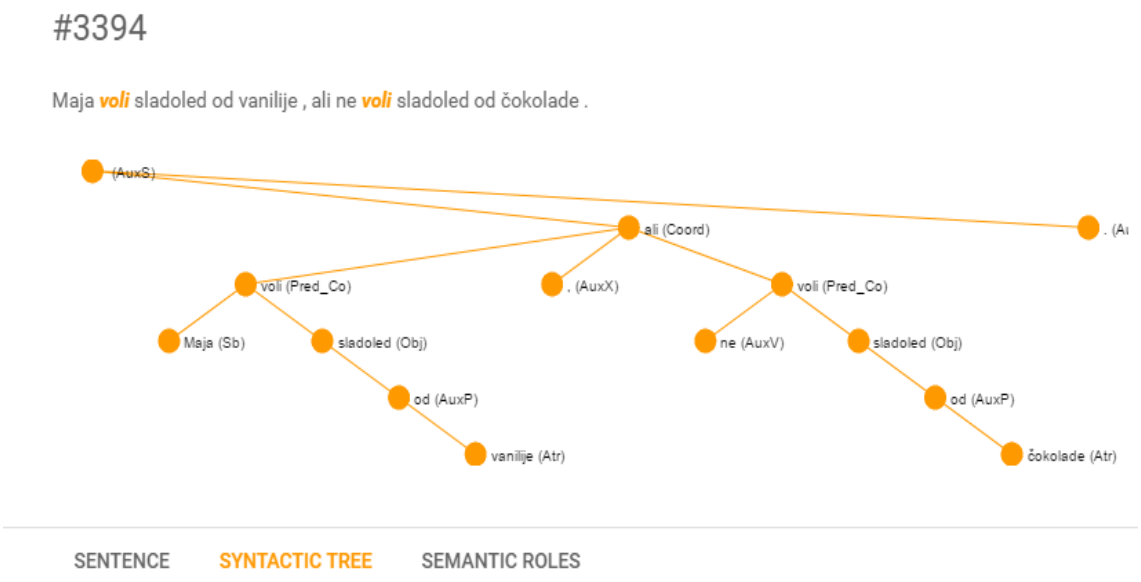

Figure 3: The example of the syntactic tree - HR4EU sub-corpus

Maja voli sladoled od vanilije, ali ne voli sladoled od čokolade.
Maja-NOMsg like-PRES3sg ice-cream-ACCsg from vanilla-GENsg,
but no like-PRES3sg ice-cream-ACCsg from chocolate-GENsg
'Maja likes vanilla ice-cream, but she doesn't like chocolate ice-cream.'

[13] The vocabulary used in the HR4EU courses is mostly based upon the corpora frequency lists, and the syntactic structure of sentences follows the grammar content which is presented in the lesson.

The challenges for the decoding of the syntactic structure stated above can be expanded to the understanding of the role of the verb arguments as well. Graphical representation of the semantic structure of the sentence can, therefore, improve the learners' accurate interpretation of semantic roles of the verb arguments, i.e. they can easily see "Who did What to Whom" (Palmer, 2010). The understanding of these basic relations in the sentence is crucial for the foreign language learners, and along with the understanding of the syntactic structure helps them to build correct sentences on their own. The example of the semantic tree of the same sentence from the HR4EU corpus is presented in Figure 4.

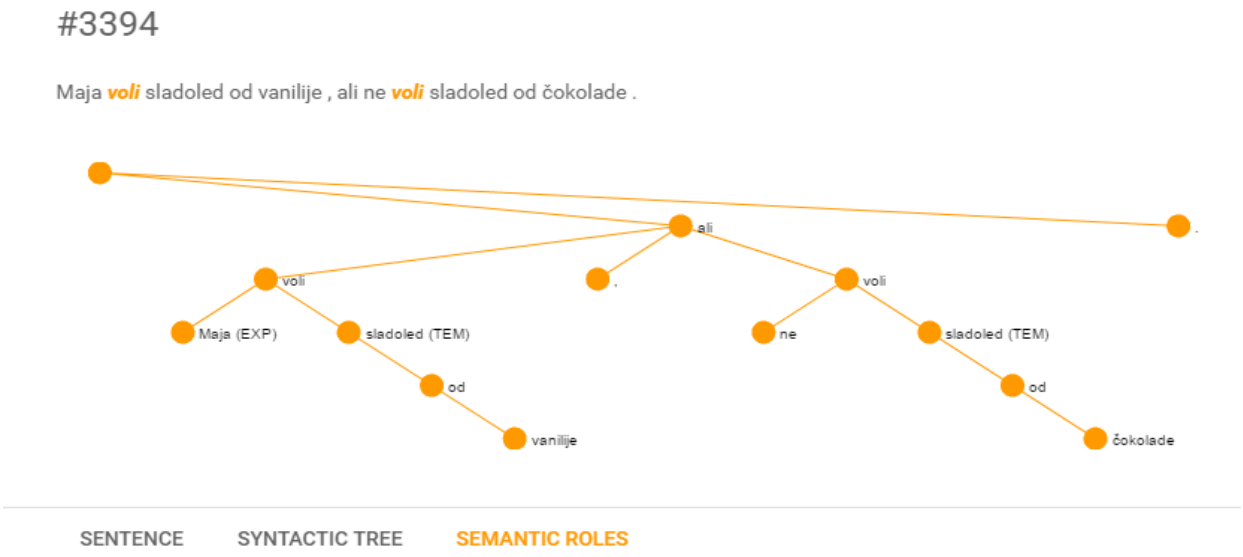

Figure 4: The example of the semantic tree - HR4EU sub-corpus (cf. Figure 3 for glosses and translation)

5. Conclusion

In this paper we have presented the HR4EU – web portal for e-learning of Croatian and its bidirectional relation to language resources for Croatian. The HR4EU is the first completely free-of charge portal with e-courses of Croatian language developed by language professionals. Moreover, it is the first portal which takes advantages of the language technologies in the computer aided language learning. The interrelation between the HR4EU and the one of the existing LRs for Croatian – the Croatian Dependency Treebank – is described in this paper.

The resources like HOBS are most commonly used in NLP tasks, e.g. parsing (syntactic layer) and automatic semantic role labelling (semantic layer). However, they can be extremely useful in the CALL as well, but they have to be modified to serve the language learners' purposes. We have applied the same model used for the HOBS to less complex sentences used in the HR4EU courses to help our users to understand the syntactic and semantic structure of Croatian sentences. They can, moreover, use this resource if they are not sure of the verbal frame, e.g. if they don't know which preposition they should use with the particular verb to express the particular argument. The other language resources, especially different morphological lexica, can also be helpful in the CALL, and further stress the importance of language technologies in the computer aided language learning. The application of LRs to other domains, along with NLP, extends their visibility and usability.

Acknowledgement

This paper and hereby presented project are fully supported by European Union, European Social Fund, under the project grant HR.3.2.01.-0037 Mrežni portal za online učenje hrvatskoga jezika HR4EU.

References

Agić, Ž. (2012). *Pristupi ovisnosnom parsanju hrvatskih tekstova.* PhD thesis, University of Zagreb, Faculty of Humanities and Social Sciences.

Agić, Ž., Berović, D., Merkler, D., Tadić, M. (2014). Croatian Dependency Treebank 2.0: New Annotation Guidelines for Improved Parsing. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 2313-2319.

Agić, Ž., Ljubešić, N. (2014). The SETimes.HR Linguistically Annotated Corpus of Croatian. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 1724-1727.

Oliver, A., Šojat, K., Srebačić, M. (2015). Enlarging the Croatian Wordnet with WN-Toolkit and CroDeriV. In Angelova, G., Bontcheva, K., Mitkov, R. (eds.) Proceedings of the International Conference Recent Advances in Natural Language Processing, Hissar, Bulgaria: BAS, pp. 480-487.

Palmer, M., Gildea, D., D. Nianwen, X. (2010). *Semantic Role Labelling*. Morgan and Claypool Publishers.

Raffaelli, I., Bekavac, B., Agić, Ž., Tadić, M. (2008). Building Croatian Wordnet. In: Tanács, Attila; Csendes, Dóra; Vincze, Veronika; Fellbaum, Christianne; Vossen, Piek (eds.) Proceedings of the Fourth Global WordNet Conference 2008, Szeged: GWC, pp. 349-359.

Šojat, K., Srebačić, M. and Štefanec, V. (2013). CroDeriV and the Morphological Analysis of Croatian Verb, *Suvremena lingvistika* 39/(75): 75-96.

Tadić, Marko (1996). Računalna obradba hrvatskoga i nacionalni korpus. *Suvremena lingvistika* 41-42.

Tadić, M., Fulgosi, S. (2003). Building the Croatian Morphological Lexicon. In: *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages* (Budapest 2003), ACL, pp. 41-46.

Tadić, M. (2006). Croatian Lemmatization Server. In*:* Vulchanova, M. D., Koeva, S., Krapova, I., Vulchanov, V. (Eds.). *Formal Approaches to South Slavic and Balkan Languages*. Sofia: Bulgarian Academy of Sciences, pp. 140-146.

Tadić, M (2007). Building the Croatian Dependency Treebank: The Initial Stages. *Suvremena lingvistika* 63: 85-92.

Appendix 1 – Syntactic specification used in HOBS – syntactic level

Table 1: List of analytical functions from PDT

afun	**Description**
Adv	adverbial
Apos	apposition
Atr	attribute
Atv	complement hung on a non-verb. element
AtvV	complement hung on a verb
AuxC	subordinating conjunction
AuxG	other graphic symbols, not terminal
AuxK	terminal punctuation of a sentence
AuxO	emotional, rhythmic particles
AuxP	preposition
AuxR	reflexive passive
AuxT	reflexivum tantum
AuxV	auxiliary verb
AuxX	comma
AuxY	some particles
AuxZ	emphasizing words
Coord	coord. node
ExD	ellipsis
Obj	object
Pnom	nominal predicate
Pred	predicate
Sb	subject

Table 2: Syntactic sub-tagset for subordinate clause annotation in HOBS

tag	**Description**
Sub_Atr	attribute
Sub_Adv	adverbial
Sub_Obj	object
Sub_Pred	predicate
Sub_Sb	subject

Table 3: Syntactic sub-tags for adverbial clause subclassification

Sub Adv	**Adverbial clause**
Sub_Adv_loc	local
Sub_Adv_temp	temporal
Sub_Adv_mod	modal
Sub_Adv_caus	causative
Sub_Adv_cons	consequential
Sub_Adv_fin	final
Sub_Adv_cond	conditional

Appendix 2 – SRL specification used in HOBS – semantic level

Table 1. SRL Tagset for Croatian

Tag	Role	Example
AGT_anim	agent_animate	**Marko** je zapjevao. ***Marko*** *started to sing.*
AGT_inanim	agent_inanimate	**SDP** je izjavio... ***SDP*** *has declared...(political party)* Atomski **udar** je uništio grad. *Nuclear* ***attack*** *has destroyed the city.*
PAT	patient	Oduzeli su **im** ljudska prava. *They took* ***them*** *the human rights.* Marko je udario **loptu**. ***Marko*** *hit the ball.*
EXP	experiencer	**Krešo** uživa u hrenu i šunki. ***Krešo*** *enjoys eating ham and horseradish.*
BEN	beneficiary	Krešo uči **studente** matematiku. *Krešo teaches* ***students*** *math.*
RES	result	On je postao **Španjolac**. *He became* ***Spanish***.
PART	participant	**Hrvatska** je potpisala ugovor **s Rusijom**. ***Croatia*** *signed the contract with* ***Russia***.
TEM	theme	On je naučio **španjolski**. *He learned* ***Spanish***. **Lopta** se odbila od zida. *The* ***ball*** *bounced from the wall.*
INS	instrument	**Ključem** je otvorio vrata. *He opened the door with the* ***key***.
SRC	source	Svjetlost dolazi **od Sunca**. *The light comes from the* ***Sun***.
QUAN	quantity	Povećali su trošarine **od 10 do 20 posto**. *They increased the exice duties* ***from 10 to 20 percent***.
TMP	time	Putovanje je trajalo **od 6 ujutro do 7 navečer**. *The trip lasted* ***from 6 a.m. until 7 p.m.*** **Prošle godine** donijeli su odluku o ukidanju ropstva. *They decided to abolish slavery* ***last year***.
LOC	location	Delegacija je otputovala **u Sarajevo**. *The delegation has departed* ***to Sarajevo***.
LOC_ap	abstract location	Hrvatska se prima **u Partnerstvo za mir**. *Croatia is becoming a member of* ***Partnership for Peace***.
CAU/FIN	cause, intention	Krešo je umoran **od naporna rada**. *Krešo is tired* ***of hard work***. Povisili su poreze **radi smanjenja deficita**. *They increased taxes* ***to reduce deficite***.
MNR	manner	**Veselo** su potpisali sporazum. *They* ***cheerfully*** *signed an agreement.* Igrao je **nepošteno**. *He played* ***unfairly***.
ATR	attribute clause	Vidio sam dijete **koje** se igralo. *I saw a child* ***that*** *was playing.*

SynTags - Web Interface for Syntactic and Semantic Annotation

Atanas Atanasov
Sofia University "St. Kliment Ohridski"
atanasow@gmail.com

Abstract

This paper presents a web tool for syntactic and semantic annotation and two of its applications. It gives the linguists the possibility to work with corpora and syntactic and semantic frames in XML format without having computer skills. The system is OS and platform independent and could be used both online and offline.

1. Introduction

This paper presents an online system for syntactic and semantic annotation. Initially it was developed as a support tool for student theses in syntax and thereafter it was upgraded and used as data processing tool in linguistic research of the prepositional phrases in predicative position in contemporary Bulgarian.

The core of the system is written in XML - it is built on the basis of XForms. In order to be accessible online, it is installed on eXist-db server (http://exist-db.org/), which supports XForms, XQuery etc. It is created using a modified version of AgenceXML's XSLTforms (http://www.agencexml.com/), which allows browsers to manipulate XForms and has a client-side implementation, preventing server overloading.

The main advantage of the system is the possibility for the user to fill and save all the data (i.e. to create complicated annotated corpora; to present the argument structure of the predicates and the semantic and subcategorization frame) in xml file without knowing xml or having computer skills at all.

Compared to other existing annotation tools (like Hydra or Chooser for example) SynTags offers a different approach. Unlike Hydra (http://dcl.bas.bg/hydra/), which is a system for browsing and editing wordnet data, SynTags serves a completely different purpose - it uses predefined synsets (that cannot be edited directly from the user interface) and the main goal is to provide an environment for manual presentation of the argument structure of the predicates and the syntactic realization and the semantic properties of these arguments.

It has more in common with Chooser (http://dcl.bas.bg/chooser-2/), but SynTags is not that powerful in semantic mark-up of elements (it is not connected to the whole wordnet database) as the aim is not the creation of semantically annotated corpus, in which all the words are connected to the corresponding synset. The annotation level in the sentences represents the argument positions, so it is more similar to the one used in the Berkeley FrameNet annotation tool (https://framenet.icsi.berkeley.edu/fndrupal/annotation_tool), but SynTags also provides an option to add and edit the framenet data as well as the subcategorization frames (both discussed more detailed in chapters 3.2 and 3.3).

2. Application in student theses

The first beta version of the software was tested as a tool for creation of student theses and it was implemented in e-learning system giving the students the possibility to work online on every browser without need to install XML editors or any other apps. The interface of this first working version looks like this:

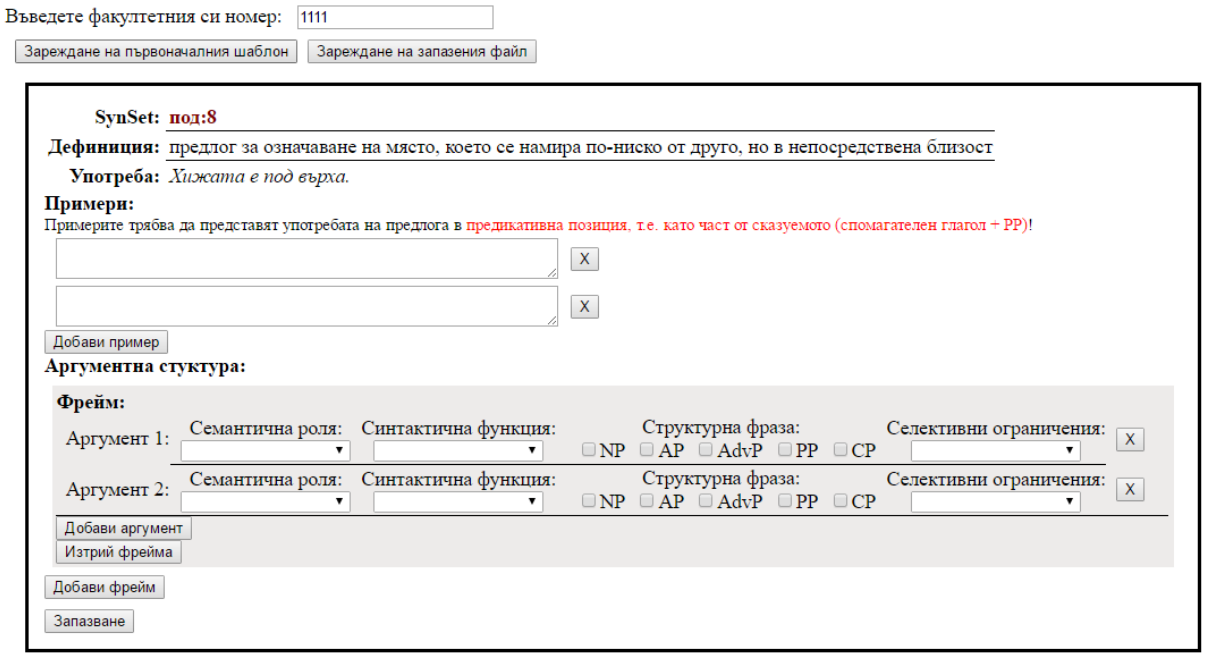

Figure 1

The students have to excerpt the corresponding examples from the Bulgarian National Corpus (BNC) and try to present their argument structure and the semantic relations between the arguments of the predicate. All the data loaded and saved in the browser is actually in XML format, visible for the professors, but not for the students. All the data visible in the web Xform will be discussed in details in the next chapter.

3. Application as an annotation tool for PPs in predicative position

After the successful try-out, the system was upgraded with more complex functions, the most important of which is the possibility to annotate the examples and to bind their arguments with the syntactic and semantic frames. Here is a screenshot of the main interface:

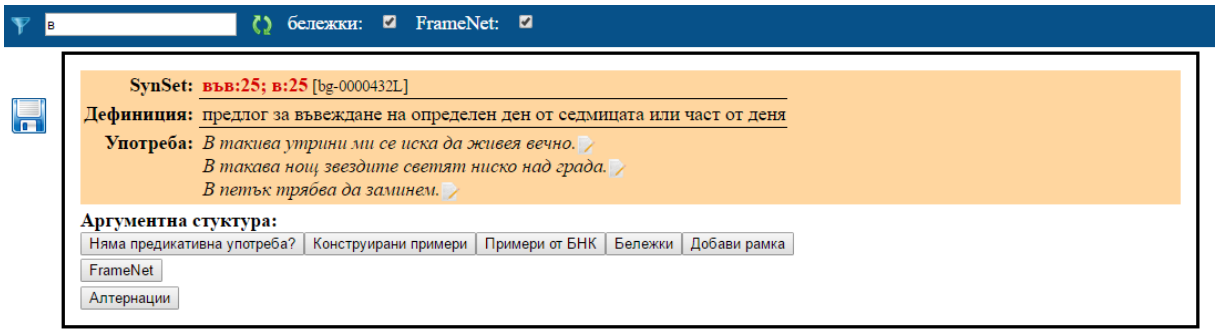

Figure 2

The header of each Synset contains the main information from the Bulgarian Wordnet - the literals (with the corresponding sense number), the ID, the definition and the usage given in BulNet (where it's applicable). This information is manually copied from BulNet 3.0 (http://dcl.bas.bg/bulnet/) in a pre-process XML file. The user has the possibility to make some personal notes for every one of the usage examples.

Below the Wordnet block there is an "Argument structure" section containing several other options: "No predicative usage", "Constructed examples", "Examples from Bulgarian National Corpus", "Notes", "Add frame", "FrameNet" and "Alternations".

The first one is used for those prepositions that could not be used as predicatives. When pressed it deletes all the information already entered (if there's any) and eliminates all the other options in the Synset. The Synset window is colored red and only one textbox that remains in it is about free text description for the reason why the preposition cannot be a part of a predicate (for example 'only attributive usage'). Also there's an option to add some additional notes. The delete button next to the textbox reverts the Synset interface to the initial state - the user can again add and edit examples, frames etc.

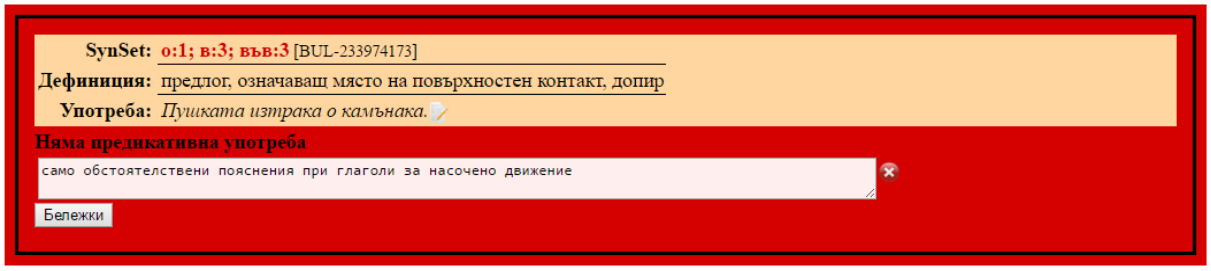

Figure 3

3.1. Corpora annotation

In order to provide evidence that the analysis is correct, every particular sense should be illustrated by as many examples as possible. In this case it is advised (following the principles stated in Koeva et al. 2008) that at least five examples should be given for every Synset. Pressing one of the next two buttons ('constructed examples' and 'examples from BNC') triggers an interactive text area, where after the example is entered, it could be annotated with the help of the buttons above the box.

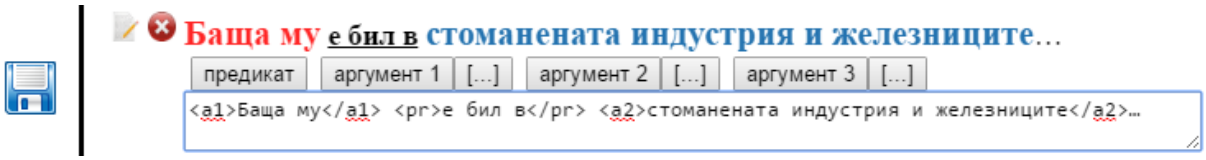

Figure 4

When a part of the text is selected, pressing a button wraps the selection in XML tag. The first one puts <pr>...</pr>, which marks the predicate (in this specific usage it actually marks a part of the predicate - the auxiliary verb and the preposition, interpreted here as the core of the predicate). The next buttons mark the arguments if they are explicit (e.g. <a1>Той</a1>) or their position if they are implicit (e.g. <a1>[...]</a1>). Any changes in the textbox appear above in real time presenting the data formatted in different style depending on the corresponding XML annotation. The styled text is interactive - clicking on it shows or hides the edit window for the example. Also saving the document makes all the edit text boxes disappear.

For each example there's also an option to add or delete a note or the whole element.

The actual data is saved in the xml file in an <example>...</example> element, so the previous example is coded in the following format:

<example><a1>Баща му</a1> <pr>е бил в</pr> <a2>стоманената индустрия и железниците</a2>…</example>,

creating this way a syntactically and semantically annotated corpus.

3.2. Argument structure

When the examples are ready the next button adds the subcategorization frames. Here the linguist has the possibility to add or remove frames and to add or delete arguments in the frames. The number of the arguments depends on the semantic properties of the predicate - they should vary from zero to three.

In the system there are two semantic levels of presentation. The first one is more generalized and it follows the well-known semantic roles in Role and Reference Grammar (Van Valin et al. 1997),

where the relations between the predicates and their arguments are presented with the following scheme:

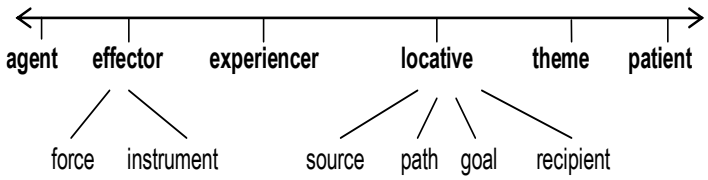

Figure 5

As all the frames in a Synset refer to the same definition they need to have the same number of arguments and in the most cases their arguments should have the same semantic roles. If the semantic roles are different, it means the definition should be divided into parts presenting more accurately the semantics of the predicate.

The other semantic level is directly connected to the Princeton Wordnet synsets and their Bulgarian correlates. The main goal is to present the exact selective restrictions of the core elements. In other words, this is an attempt for more precise description of the semantic properties of the arguments. In a separate XML file are extracted the main concepts from the Wordnet hierarchy - about 65 synsets considered as a “skeleton” and they are dynamic – every time when an argument requiring a synset not included in the file is found, it has to be added. This file is published and accessible online as a HTML page and the user could go to this interactive web page (fig. 6) for a quick reference of the hierarchical relations, definitions, examples and so on. Of course, if more detailed information is needed, the linguist should check the official BulNet/WordNet website.

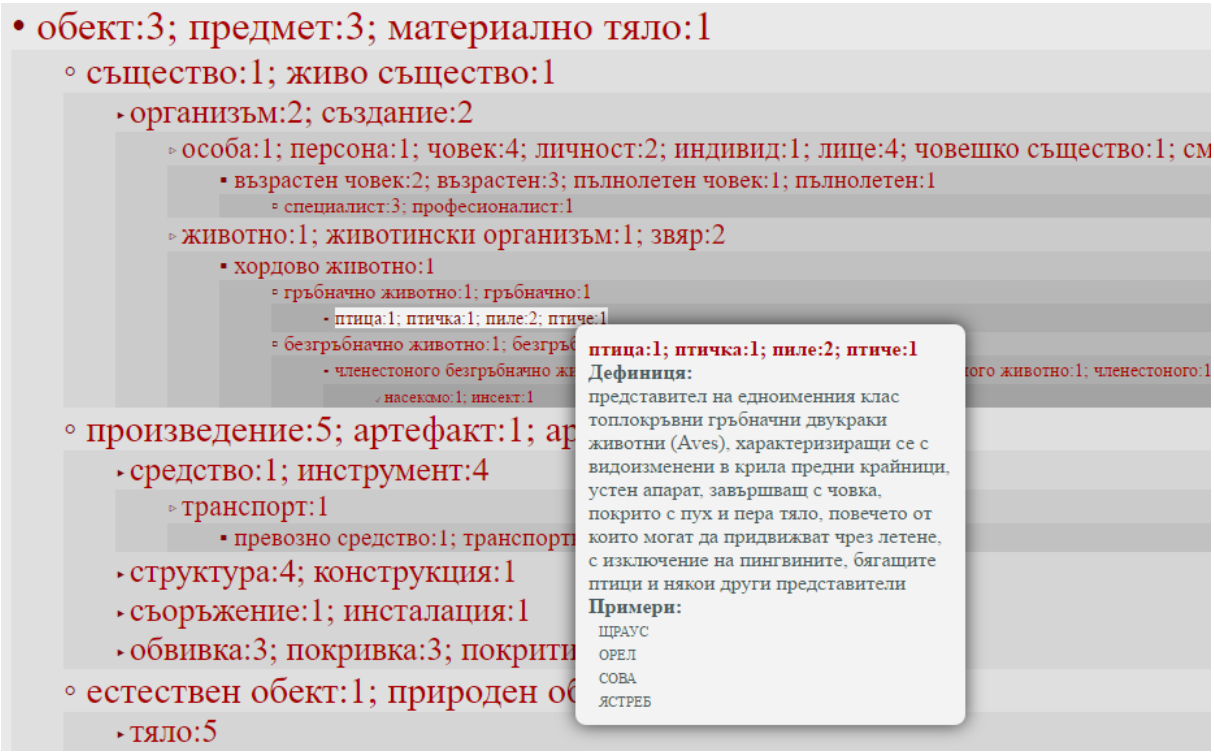

Figure 6

The syntactic function of the arguments also should be presented, following the traditional classification: subject, predicative (not an argument of the predicate, as it is a part of it together with the preposition and the copula - it is considered to be an argument of the preposition itself), direct and indirect object, adjunct and small clause.

The following figure illustrates a sysnset’s argument structure presentation:

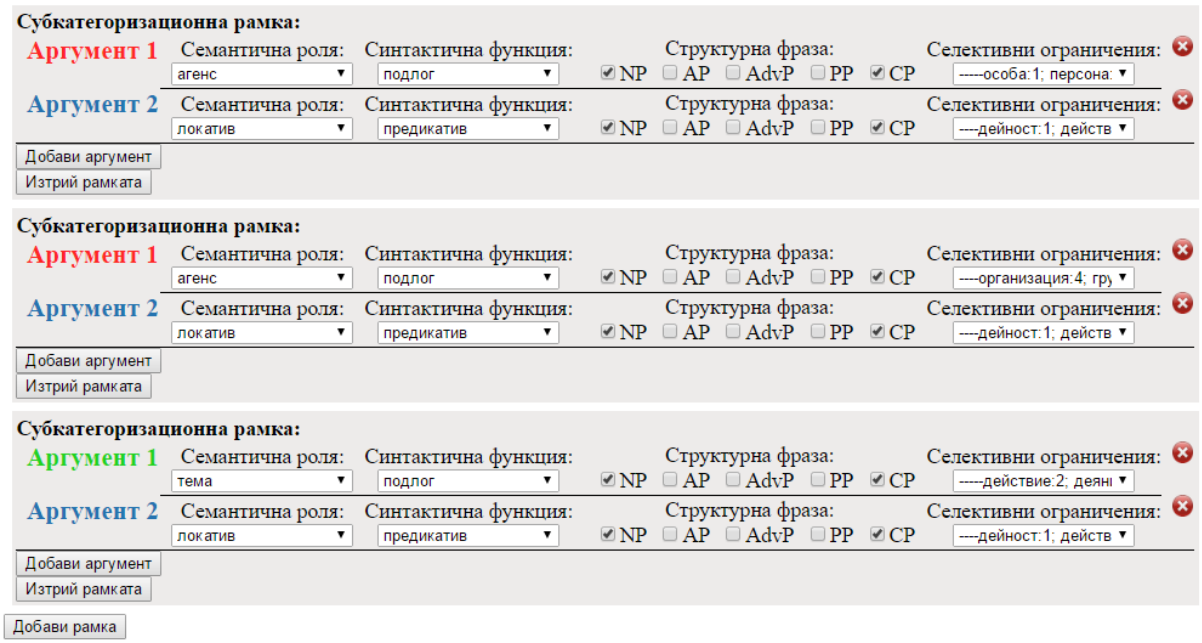

Figure 7

The number of the frames in a Synset depends mainly on the selective restrictions of the arguments. The predicate - representing a real situation - should have a fixed number of core elements, but they could have different realization - syntactic or semantic. The main phrase type (NP, AP, AdvP, PP or CP) have to be chosen for each element. If an argument with the same meaning could be realized as more than one type of structure phrase, there is an option all of them to be presented in the same frame. For example the subject in Bulgarian sentence always could be expressed with NP or with CP and the locative adjuncts can be expressed with AdvP or PP. Since this alternations are consistent there is no need adding a second frame - it is enough to check both in the same frame.

There should be more than one frame when the selective restrictions belong to different categories, e.g. the predicates that require a person (physical entity) or an organization (abstract entity) in the same argument position.

3.3. FrameNet

In the system there is also an option to connect the predicate meaning (the synset definition) with the corresponding frame from Berkeley FrameNet Project. As the Bulgarian FrameNet is still in working stage and is not accessible yet, this binding for now is only manual and presents the only the core frame elements. The frame and the frame core elements names and definitions and translated in Bulgarian and aligned with the original data (FameNet 1.6).

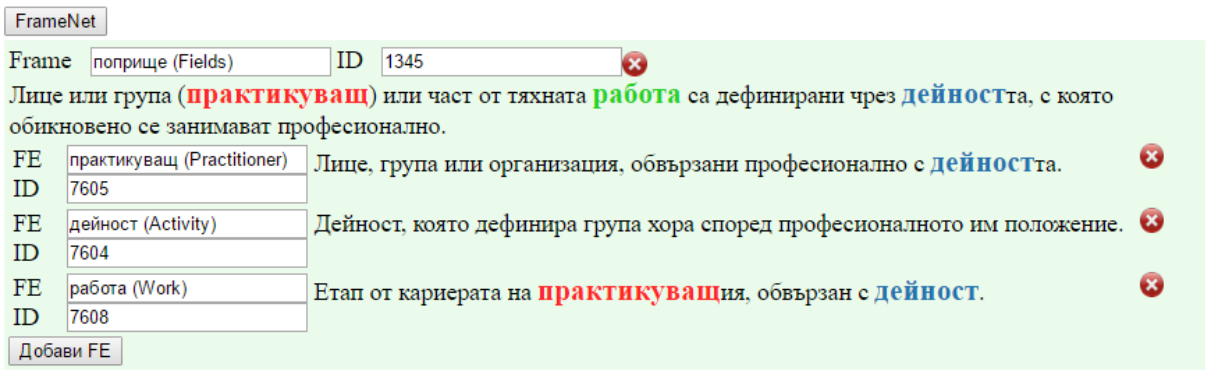

Figure 8

All the frame elements, the arguments in the subcategorization frames and their realization in the examples are bound to each other and styled the same way (cf. fig. 4, 7 & 8).

In this particular application (for description of predicative PPs) another experimental function is available – presenting the possible substitutions of the auxiliary verb with a lexical verb or the PP with AdvP.

3.4. Filtering and search

At the top of the web page there are several filter options. It is possible to search for a literal and display only the synsets containing it, to show or hide the user notes and also to activate or stop the FrameNet functionality.

4. Advantages

This are the main pluses of the SynTags system:

- *Universal tool for corpora and syntax frame annotation*. The system can be easily modified (for now only by changing a few lines in the source code) in order to satisfy the needs of any particular linguistic task related to corpus annotation or semantic and syntactic presentation.
- *Easy collaboration*. The tool can be used by many developers working on the same xml database.
- *Easy access*. It is platform and operating system independent - the only requirement is a current web browser.
- *Comfortable user interface*. Not special programming knowledge is required, so everybody could use the tool without having advanced computer skills.
- *Online and offline usage*. The tool is accessible online, but it also could be easily installed locally on a free open source eXist-db server.

5. What's next?

- *Optimization for large data processing*. The current version has some issues concerning the processing of very big files, so in the future the efforts will be concentrated mainly on improving the stability and the speed of the system.
- *Adding a more complex search and filter functionality*. Now the system can search only by xPath expressions - it is planned to improve this functionality by adding a full xQuery support.
- *Adding options for advanced user settings*. SynTags currently works with predefined XML and DTD files – the next step will be to give the uses the opportunity to modify them partially from the user interface.
- *Implementation of full FrameNet support*. It was mentioned that the FrameNet data could be entered manually. The future plans include full implementation of FrameNet 1.6 in the system database.

References

Baker, C. F., Fillmore, C. J. and Lowe, J. B. (2006). The Berkeley FrameNet Project. In *Proceedings of the 17th International Conference on Computational Linguistics, Volume 1*. Association for Computational Linguistics, 1998, pp. 86-90.

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

Fillmore, C. J. and Baker, C. F. (2009). A Frames Approach to Semantic Description. In B. Heine and H. Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press.

Koeva, S. (2010). *Balgarskiyat FrameNet 2010*. Sofia, 2010.

Koeva, S., Vlahova, R., Dekova, R., Nestorova, P. and Atanasov, A. (2008). *Balgarskiyat FrameNet. Semantiko-sintaktichen rechnik na balgarskiya ezik*. sastavitel Svetla Koeva, Sofia, 2008.

Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson C.R. and Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California.

Van Valin, R. D., Jr. and LaPolla, R. J. (2007). *Syntax: Structure, Meaning and Function*. Cambridge: Cambridge University Press.

Resources

Bulgarian National Corpus: http://search.dcl.bas.bg/

BulNet: http://dcl.bas.bg/bulnet/

Chooser: http://dcl.bas.bg/chooser-2/

FrameNet: https://framenet.icsi.berkeley.edu/fndrupal/

FrameNet Annotation Tool: https://framenet.icsi.berkeley.edu/fndrupal/annotation_tool

Hydra: http://dcl.bas.bg/hydra/

WordNet: http://wordnet.princeton.edu/

Finding Good Answers in Online Forums: Community Question Answering for Bulgarian

Tsvetomila Mihaylova, Ivan Koychev
FMI, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria

Preslav Nakov
ALT Research Group, Qatar Computing Research Institute, HBKU, Doha, Qatar

Ivelina Nikolova
IICT, Bulgarian Academy of Sciences, Sofia, Bulgaria

Abstract

Community Question Answering (CQA) is a form of question answering that is getting increasingly popular as a research direction recently. Given a question posted in an online community forum and the thread of answers to it, a common formulation of the task is to rank automatically the answers, so that the good ones are ranked higher than the bad ones. Despite the vast research in CQA for English, very little attention has been paid to other languages. To bridge this gap, here we present our method for Community Question Answering in Bulgarian. We create annotated training and testing datasets for Bulgarian, and we further explore the applicability of machine translation for reusing English CQA data for building a Bulgarian system. The evaluation results show improvement over the baseline and can serve as a basis for further research.

1. Introduction

With the ever growing user-generated content, it is becoming increasingly harder and time-consuming for users to find valuable information. This is especially true for web forums, where a question can generate a thread of hundreds of answers. Thus, there is a need to filter the answer thread and to present to the user the most relevant answers first, i.e., to rerank the answers in a forum not chronologically as they naturally occur, but based on how well they answer the original forum question.

Community Question Answering (CQA) is a special case of the more general problem of Question Answering (QA), which has been an active research area for years (Webber and Webb, 2010). The TREC conference has had QA tasks since 1999 (Voorhees, 1999), focusing on various aspects of the problem.

CQA is a topic with growing research interest. The specifics of CQA include user-generated content in free text, without necessarily following strict rules. Important difference between the traditional content and the user-generated content is that the latter shows higher variance in quality (Agichtein et al., 2008; Ahn et al., 2013; Baltadzhieva and Chrupała, 2015). This problem is well-studied for English, e.g., there has been a shared tasks for CQA at SemEval-2015 (Nakov et al., 2015) and SemEval-2016 (Nakov et al., 2016b).

However, the field is not explored for Bulgarian yet. To bridge this gap, in this paper, we experiment with CQA data from the biggest online forum in Bulgaria - BGMamma.[1] We create annotated training and testing datasets for Bulgarian. While annotating data for testing is not that hard, annotating a lot of data for training is a rather time-consuming task. Therefore, we annotate small sets for training and testing, and we translate them from Bulgarian to English, and we train a system that works for English. In order to make a larger training set, we use additional publicly available annotated data for English. Then we apply domain adaptation to combine the small translated in-domain data with the large out-of-domain data.

[1] BG Mamma: `http://www.bg-mamma.com/`

The remainder of the paper is organized as follows: Section 2. introduces the research in the field related to our task. Section 3. describes the features and the method used for classification. Section 4. contains the result of our experiments, where we compare the results from using different training sets and different feature groups. In section 5., we conclude and we point to possible directions for future work.

2. Related Work

Community Question Answering is a topic of great research interest. For example, there has been a shared task for CQA in SemEval-2015 (Nakov et al., 2015) and SemEval-2016 (Nakov et al., 2016b) editions. Various approaches for CQA have been explored by the systems in those competitions. For example, Belinkov et al. (2015) used vectors of the question and of the comment, metadata features, and text-based similarities. Nicosia et al. (2015) used similarity measures, URLs in the comment text and statistics about the user profile: number of good, bad, and potentially useful comments. In our system, we use similar features to those systems, such as the number of posts by the same user in the thread, topic model-based feature, special words, URLs, word embeddings of the question and comment, metadata features, and text similarities.

Other approaches for CQA, used in the top systems in SemEval-2016 Task 3 on CQA (Nakov et al., 2016b) include troll user features by (Mihaylov et al., 2015a; Mihaylov et al., 2015b; Mihaylov and Nakov, 2016a), fine-tuned word embeddings as in the SemanticZ system (Mihaylov and Nakov, 2016b), and PMI-based goodness polarity lexicons as in the PMI-cool system (Balchev et al., 2016), as well as sentiment polarity features (Nicosia et al., 2015). Other systems are based on a deep learning architecture, e.g., as in the MTE-NN system (Guzmán et al., 2016a; Guzmán et al., 2016b; Nakov et al., 2016a), which borrowed an entire neural network framework and architecture from previous work on machine translation evaluation (Guzmán et al., 2015). We do not currently use such kinds of features in our system.

The current study is based on our previous work for Task 3 of SemEval-2016[2] (Mihaylova et al., 2016). We rank a set of comments according to their relevance to a question. The task is solved with a classification approach where each question-comment pair is tagged as *Good* or *Bad* and the rank of a comment is a function of the probability that the pair is *Good*. The following feature groups are considered during classification: metadata, semantic, lexical, credibility and user features. In the current research for Bulgarian, we only apply metadata, semantic and lexical features. The user features are still applicable and will be included in future experiments.

Using machine translation for solving tasks in languages different from English is used in various domains. For example, Mohammad et al. (2016) translated text from Arabic to English for sentiment analysis. Balahur and Turchi (2014) used machine translation for sentiment analysis of tweets. In cross-lingual and multilingual information retrieval, machine translation is often applied to the query, to the results or to both (PothulaSujatha and Dhavachelvan, 2011). The experiments show that using machine translation in such settings yields meaningful results and could be applied for translating the target documents to languages rich in resources such as English as we do in the current study. Translation is used for CQA by Zhou et al. (2011; 2012) who experiment with machine translation for question retrieval.

The problem with no sufficient data available for classification in a target domain is solved by using domain adaptation (Daume III, 2007). For the current study, we do not have sufficient data for Bulgarian, so we are using domain adaptation for including publicily available annotated data to expand the training set.

In our research of CQA for Bulgarian, we use machine translation to translate the collected training and testing target data from Bulgarian to English. We further use domain adaptation to expand the insufficient training set. After that, we use an existing pipeline developed for CQA for English, and we run it on the translated data together with the additional training data.

[2]SemEval-2016, Task 3: `http://alt.qcri.org/semeval2016/task3/`

3. Method

The purpose of our experiments is to see whether a system that performed well for CQA in English (ranked 1st on the main Substask C of Task 3 on SemEval 2016), would deliver strong results for Bulgarian too.

The experimental environment is built on top of the framework developed for CQA in English (Mihaylova et al., 2016). It solves the task of ranking comments with respect to their relevance to a given question. The pipeline includes variety of features, some of which are extracted from external data sources, i.e., the Qatar Living (QL) forum.

In this work, we use all features of the system for English, excluding user statistics from the QL forum, pointwise mutual information (PMI) and credibility features. Those features are still relevant for our study in Bulgarian, and we plan to add them in future experiments.

3.1. Features

In our experiments with data in Bulgarian, we use lexical, semantic and metadata features in the way they are described in (Mihaylova et al., 2016).

Metadata Features

These features present observations for the thread and for the comment structure and properties.

- Whether the comment is written by the author of the question.
- Comment's rank in the thread.
- Ratio of the comment length to the question length (in terms of number of tokens).
- Number and order of comments from the same user in a particular thread.
- Presence and the number of URLs in the question and in the comment.

Lexical Features

For obtaining the lexical features, the question and the comment texts are annotated with GATE (Cunningham et al., 2002; Cunningham et al., 2011).

- Number of each question word (*where*, *who*, *what*, etc.) in the question and in the comment text.
- Whether the comment contains an answer to a wh-question (*where*, *who*, *what*, etc.). For example, if the question contains *where* and the comment contains an address or location, this is considered as a response to such a question.
- Number of verbs, nouns, pronouns, and adjectives in the question and in the comment.
- Number of question marks and question words in the question and in the comment.
- Comment contains smileys, currency units, e-mails, phone numbers, only laughter, "thank you" phrases, personal opinions, disagreement.
- Number of misspelled words and offensive words from a dictionary.
- Dictionary of unigram and bigram occurrences across the classes.
- Lexical similarity between a question and a comment using *SimHash* (Sadowski and Levin, 2007).
- Level of readability and complexity of the text (Aluisio et al., 2010). The standard readability measures include Automated Readability Index, Coleman-Liau Index, Flesch Reading Ease, Gunning Fog Index, Flesch-Kincaid Grade Level, LIX, SMOG grade. We also use statistics about the average number of words per sentence in the comment or in the question, and type-to-token ratios.
- Average number of words per sentence in the comment or in the question.
- Type-to-token ratios in the question and in the comment.

Semantic Features

This group of features aims to find the similarity of the question and of the comment meaning.

- Topic Modeling with Mallet (McCallum, 2002) is used for training of 100 topics from questions and comments from the QL training data.
- Word Embeddings trained with Word2Vec(Mikolov et al., 2013) on the QL forum data.
- Cosine distances between the text of the question and of the comment: between vectors of all words, between different parts of speech (nouns, verbs, adjectives). The cosine distance was calculated between the sum of the embeddings of all words in the question and in the comment text.

3.2. Domain Adaptation

Since the training data we prepared for Bulgarian is relatively small, we expanded the training set by using domain adaptation (Daume III, 2007). The idea of domain adaptation is, when insufficient training data exists for a target domain, to use available data from another domain as an additional training set, called the source set. Suppose we have a set of features we can extract from the target and from the source data. We can perform domain adaptation using equation 1.

$$\Phi^s = \langle x, x, 0 \rangle, \Phi^t = \langle x, 0, x \rangle \tag{1}$$

In this equation, x is the vector of features and $0 = \langle 0, 0, ... \rangle$ is the zero vector. In order to put the features extracted from the target and from the source domain into one classifier, we expand the feature space and we use three parts for the feature vector. The first part contains features extracted from both the target and the source sets. The second part contains features extracted from the source set only. The third part contains features extracted from the target set only.

For the domain adaptation, we construct a training set that contains two subsets: the features from Qatar Living as a source set formatted as Φ^s, and the features from the BG-Mamma training set as a target set (Φ^t). The test set is the test set from BG-Mamma, again formatted as a target set (Φ^t).

3.3. Classifier

The task of ordering the comments with respect to the question is a ranking problem. It aims to order the comments according to their relevance as a response to the given question. It is important that the *Good* comments are ranked higher than the *Bad* ones. We approach the problem as a classification task. Each example is a question-answer pair, and the following feature vector is formed for the examples:

$$v_{q_1}, ..., v_{q_k}, v_{c_1}, ..., v_{c_k}, f_1, ..., f_m \tag{2}$$

where v_q and v_c are the k-sized vectors of the word embeddings of the question and of the comment, and f is a vector of the non-embedding features.

We used LibSVM (Chang and Lin, 2011; Hsu et al., 2003) for the classification. The results from the classification give probability for each class. The probability for the *Good* class is used as a ranking score for the question-comment pair. We experimented with different kernels, but the best results are achieved with an RBF kernel. Thus, we only report results when using an RBF kernel.

4. Experiments

4.1. Data

The data for the current study is collected from the largest online forum in Bulgaria - BGMamma. The forum has topics in various categories, each topic is a thread with comments from different users. In order to prepare the data in a format suitable for our task, we first selected topics with titles containing 'въпрос' (the Bulgarian word for 'question'). The first comment in the topic is considered as a question. The next five comments in the topic are considered as answers to this question. We annotated manually 80 questions with the first 5 answers from the thread for each of them, i.e., 400 question-comment pairs. Each answer is annotated as either *Good* (it gives a direct answer to the given question) or *Bad* (it does not give a direct answer to the question).

We split the annotated questions into training and test set. The training set has 50 questions with 5 answers each, i.e., 250 question-answer pairs. The test set has 30 questions with 5 answers each, i.e., 150 question-answer pairs. Table 1 shows more detailed statistics for the training and test sets.

After the data was annotated, the topic categories, question texts, question subjects and comment texts were translated from Bulgarian to English with the *Microsoft Translation API*.[3] As an additional training data we use the Train-1 set from SemEval-2016 Task 3. From them, we took only the comments on positions from 1 to 5 in the forum thread. The difference of the SemEval labeling of the comments is that they also include *Potentially Useful* labels. We consider those labels *Bad*.

[3] https://www.microsoft.com/en-us/translator/translatorapi.aspx

	Questions Count	Comments Count	Good Comments	Bad Comments
Test Set from BG-Mamma	30	150	49	110
Train Set from BG-Mamma	50	250	84	166
Additional Train Set from QL	1411	7055	3021	4034

Table 1: Statistics about the data sets.

Table 2 shows an example of a question thread and its comments from the forum (the translation in English is presented in Table 3). It also illustrates the difference between the relevant vs. non-relevant answers. The comments marked as *Good* give a direct answer to the asked question. The answers marked as *Bad* can be for example a 'Thank you' statement, an irrelevant comment, could be a new question or a reply to some question in the comment thread rather than to the original question.

Question Subject		**Question Text**
Въпроси относно камина Ерато		Моля тези от вас, които имат такава камина да се включат с отговор на няколко въпроса: 1. Запалихте ли вече камините. Първоначално само вечер ли? 2. Какви настройки сте направили? 3. Имате ли някакво ръководство? 4. Какви пелети ползвате? Предварително благодаря на всички :lol:
Comments		
Position	**Relevance**	**Comment Text**
1	Good	Имам Пони9 на Ерато. Днес я запалих за 2 часа. Пелетите са български от Разлог, но имаме още 2-3 торби от тях. За тази зима сме поръчали етрополски пелети 2,5 тона. Засега не сме настройвали нищо. Миналата година бяхме настроили да се включва сутрин в 5:30, после по някое време се изключваше, пак се включваше и т.н., но не помня подробности. Имам книжка с инструкции.
2	Bad	Благодаря за отговора. Използвали ли сте някакъв екорецим?
3	Bad	Нямам идея какво е това. :shock:
4	Bad	Бухахаха :D ей такива смешки стават, когато пишеш през телефона. Имах предвид ЕКО РЕЖИМ :hug: Междудругото вашата камина, когато достигне определена темп спира ли работа?
5	Bad	Първата зима спираше, но после от фирмата, откъдето я купихме, й промениха настройките и сега не спира. Проблемът със самоизключването бе, че трябва температурата да падне с 2 градуса под зададената, за да се включи. По този начин се получаваха големи температурни амплитуди.

Table 2: Example of question and comments from the forum in Bulgarian.

The feature extraction pipeline includes word embeddings trained on the Qatar Living[4] forum with Word2Vec (Mikolov et al., 2013). This data was provided as an unannotated data for SemEval-2016 Task 3 and it includes 200,000 questions and 2 million comments. The vectors were trained with Gensim (Řehůřek and Sojka, 2010).

4.2. Experiment Setup

We train our models on several different training sets and we measure which one achieves best results when testing on the test set. The first one is the training set from the Bulgarian forum, translated to English. The second one is the training set from the QL forum - questions and comments originally written in English.

[4]Qatar Living: `http://www.qatarliving.com/forum`

Question Subject		**Question Text**
Questions about fireplace Erato		Please those of you who have that fireplace to get involved with the answer to a few questions: 1. You lit the fireplaces. Initially only night? 2. What settings have you done? 3. Do you have any guidance? 4. What pellets you use? thanks in advance to all : lol:
Comments		
Position	**Relevance**	**Comment Text**
1	Good	I have Poni9 on Erato. Today I lit it for 2 a.m. pellets are Bulgarian from Razlog, but we still have 2-3 bags of them. For this winter we ordered etropole pellets 2.5 tons. so far, we haven't set up any thing. Last year we were set up to turn on at 5:30 in the morning, then at some time is excluded, it still included, etc, but I don't remember the details. Have book with instructions.
2	Bad	Thanks for the reply. Have you used any ekorecim?
3	Bad	I have no idea what that is. :shock:
4	Bad	Buhahaha :D These jokes become, when you write in the phone. I meant the ECO MODE : hug: by the way your fireplace when it reaches a certain temp stops work?
5	Bad	The first winter, but then stopped by the company where we bought it, I changed the settings and now I can't stop. problem with turning itself off, you need the temperature to drop to 2 degrees below the set to be turned on. Thus received large temperature amplitudes.

Table 3: Example of question and comments from the previous table, translated to English.

To construct the third training set, we use domain adaptation as described in (Daume III, 2007). The details were described in Section 3.2. above. As the source set, we use the training set from QL. The training set in Bulgarian is included as target in the training data and the test set on the Bulgarian forum is also processed as target. Comparison of the results is shown in Section 4.4..

The baseline is calculated by ranking the comment with respect to their chronological position in the question-comment thread. The first posted comment in the thread has position 1, the second one has position 2 etc. For the baseline, *1/comment position* is used as the ranking score for the comment in the thread.

4.3. Evaluation

In Section 4., we present the results of our experiments. We first compare the test results when the classifier was trained on different training sets with all features. After that, we compare different feature groups to find the most important ones for our task.

As a main evaluation measure, we use Mean Average Precision at 5 (MAP@5), as we are interested in the most useful answers appearing at the top of the result. As an additional measure, we use accuracy. When a ranked result is given, *MAP* (formula 3) calculates the mean of the average precision for each query (question) q. Average precision *AveP(q)* takes the precision at each position for the given question (i.e., for the first 1 result, for the first 2 results) and then takes the average of those values (precision *P(k)* measures the ratio of the positively classified - *Good* examples to all given examples up to position k). Finally, *accuracy* measures the ratio of the number of correctly classified examples to the total number of examples.

$$MAP@5 = \sum_{q=1}^{Q} AveP(q)/Q, AveP@5 \sum_{k=1}^{5} P(k)/5 \quad (3)$$

4.4. Results

Table 4 shows the results when training the classifier with different training sets: from BG-Mamma, from Qatar Living (including all comments and only the first 5 comments), and using domain adaptation. For this comparison, all features are used. The results show that only using the data from Qatar Living as a training set does not yield very good results. The best results are achieved when the training set from BG-Mamma is used, as well as when domain adaptation is applied. For further experiments, we use only the training set from BG-Mamma, as is yields comparable results to domain adaptation, but training the classifier is faster because of the smaller feature space and the smaller set size.

Training Set	MAP	Accuracy (%)
Baseline	70.76	–
Training data from BG-Mamma	90.39	78.67
Training data from Qatar Living - all data	83.67	73.33
Training data from Qatar Living - only answers up to 5	87.06	74.67
Domain adaptation - data from Qatar living and BG-Mamma	90.39	79.33

Table 4: Comparison of different training sets. The shown results are trained on the corresponding training set with all features.

For our next experiments, we wanted to determine which groups of features are significant for the results and which ones are not. Tables 5 and 6 show experiments with different features groups. The classifier for those experiments was trained on the training set from BG-Mamma, translated to English. The compared feature groups contain logically related features, described in Section 3.1.. The results show that the most significant features are the word embeddings and the metadata of question and comment. Those feature groups improve the baseline when used on their own and the result is lower when they are excluded from the feature set. In our previous work (Mihaylova et al., 2016), the word embeddings and the metadata also turned out to be among the most significant features.

The described experiments show that the approach of using machine translation and a pipeline prepared for English works well. The achieved results significantly improve the baseline.

	MAP	Accuracy (%)
All Features	90.39	78.67
only Semantic features / Word embeddings	81.06	67.33
only Metadata features / Thread structure	76.89	72.00
only Metadata features / Comment structure	74.28	67.33
only Semantic features / Cosine distances	66.42	67.33
only Metadata features / URLs	68.15	67.33
only Lexical features / Question words	59.13	67.33
only Lexical features / Parts of speech	69.67	66.00

Table 5: Experiments with different feature groups. The results are obtained when only the features from the given group are used for classification.

	MAP	Accuracy (%)
All Features	90.39	78.67
All − Semantic features / Word embeddings	86.22	72.67
All − Metadata features / Thread structure	85.83	78.00
All − Metadata features / Comment structure	90.11	77.33
All − Semantic features / Cosine distances	88.72	76.67
All − Metadata features / URLs	90.39	78.67
All − Lexical features / Question words	90.39	74.00
All − Lexical features / Parts of speech	90.94	76.67

Table 6: Experiments with different feature groups. The results are obtained when all the features are used, excluding the features in the given group.

5. Conclusion and Future Work

We have presented our research on Community Question Answering for Bulgarian using machine translation. First, we translate the text of the questions and answers from Bulgarian to English and then run a pipeline tested for English with the translated texts. The experiments show that this approach works very well and the improvement over the baseline is comparable to the one used in the original system tested in English. The results show that this approach can be used for further work in CQA for Bulgarian.

In future work, we plan to try ideas from the top systems that participated in SemEval-2016 Task 3 on CQA (Nakov et al., 2016b). In particular, we want to incorporate several rich knowledge sources, e.g., as in the SUper Team system (Mihaylova et al., 2016), including troll user features as inspired by (Mihaylov et al., 2015a; Mihaylov et al., 2015b; Mihaylov and Nakov, 2016a), fine-tuned word embeddings as in the SemanticZ system (Mihaylov and Nakov, 2016b), and PMI-based goodness polarity lexicons as in the PMI-cool system (Balchev et al., 2016), as well as sentiment polarity features (Nicosia et al., 2015).

We further want to use our features in a deep learning architecture, e.g., as in the MTE-NN system (Guzmán et al., 2016a; Guzmán et al., 2016b; Nakov et al., 2016a), which borrowed an entire neural network framework and architecture from previous work on machine translation evaluation (Guzmán et al., 2015).

Moreover, we plan to use information from entire threads as well as from other question-answer threads to make better predictions, as using thread-level information for answer classification has already been shown useful for SemEval-2015 Task 3, subtask A, e.g., by using features modeling the thread structure and dialogue (Nicosia et al., 2015; Barrón-Cedeño et al., 2015), or by applying thread-level inference using the predictions of local classifiers (Joty et al., 2015; Joty et al., 2016). How to use such models efficiently in our ranking evaluation setup is an interesting research question.

Finally, we plan to experiment with different CQA tasks, such as ranking similar questions to a given question and finding useful answer to a new question entered by a user of the forum as in SemEval-2016 Task 3. We could run a pipeline for Bulgarian using the same features and we will can compare the results to the current approach. This can include translation of the English resources to Bulgarian.

Acknowledgments. This research was performed by Tsvetomila Mihaylova, a M.Sc. student in Computer Science in the Sofia University "St Kliment Ohridski". It is also part of the Interactive sYstems for Answer Search (Iyas) project, which is developed by the Arabic Language Technologies (ALT) group at the Qatar Computing Research Institute (QCRI), HBKU, part of Qatar Foundation in collaboration with MIT-CSAIL.

References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding High-quality Content in Social Media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 183–194, Palo Alto, California, USA.

Ahn, J., Butler, B. S., Weng, C., and Webster, S. (2013). Learning to Be a Better Q'Er in Social Q&A Sites: Social Norms and Information Artifacts. In *Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries*, ASIST '13, pages 4:1–4:10, Montreal, Quebec, Canada.

Aluisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). Readability Assessment for Text Simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, IUNLPBEA '10, pages 1–9, Los Angeles, California, USA.

Balahur, A. and Turchi, M. (2014). Comparative Experiments Using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis. *Comput. Speech Lang.*, 28(1):56–75.

Balchev, D., Kiprov, Y., Koychev, I., and Nakov, P. (2016). PMI-cool at SemEval-2016 Task 3: Experiments with PMI and Goodness Polarity Lexicons for Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Baltadzhieva, A. and Chrupała, G. (2015). Question Quality in Community Question Answering Forums: A Survey. *SIGKDD Explor. Newsl.*, 17(1):8–13.

Barrón-Cedeño, A., Filice, S., Da San Martino, G., Joty, S., Màrquez, L., Nakov, P., and Moschitti, A. (2015). Thread-Level Information for Comment Classification in Community Question Answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '15, pages 687–693, Beijing, China.

Belinkov, Y., Mohtarami, M., Cyphers, S., and Glass, J. (2015). VectorSLU: A Continuous Word Vector Approach to Answer Selection in Community Question Answering Systems. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 282–287, Denver, Colorado, USA.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE: an Architecture for Development of Robust HLT applications. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, ACL '12, pages 168–175, Philadelphia, Pennsylvania, USA.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., and Peters, W. (2011). *Text Processing with GATE (Version 6)*.

Daume III, H. (2007). Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, ACL '07, pages 256–263, Prague, Czech Republic.

Guzmán, F., Joty, S., Màrquez, L., and Nakov, P. (2015). Pairwise Neural Machine Translation Evaluation. 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)*, ACL-IJCNLP '15, pages 805–814, Beijing, China.

Guzmán, F., Màrquez, L., and Nakov, P. (2016a). Machine Translation Evaluation Meets Community Question Answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, Berlin, Germany.

Guzmán, F., Màrquez, L., and Nakov, P. (2016b). MTE-NN at SemEval-2016 Task 3: Can Machine Translation Evaluation Help Community Question Answering? In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2003). *A Practical Guide to Support Vector Classification*. Technical report, Department of Computer Science, National Taiwan University.

Joty, S., Barrón-Cedeño, A., Da San Martino, G., Filice, S., Màrquez, L., Moschitti, A., and Nakov, P. (2015). Global Thread-level Inference for Comment Classification in Community Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 573–578, Lisbon, Portugal.

Joty, S., Màrquez, L., and Nakov, P. (2016). Joint Learning with Global Inference for Comment Classification in Community Question Answering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '16, San Diego, California, USA.

McCallum, A. K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Mihaylov, T. and Nakov, P. (2016a). Hunting for Troll Comments in News Community Forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL '16, Berlin, Germany.

Mihaylov, T. and Nakov, P. (2016b). SemanticZ at SemEval-2016 Task 3: Ranking Relevant Answers in Community Question Answering Using Semantic Similarity Based on Fine-tuned Word Embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Mihaylov, T., Georgiev, G., and Nakov, P. (2015a). Finding Opinion Manipulation Trolls in News Community Forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, CoNLL '15, pages 310–314, Beijing, China.

Mihaylov, T., Koychev, I., Georgiev, G., and Nakov, P. (2015b). Exposing Paid Opinion Manipulation Trolls. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP '15, pages 443–450, Hissar, Bulgaria.

Mihaylova, T., Gencheva, P., Boyanov, M., Yovcheva, I., Mihaylov, T., Hardalov, M., Kiprov, Y., Balchev, D., Koychev, I., Nakov, P., Nikolova, I., and Angelova, G. (2016). SUper Team at SemEval-2016 Task 3: Building a Feature-Rich System for Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '13, pages 746–751, Atlanta, Georgia, USA.

Mohammad, S. M., Salameh, M., and Kiritchenko, S. (2016). How Translation Alters Sentiment. In *Journal of Artificial Intelligence Research*, volume 55, pages 95–130.

Nakov, P., Màrquez, L., Magdy, W., Moschitti, A., Glass, J., and Randeree, B. (2015). SemEval-2015 Task 3: Answer Selection in Community Question Answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 269–281, Denver, Colorado, USA.

Nakov, P., Guzmán, F., and Màrquez, L. (2016a). It Takes Three to Tango: Triangulation Approach to Answer Ranking in Community Question Answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '16, Austin, Texas, USA.

Nakov, P., Màrquez, L., Moschitti, A., Magdy, W., Mubarak, H., Freihat, A. A., Glass, J., and Randeree, B. (2016b). SemEval-2016 Task 3: Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California, USA.

Nicosia, M., Filice, S., Barrón-Cedeño, A., Saleh, I., Mubarak, H., Gao, W., Nakov, P., Da San Martino, G., Moschitti, A., Darwish, K., Màrquez, L., Joty, S., and Magdy, W. (2015). QCRI: Answer Selection for Community Question Answering - Experiments for Arabic and English. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '15, pages 203–209, Denver, Colorado, USA.

PothulaSujatha and Dhavachelvan, P. (2011). A Review on the Cross and Multilingual Information Retrieval. *International Journal of Web & Semantic Technology (IJWesT)*, 2(4):115–124.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta.

Sadowski, C. and Levin, G. (2007). *SimiHash: Hash-based similarity detection*. Technical Report UCSC-SOE-11-07, University of California, Santa Cruz, USA.

Voorhees, E. M. (1999). The TREC-8 Question Answering Track Report. In *In Proceedings of TREC-8*, pages 77–82.

Webber, B. and Webb, N. (2010). Question Answering. In Alexander Clark, Chris Fox, S. L., Ed., *The Handbook of Computational Linguistics and Natural Language Processing*, chapter 22, pages 630–654.

Zhou, G., Cai, L., Zhao, J., and Liu, K. (2011). Phrase-based Translation Model for Question Retrieval in Community Question Answer Archives. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 653–662, Portland, Oregon.

Zhou, G., Liu, K., and Zhao, J. (2012). Exploiting Bilingual Translation for Question Retrieval in Community-Based Question Answering. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 3153–3170.

Quotation Retrieval System for Bulgarian Media Content

Ivelina Stoyanova, Martin Yalamov, Svetla Koeva
Department of Computational Linguistics
Institute for Bulgarian Language, Bulgarian Academy of Sciences
{iva,martin,svetla}@dcl.bas.bg

Abstract

This paper presents a method for automatic retrieval and attribution of quotations from media texts in Bulgarian. It involves recognition of report verbs (including their analytical forms) and syntactic patterns introducing quotations, as well as source attribution of the quote by identification of personal names, descriptors, and anaphora.

The method is implemented in a fully-functional online system which offers a live service processing media content and extracting quotations on a daily basis. The system collects and processes written news texts from six Bulgarian media websites. The results are presented in a structured way with description, as well as sorting and filtering functionalities which facilitate the monitoring and analysis of media content.

The method has been applied to extract quotations from English texts as well and can be adapted to work with other languages, provided that the respective language specific resources are supplied.

1. Introduction

In the age of digital technologies the daily amount of information made available on the internet has increased significantly. That is why information extraction, media monitoring, and opinion mining have become the focus of active research in NLP.

Retrieval of quotes from media content and identifying their author can be important for analysing the behaviour of various actors in the political or social life. This can help provide context for actions, events or statements, clarify the standing of certain figures regarding topics or issues, make a comparison of opinions. Research in this area has applications in social sciences, political sciences, journalism, etc.

There are three types of quotes – direct (literal presentation of someone's words), indirect (paraphrased speech) and mixed (where part of the statement is presented directly, while another part is paraphrased). They exhibit different features – word order, punctuation, grammatical dependencies (e.g., use of particular verb tense, voice and evidentiality forms), some of which are language specific (e.g., use of punctuation for subordinate clauses).

Although it may look like a trivial task, simple approaches for quotation retrieval do not perform particularly well and need improving. Actually, the task of quotation retrieval involves subtasks that still pose challenges to NLP, such as named entity (NE) recognition, including multiword NEs, anaphora resolution, syntactic parsing. Information retrieval and structured presentation of extracted information is also essential to ensure applicability of results for various research purposes.

The paper presents a system for automatic quotation retrieval from media content in Bulgarian. Our purpose is three-fold: (a) to elaborate on the practical aspects of quotation retrieval and attribution; (b) to offer a meaningful, structured representation of quotes which facilitates analysis of media content; and (c) to offer a live service which processes media content and extracts quotations.

In section 2. we discuss related work in the field. Section 3. presents in detail the features of quotation description and the method for quotation retrieval and attribution. The following section 4. is focused on the implementation of the online service for quotation retrieval and the structured representation of results. Section 5. shows some directions for extending the description of quotations by information extraction. The paper concludes by outlining some directions for future work.

2. Related Work

Approaches to quotation recognition vary in terms of: (a) coverage (direct, indirect and/or mixed quotations); (b) techniques (syntactic patterns, heuristics, machine learning); (c) applications (whether they were theoretical or have been implemented in a fully functional system).

Pouliquen et al. (2007) present the system `NewsExplorer` that extracts quotations from multilingual news, their author, as well as named entities occurring in the quote. The system also recognises variants of personal names.

Sagot et al. (2010) describes a corpus-based approach to quotation extraction based on the study of quotation verbs, their features and sentential categorisation frames. Krestel et al. (2008) developed a quotation extraction and attribution system that combines a lexicon of reporting verbs and a manually constructed grammar to detect specific constructions satisfying lexical constraints. Similarly, de La Clergerie et al. (2011) employ syntactic patterns to identify quotes in French news texts.

Sarmento and Nunes (2009) present an online service `verbatim` working on data from the Portuguese mainstream media. The authors outline some generic tasks: data acquisition and parsing, quotes extraction, removal of duplicates, topic distillation and classification, and interface design for presentation and navigation. They apply syntactic patterns on text to identify and attribute quotes to a speaker.

Schneider et al. (2010) present a system called `PICTOR`, that queries a large news corpus for topical quotations and then visualises them over time. Alongside identification of quotes and speakers in an article, authors select quotes relevant to a user query, scoring quote similarity in order to filter and cluster related quotes, and present a graph-based visualisation for plotting relevant quotes over time.

Atteveldt (2013) uses syntactic analysis and topic models to identify quotations from politicians. His method relies substantially on lexical resources. The author uses a dictionary to identify the sources (the person who is being quoted), and a list of verbs (e.g., *say, state*) and attribution phrases (e.g., *according to*).

Pareti et al. (2013) note the low portion of direct quotes (30-52% in the corpora they use) and focus on extraction of indirect and mixed quotes as well. They report on the results and evaluation of the extraction and attribution of direct, indirect and mixed quotations over two large news corpora.

Machine learning approaches for quotation retrieval have been suggested by Fernandes et al. (2011), O'Keefe et al. (2012), Pareti et al. (2013).

More often quotation retrieval is implemented as part of a more complex task, such as opinion mining and sentiment analysis (O'Keefe et al., 2013), or comparative analysis of political statements (Atteveldt, 2013).

Based on the review of related works we note several possible directions in which quotation retrieval can be further extended: (a) to develop fully functional retrieval systems on media content rather than applications for purely research purposes; (b) to perform analysis on dynamic media content on a daily basis rather than a fixed text corpus; and (c) to provide efficient description of quotations with filtering and sorting functionalities. Still, not many quotation retrieval systems are available online and on live media content. To the best of our knowledge, no system for quotation retrieval exists for Bulgarian.

3. Quotation Retrieval

3.1. Outline of the Task

The main task includes identification of the quotations and their attribution to a source. Here we cover direct and indirect quotes, while the (direct and indirect) components of mixed quotes are handled as separate entities. Since in the presentation of results quotations from the same news text and attributed

to the same person are grouped together, treatment of mixed quotes in this way does not affect their information representation.

Pareti (2012) and Pareti et al. (2013) define a quotation attribution relation by four components: a source span (the entity the content is attributed to); a cue span (the lexical anchor of the relation, e.g. a report verb); a content span (the quoted text); and a supplement span (any additional elements relevant to the interpretation of the attribution relation).

Taking into account the features above, we extend and organise the description of a quotation into the following sets of features: structural features used to extract the quotation, informational content features which characterise the content of the quotation, and external, or metadata-based, features which provide editorial information about the text the quotation appears in.

1. **Structural features**

   - **Span.** The quote can be contained within a sentence (Example 1a), or span over several sentences (Example 1b).
   - **Syntactic patterns and lexical elements.** In most cases the quotation is introduced by a reporting verb and a specific syntactic pattern. Indirect speech is also marked by subordinate conjunctions and linking words.
   - **Punctuation.** Punctuation is an essential feature for quotation retrieval. Direct quotes in Bulgarian (and many other languages) are introduced by colons and/or surrounded by quotation marks, or (rarely in news) introduced by a dash on a new line. Indirect speech is usually expressed as a subordinate (object) clause within the sentence without any distinctive punctuation.
   - **Source.** A quote is attributed to a speaker, who can be represented in the text by his name (e.g., *Boyko Borisov, Borisov*, Examples 2a and 2b), a descriptor (e.g., *the prime minister*, Example 2c), or an anaphora (e.g., *he*). In some cases the source can be an organisation or group presented by its name (e.g., *Bulgarian Socialist Party*) or an abbreviation (e.g., *BSP*, Example 2d).

2. **Informational content features**

   The content features include essential elements characterising the informational content and the topic of the quoted text. See Section 5. for more details on the techniques used for information extraction. Since in many cases the quoted text is short, it is not always possible to detect a particular topic in it. The content features include:

   - **Named entities of persons, places, and organisations within the quoted text.**
   - **Temporal expressions in the quoted text.** We identify dates and times which can be used to describe the topic of the quotation and to find relations with other quotations.
   - **Keywords** which relate to the overall content of the news text or are significant for describing the content of the quoted text.
   - **Opinion and sentiment features.** The reporting verb often reflects evaluation of the quote's content made by the author of the text, i.e. external for the quote itself, which is crucial for the analysis of its content. This can be expressed lexically, for example using verbs for negation (*The prime minister denied that ...*) or modal verbs (*The Bulgarian Socialist Party should state that it wants to get the power*), or morphologically using negative forms (*The prime minister did not state that ...*) or conditional mood (*The prime minister could have said: "..."*).

3. **External (metadata-based) features**

   - **Publication time of the news text.** This is the date (and possibly time) of the publication. It is useful in order to enable filtering or sorting by time, or creating a timeline for the quotations on a given topic.

- **Media source.** The media source is an essential part of the description of the quote. It allows to filter by source or to follow how various topics or events are presented across different media.
- **URL to the publication.** The publication can provide context of the quotation and it is a way to avoid copyright issues by (a) linking to the source, and (b) not publishing the whole text or large excerpts from it.

Example 1.
(a)
Gunlaugson tvardi, [*che zakonite ne sa narusheni i saprugata mu ne se e oblagodetelstvala finansovo.*]
Gunlaugston states [*that laws have not been broken and his wife has not benefited financially.*]
(b)
[*"Tova beshe edin dostoen mach za final. Tryabva da prodalzhim da se razvivame kato tehnika",*] *zayavi trenyorat Miroslav Zhivkov sled dvuboya.*
[*"This was a decent match for the final. We should continue to develop our technique",*] *said coach Miroslav Zhivkov after the game.*

Example 2.
(a)
"Tolkova parkove i gradini se napraviha v Sofia", kaza oshte ***Boyko Borisov****.*
"So many parks and gardens have been built in Sofia", added ***Boyko Borisov****.*
(b)
Borisov *potvarzhdava, che Balgariya shte podkrepya evropeyskata perspektiva na Sarbiya.*
Borisov *confirms that Bulgaria will support the European prospects of Serbia.*
(c)
V profila si vav Facebook ***premierat*** *napisa: "Edna naistina otlichna vecher za balgarskiya sport."*
On Facebook the ***prime minister*** *has posted: "One really excellent evening for Bulgarian sport."*
(d)
Ot ***BSP*** *zayaviha, che partiyata ne e saglasna s proekta za koalitsionno sporazumenie.*
From ***BSP*** *announced that the party does not agree with the proposal for a coalition agreement.*

3.2. Method for Quotation Retrieval and Attribution

The method for quotation retrieval relies on the following language specific resources for Bulgarian: dictionary of verbs used for reporting speech; list of patterns defining the analytical verb forms in order to identify the form of the reported verb and its tense, voice and mood (Leseva et al., 2015); dictionary of correspondences between names and titles or descriptors (e.g., *Boyko Borisov* and *prime minister*).

Initially, texts are annotated with POS and lemma. Taking into account the free word order in Bulgarian, the implementation of rigid syntactic patterns is not efficient. Instead, similarly to Pouliquen et al. (2007), we identify each quote as a triple of quoted text, reporting verb, and source (person) with the restriction that the verb and the source are both either on the left or on the right of the quoted text.

We perform pattern matching to identify the quoted text, as well as the source and the reporting verb. In direct quotations, the quoted text is introduced by punctuation (quotation marks, colons, dash, new line) and can span over several sentences. Indirect quotations are found within a sentence and are identified as subordinate clauses introduced by a report verb, subordinate conjunction and/or punctuation.

In Bulgarian, most NEs and more specifically, names of persons and organisation, are tagged by the POS and grammatical tagger and lemmatiser. Additional rules for identification of NEs were manually crafted. A sequence of single word personal NEs (e.g., first name followed by a surname) are combined and annotated as one NE. Special check is performed in a dictionary of categorised NEs from Wikipedia in order to separate geographic or organisational NEs from adjacent personal name.

A dictionary of correspondences between names and descriptors (such as titles, job posts, etc.) in Bulgarian has been automatically compiled from Wikipedia. Currently, it includes 31,446 personal names with a corresponding set of descriptors. The names include popular Bulgarian and foreign politicians, artists, sportsmen and public figures. All names and descriptors are matched to a canonical form,

usually the full name (e.g., *Borisov* is matched to *Boyko Borisov*). To avoid mismatches, the canonical form should occur at least once in the same text.

We apply a set of simple rules for anaphora resolution which cover only a selected number of cases in order to improve the recall of the method. We first identify third person singular pronouns in nominative (*toy – he*, and *tya – she*) which immediately precede the reporting verb. The attempt to resolve them includes looking backwards in the current and the previous sentence for a noun, including personal names, which agrees in gender and number with the anaphora. It is resolved only if the first agreeing noun is a NE. If the anaphora is matched with a common noun before reaching a NE, the anaphora is regarded as unresolved.

A dictionary of 114 reporting verbs in Bulgarian is used to identify the quotations in the text. The dictionary is extracted from the Bulgarian wordnet[1] by exploiting the semantic relations of synonymity and hypernymity – all synonyms and hyponyms of the synsets containing *govorya* (*speak*).

Based on the distances measured in number of tokens between any pair of the triple (quoted text Q – report verb V – potential source S), we evaluate a simple confidence measure for the validity of the retrieved quotation where the confidence (C) is reduced for any extra position separating any pair of the three components:

$$C(Q,S,V) = 0.99 - \frac{d(Q,V) + d(Q,S) + d(S,V) - 1}{3} \times 0.07$$

A set of 'penalties' is also introduced to adjust the score in some specific cases. They are applied in the following order:

- If the identified source (NE, descriptor, anaphora) and the reporting verb are on different sides of the quoted text, the score is reduced by 70%. This effectively excludes such cases.
- If the reporting verb and the source precede the quoted text, and the reporting verb is in active voice and precedes the source, the score is reduced by 30%.
- If the reporting verb is in active voice and the source (including any adjectives in front of it if it is a descriptor noun) is preceded by a preposition, the score is reduced by 20%.
- If the reporting verb is in passive voice and the source (including any adjectives in front of it) is not preceded by the preposition *ot* (*by*), the score is reduced by 20%.

The score is used for filtering out direct quotations attributed to the wrong source. The score is also applied to rank possible triples from the same sentence and select the most reliable from conflicting quotations. Example 3 shows the scores for three possible attributions of the indirect quotation in the sentence, the first attribution is disregarded as it is below the threshold of 0.5, and the attribution with the higher score (*Emil Radev*) is selected.

Example 3.
Po povod kandidata na GERB i dumite na Boyko Borisov evrodeputatat Emil Radev $_V$*komentira,* $[_Q$ *che tryabva da se promenyat pravilata za izdigane i izbor na prezident.*$]$
With respect to the candidate of GERB and the words of Boyko Borisov, the European MP Emil Radev $_V$*commented* $[_Q$ *that the rules for president nominations and elections should be changed.*$]$

$$C(Q, \text{GERB, V}) = 0.4000, \quad C(Q, \text{Boyko Borisov, V}) = 0.5867, \quad C(Q, \text{Emil Radev, V}) = 0.9207$$

The method is applied on Bulgarian media content collected from six major news websites. On average, daily about 3,200 potential quotations are identified, which are further filtered based on: (a) attribution to a named source – we exclude quotations that cannot be matched to named entities directly (a name is identified in the sentence) or indirectly (a descriptor or anaphora is identified in the sentence

[1] http://dcl.bas.bg/bulnet/

which is matched to a name); and (b) confidence score – we set a threshold of 0.5 for both direct and indirect quotes. Further, in the presentation of results we combine separate quotations, both direct and indirect, attributed to the same source within a single text (see Section 4.2.).

3.3. Evaluation

The evaluation of the method is based on a manually verified set of 200 quotations (79 direct and 121 indirect). We evaluate the precision and recall of discovering the full quotations (both boundaries) or only the start of the quotation. The evaluation of source attribution is performed on all identified quotations and includes NEs, descriptors and anaphoras. Only fully recognised names and matches to NEs are considered as correct. We perform experiments with different confidence thresholds, the results of which are presented in Table 1.

Type	**Confidence threshold**	**Full quotation**		**Start of quotation**		**Source attribution**	
		Precision	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
Direct	0.3	0.97	0.77	1.00	0.80	0.94	0.73
	0.5	1.00	0.63	1.00	0.63	0.97	0.65
	0.7	1.00	0.53	1.00	0.53	1.00	0.60
Indirect	0.3	0.81	0.58	0.89	0.66	0.82	0.68
	0.5	0.88	0.55	0.89	0.56	0.83	0.62
	0.7	0.90	0.50	0.92	0.51	0.87	0.61

Table 1: Evaluation of the results (precision and recall) in terms of: (a) the full quotation, (b) the identification of the start of the quotation, and (c) source attribution.

4. Online System for Quotation Retrieval

4.1. Workflow

The online quotation retrieval system is part of a complex system for collection and analysis of media content in Bulgarian. The results are available at `http://dcl.bas.bg/quotations/` (Figure 1). The workflow includes the following components:

1. **Download of texts from several news agencies**. Two approaches were implemented: (a) monitoring of RSS feeds; or (b) focused crawling with pre-crawl data mining. Metadata are extracted from the original webpage and stored separately from the text according to the principles of the Bulgarian National Corpus (Koeva et al., 2012).

2. **Processing and linguistic annotation** on Bulgarian texts was performed using the Bulgarian Language Processing Chain (Koeva and Genov, 2011) through a RESTful service. Downloaded and processed texts are added to the Bulgarian National Corpus and can be used for other applications, such as neologism detection[2].

3. **Quotation retrieval and text analysis** to describe quotation features as outlined in Section 3. The application for quotation retrieval is implemented in Java 7.

4. **Presentation of results**. The results are represented online in a structured manner and with a search and sorting functionality.

5. **Update routine**. Results are automatically updated on regular intervals throughout the day after newly downloaded data have been processed.

[2]http://dcl.bas.bg/neologisms/

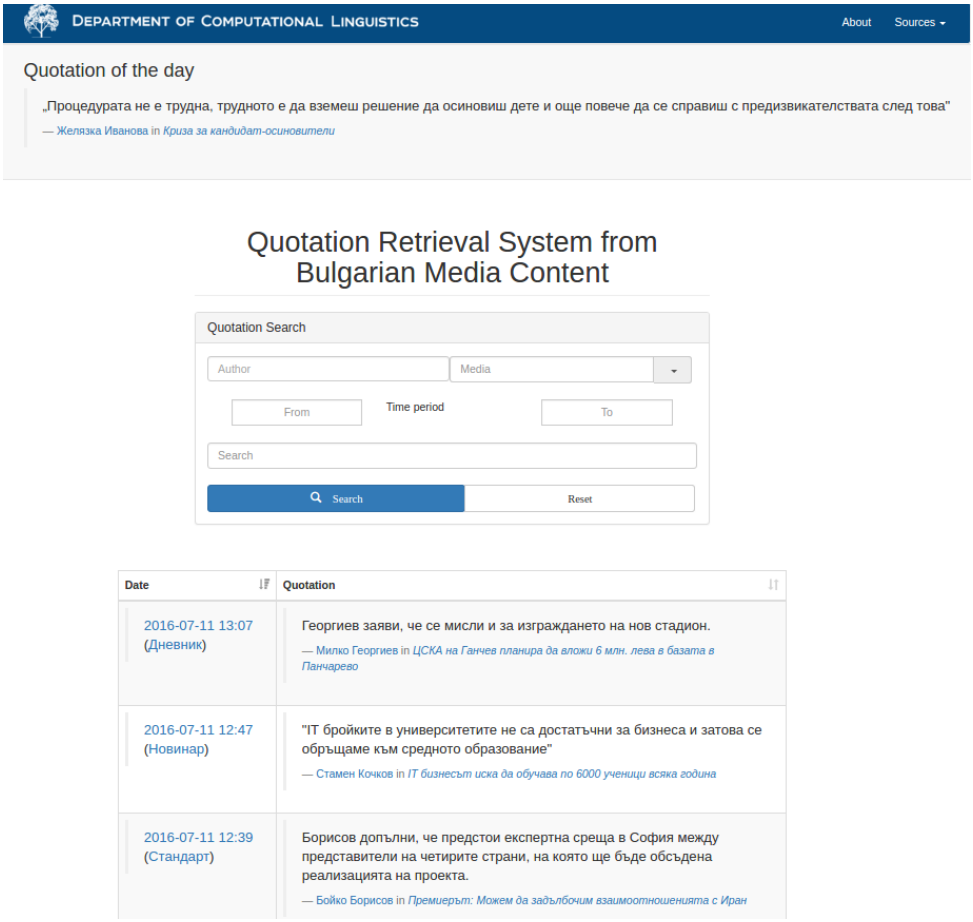

Figure 1: Results displayed online

4.2. Structured presentation of results

Retrieved quotations are put into a database and are presented online in a structured manner to facilitate their viewing and analysis. Each quote is presented with the following information: quoted text, source and link to the original news article. Quotations attributed to the same source within a text are combined together.

Quotations can be sorted by date of the news articles. There is also a searching and filtering functionality based on: (a) source – name of the person; (b) period of time; (c) media; and (d) query words within the quoted text. Each field has an autocomplete dropdown list which shows possible values and updates upon typing.

Further, we offer a 'Quote of the day' on a selected popular topic (e.g., on 5 April the most frequent topic was Panama papers). The selected quote needs to satisfy the following conditions: (i) to contain as many as possible of the top 10 most frequent keywords discovered within all texts of the day; and (ii) to have high confidence measure (above 0.9, or the highest available) in order to ensure that it is correctly identified and attributed.

5. Towards Topic Detection

In recent years topic modelling is gaining popularity as a way to discover and represent the abstract topics in a collection of documents, including in conversational texts such as emails and social media posts (Carenini et al., 2011) and news (Blei, 2012). Various well developed approaches have been applied, such as Latent Semantic Analysis or Latent Dirichlet Allocation. Recently, neural networks have been employed for the task of topic modelling (Mikolov and Zweig, 2012). Toolkits for topic modelling have also been developed, e.g. MALLET (Graham et al., 2012) or Stanford Topic Modelling Toolkit (Ramage

et al., 2009).

Here we perform the first steps towards topic modelling by identifying significant components within the set of quotes by the same source within a single text document. The following elements are extracted from the quoted text: (a) named entities of persons, places, and organisations; (b) temporal expressions; and (c) a set of keywords. Basically, we answer the set of questions *who, what, when, where* and define the topic in a very narrow sense. The list of identified elements can include proper names and temporal words (e.g., *Theresa May, Boris Johnson, London, Brexit, Great Britain, Wednesday*), as well as concrete and abstract nouns (*borders, minister, foreign affairs, politics*).

For NE recognition we use the same module applied for quotation attribution (see Section 3.2.). While in attribution we are only interested in named entities of persons or organisations (to whom quotations can be attributed), here we also identify geographical entities and event names. Categorisation of NEs is performed using a dictionary of NEs derived from Wikipedia and other sources divided into semantic categories – personal names, organisations, places, and events (Koeva et al., 2016).

Temporal expressions include dates (*14 July*, *14/07/2016*, etc.), time (e.g., *18:00*), concrete or relative temporal expressions (e.g., *on Tuesday, in April, yesterday, last year*). Temporal relations can also be expressed morphologically (e.g., by the verb form). So far we only consider explicit dates and time.

Keywords extraction on the quoted text is based on: (i) predefined dictionary of 139 domain-specific words which point to a domain (e.g., budget – Economy; parliament – Politics); and (ii) frequency analysis (words, except stop words, with frequency above a threshold are identified as keywords). The dictionary in (i) is applicable in the cases of short texts where frequency analysis is not informative.

The topic detection is essential for providing more functionalities in the online system for quotation retrieval in terms of grouping of results, finding quotations about the same or similar topic, or discovering relations between quotation from different media sources.

6. Future Work

The method for quotation retrieval has been also applied on English news texts collected through the BBC RSS feed and annotated using Stanford CoreNLP (Manning et al., 2014). We compile an English dictionary of reporting verbs containing 43 unique verbs derived from the Princeton WordNet in a similar way to that of the Bulgarian reporting verbs (see Section 3.2.). Essentially, the methods for NE recognition are the same with the use of some language specific resources such as the dictionary of English NEs from Wikipedia. The same patterns for matching quotations are applied.

In the future our efforts will be focused on improving the method for quotation retrieval and its results. At present, quotation attribution is performed only for named sources, i.e. either labelled as or matched directly to NEs. However, these depend on the quality of the modules for anaphora resolution and the coverage of the dictionaries matching descriptors to NEs. Moreover, we are looking into ways to establish more matches (e.g., based on previous occurrences in media texts) and to increase significantly the recall of the system. Machine learning methods also look promising for the purposes of quotation identification and source attribution.

Furthermore, we could use information about whether the report verb is a marker of opinionated content and of what polarity (Esuli and Sebastiani, 2006). Some of these verbs are neutral (e.g., *say, tell, explain*) while others express opinion about particular features of the quotation such as its truth value (e.g., *deny*) or importance and validity (e.g., *emphasise, hint*). SentiWordNet (Baccianella et al., 2010) is used to obtain the positivity and the negativity scores of verbs for the purposes of opinion mining and sentiment analysis. The analysis based on the verb semantic features falls outside of the scope of the present study. Here we use the reporting verb purely as a lexical marker introducing the quotation.

The work on the online system for quotation retrieval is ongoing. Our aim is to cover more web sources and possibly extend the data beyond news and media domain. Finally, improvement in the presentation of results is also among our future tasks – including more information in the quote description, filtering on more features, etc. User feedback will also be valuable in this respect.

References

Atteveldt, W. V. (2013). Quotes as Data: Extracting Political Statements from Dutch Newspapers. In *New Directions in Analyzing Text as Data Workshop*.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pages 2200–2204, Valletta, Malta. European Language Resources Association (ELRA).

Blei, D. (2012). Topic Modeling and Digital Humanities. *Journal of Digital Humanities*, 2(1).

Carenini, G., Ng, R., and Murray, G. (2011). *Methods for Mining and Summarizing Text Conversations*. Syntesis Lectures on Data Management.

de La Clergerie, E., Sagot, B., Stern, R., Denis, P., Recource, G., and Mignot, V. (2011). Extracting and Visualizing Quotations from News Wires. In Vetulani, Z., Ed., *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 522–532.

Esuli, A. and Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, pages 417–422.

Fernandes, W. P. D., Motta, E., and Milidiu, R. L. (2011). Quotation Extraction for Portuguese. In *Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (STIL 2011)*, pages 204–208, Cuiaba.

Graham, S., Weingart, S., and Milligan, I. (2012). Getting Started with Topic Modeling and MALLET. Programming Historian (02 September 2012), http://programminghistorian.org/lessons/topic-modeling-and-mallet.

Koeva, S. and Genov, A. (2011). Bulgarian Language Processing Chain. In *Proceeding to The Integration of multilingual resources and tools in Web applications Workshop in conjunction with GSCL 2011*. University of Hamburg.

Koeva, S., Stoyanova, I., Leseva, S., Dekova, R., Dimitrova, T., and Tarpomanova, E. (2012). The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, 0(1):65–110.

Koeva, S., Stoyanova, I., Todorova, M., and Leseva, S. (2016). Semi-automatic Compilation of the Dictionary of Bulgarian Multiword Expressions. In *Proceedings of the GLOBALEX 2016 Workshop: Lexicographic Resources for Human Language Technology, LREC*, pages 86–95.

Krestel, R., Bergler, S., and Witte, R. (2008). Minding the Source: Automatic Tagging of Reported Speech in Newspaper Articles. In *Proceedings of the Sixth International Language Resources and Evaluation*.

Leseva, S., Stoyanova, I., and Koeva, S. (2015). Automatic Recognition of Verb Forms in Bulgarian. In *Paisievi Cheteniya*, Plovdiv.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Mikolov, T. and Zweig, G. (2012). Context Dependent Recurrent Neural Network Language Model. In *Proceedings of the 2012 IEEE Workshop on Spoken Language Technologies*, pages 234–239, Miami, USA, December.

O'Keefe, T., Pareti, S., Curran, J. R., Koprinska, I., and Honnibal, M. (2012). A Sequence Labelling Approach to Quote Attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 790–799.

O'Keefe, T., Curran, J. R., Ashwell, P., and Koprinska, I. (2013). An Annotated Corpus of Quoted Opinions in News Articles. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 516–520. ACL.

Pareti, S., O'Keefe, T., Konstas, I., Curran, J. R., and Koprinska, I. (2013). Automatically Detecting and Attributing Indirect Quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999. ACL.

Pareti, S. (2012). A Database of Attribution Relations. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 3213–3217.

Pouliquen, B., Steinberger, R., and Best, C. (2007). Automatic Detection of Quotations in Multilingual News. In *Proceedings of Recent Advances in Natural Language Processing*, pages 487–492.

Ramage, D., Rosen, E., Chuang, J., Manning, C. D., and McFarland, D. A. (2009). Topic Modeling for the Social Sciences. In *Neural Information Processing Systems (NIPS) Workshop on Applications for Topic Models: Text and Beyond*, Whistler, Canada, December.

Sagot, B., Danlos, L., and Stern, R. (2010). A Lexicon of French Quotation Verbs for Automatic Quotation Extraction. In Calzolari, N., 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*, Valletta, Malta. European Language Resources Association (ELRA).

Sarmento, L. and Nunes, S. (2009). Automatic Extraction of Quotes and Topics from News Feeds. In *4th Doctoral Symposium on Informatics Engineering*.

Schneider, N., Hwa, R., Gianfortoni, P., Das, D., Heilman, M., Black, A. W., Crabbe, F. L., and Smith, N. A. (2010). *Visualizing Topical Quotations Over Time to Understand News Discourse*. Technical report. T.R. CMU-LTI-10-013, Carnegie Mellon University, Pittsburgh, PA.

Stress Patterns of Compounds and MWEs in English and Bulgarian

Bistra Popovska
Department for English Studies
Faculty of Philology
The Paisii Hilendarski University of Plovdiv
bistra63@abv.bg

Rositsa Dekova
Department for English Studies
Faculty of Philology
The Paisii Hilendarski University of Plovdiv
rosdek@gmail.com

Abstract

The paper presents an ongoing research on the stress patterns of compounds and MWEs of the type ADJ+N and their corresponding free NPs in English and Bulgarian. The research focuses on the identification and the formal representation of the possible stress patterns of compounds and MWEs and free NPs. During our research so far, we have compiled a corpus of over 2000 compounds and MWEs, approx. 1000 for each language – English and Bulgarian. Our theoretical framework includes elements from different theories, i.e. the Generative Phonology Theory, the Metrical Theory, and the Theory of Primary accent first which all define the stress as a prosodic element. Our main goals are to specify the prosodic region where the stress is defined in English and Bulgarian MWEs and noun phrases and to define the main features of the stress in MWEs and free NPs in English and Bulgarian. The results of our research can serve for implementation into NLP modules for spoken language processing and generation.

1. Introduction

The opportunities of modern language technologies resulting in the accumulation of huge data bases and large-scale theoretical generalizations relate to a rapid development of phonological and cognitive-linguistic prosodic studies and particularly those focused on stress (see Patseva, 2011, among others). One phenomenon which has received little attention in previous linguistic work on speech prosody is the use of contrastive stress patterns to distinguish meanings at the suprasegmental levels. The types of stress contrast can be exemplified by minimal pairs, such as the compound *gréenhouse* vs. the phrase *green hóuse* in English and *червеношийка* (*tchervenoshéeka* 'robin') and *червéна шийка* (*tchervéna sheeka* 'red neck') in Bulgarian. Previous studies using behavioural methods (e.g. Cutler and Otake 1999, Cutler and Van Donselaar 2001) and electrophysiology (e.g. Friedrich et al. 2001) suggest that listeners use lexical stress information during spoken word identification. However, the distinction between compound and phrasal stress and the role it plays in online comprehension remain relatively unexplored, and represent the focus of the present study.

Our corpus includes more than 1000 MWEs for each of the languages English and Bulgarian. The English compounds and phrases have been compiled by hand, while the Bulgarian MWEs were extracted from *The Bulgarian Dictionary of MWEs* (Koeva et al., 2016; Stoyanova and Todorova, 2014).

2. Stress as prosodic element

In linguistic terms speech prosody studies stress and intonation. In the present research we focus on the stress patterns of compounds and MWEs of the type ADJ+N and their corresponding free NPs in English and Bulgarian.

Firstly, we have examined the main prosodic concepts relating to stress, starting with the prosodic hierarchy, phonological aspect of the syllable and the main phonetic and phonological phenomena of rhythm as discussed by Dimitrova (1998, 1999), Giegerich (2005, 2011), Halle and Vergnaud (1987), Prince and Smolensky (2002), Savitska and Boyadzhiev (1988), Tilkov (1982), among others.

Secondly, we have outlined the prosodic area in which stress is defined using the terminology apparatus of the contemporary approaches to prosody. We have studied the phonetic and phonological aspects of stress, following the logic of scientific research, which has led us to the paradigm of the linguistic tendencies of the last decades. For the purposes of our research, we have considered different theoretical approaches, i.e. the Generative Phonology Theory, the Metrical Theory, and the Theory of Primary accent first.

3. Characteristics, functions and position of stress in English and Bulgarian

In our study, we follow traditional ways of defining and describing the characteristics and functions of Bulgarian stress – phonetic, positional and phonological (word stress and phrasal stress) as discussed by Kurlova (1997), Misheva (1991), and Tilkov and Boyadzhiev (1978), among others. However, we also introduce an analysis from the point of view of the Primary Accent First theory of Van der Hulst (2002, 2009).

A far as the stress in English is concerned, we have described it according to the four variable indicators: intensity, pitch, vowel quality and vowel duration, following Collins and Mees (2003).

The study investigates firstly word stress: the degrees of stress, lexically designated stress in English and then switch stress. We have formulated some word stress guidelines concerning words consisting of two or three syllables (in most of the cases stress falls on the first syllable, e.g. *súmmer*) , longer words having four or more syllables (in most cases stress falls on the antepenultimate syllable, e.g. *solubílity*), prefix words (in shorter words the main stress generally falls on the syllable after the prefix, e.g. *repláy*), word endings (on ending itself, e.g. *himsélf*, on syllable preceding ending, e.g. *defícient*).

Secondly, we have investigated stress in compounds and MWEs (Initial Element Stress, e.g. *Rússian class*. Final Element stress, e.g *Russian roulétte*). Collins and Mees (2003) have formulated some stress guidelines for compounds, defining word shape: The Manufactures Rule, according to which if the compound contains the material from which the item is manufactured, the stress falls on the final element, e.g. *apricot brándy*; and the Location Rule which dictates that if the compound contains location, the stress usually falls on the final element, e.g. *London Éye* (Collins and Mees 2003). They have formulated some further stress pattern guides related to the above (for example food labels generally have stress on the final element, e.g. *scrambled éggs*).

Finally, we have also investigated the phonetic and phonological characteristics of English stress according to Giegerich (2005) from the point of view of the Metrical phonology.

4. Stress patterns of compounds, MWEs and Phrases in English and Bulgarian

4.1. The importance of the contrast

Contrastive word stress plays an important role in the differentiation between compounds and phrases. The current research investigates the development of compound and phrasal stress in both production and perception, an area considerably neglected by previous studies on linguistic stress. By comparing directly production and perception we aim to prove that compounds are generally not mistaken for phrases while phrases are often mistaken for compounds.

According to some authors (Chomsky and Halle 1968:15) compound and phrasal stress can be assigned in English by the Compound Stress Rule according to which stress is placed on the first segment of the compound. In contrast, the Nuclear Stress Rule dictates that phrasal stress is assigned to the rightmost phrase segment. The abovementioned difference in the placement of stress generally allows listeners to discriminate between compounds and phrases with identical constituents.

According to others (e.g. Bloomfield 1933:228, Giegerich 2006), however, the stress criterion in English, commonly invoked in attempts to draw the compound-phrase distinction, is getting less reliable: it fails to correlate with other (semantic, syntactic) criteria; it draws on incomplete and flawed generalizations regarding stress in compounds and phrases. A fictitious category distinction arises for pairs of semantically very similar constructions such as *Christmas púdding* and *Chrístmas cake*. *Ice cream* has a variable stress pattern – for some speakers it is a compound and for others it is a phrase. Then that distinction needs to be revisited.

4.2. Stress patterns of compounds and MWEs in English and Bulgarian

The current study describes the types of noun-centred (e.g. *wíndmill*) and verb-centred (e.g. *táxi-driver*) compound nouns in English. It aims to define the labels for stress in compound nouns and collocations. In addition, we investigate the single stress pattern and the double stress pattern in compounds and collocations together with the stress pattern in three-word compound nouns.

We define the types of compound nouns in Bulgarian with or without a linking element (e.g. *злобо**о**днéвка* (*zlob**o**dnévka* 'burning topic') and *свръхпроизвóдство* (*svruhproizvódstvo* 'overproduction'), respectively); with or without suffixation (e.g. *първокласник* (*purvoklás**nik*** 'first-grader') vs. *буквоя́д* (*bukvoyád* 'verbalist')). As far as the stress patterns of compound nouns are concerned, we point out the reasons for their formation: extralinguistic and linguistic (semantic, syntactic and morphological).

4.3. Stress patterns of free phrases in English and Bulgarian

The research investigates the types of free phrases in English and Bulgarian together with their stress patterns. The terms *theme* and *rheme* are introduced together with the way they relate to stress. We aim to describe the factors, determining the relation between *rheme* and stress. For example, the reason can be syntactic, as the noun modifier can bear stronger stress, e.g ***бял*** *кон* (*byál kon* 'white horse').

5. Conclusions and future work

Following our theoretical approach, we have planned some experimental research focusing on compound/phrasal distinction. The experiment is designed to satisfy two experimental task types: production and perception, using as stimuli minimal pairs of segmentally identical but prosodically distinct phrases and compounds such as *blackboard* and *black bóard.* A statistical analysis of the results of the experiments will provide empirical evidence to support our theoretical model.

Finally, we believe that a thorough investigation and a proper formal representation of stress patterns of compounds, and especially MWEs as opposed to free phrases will contribute greatly to the tasks of spoken language processing and generation.

References

Bloomfield, L. (1933). *Language*. New York: Henry Holt.

Chomsky, N. and Halle, M. (1968). *The Sound pattern of English*, NY: Harper & Row.

Collins, B. and Mees M.I. (2003). *Practical Phonetics and Phonology, A Resource Book for Students*. Routledge, pages 109–116.

Cutler, A. and Van Donselaar, W. (2001). Voornaam is Not a Homophone: Lexical Prosody and Lexical Access in Dutch. *Language and Speech*, 44:171–195.

Cutler, A. and Otake, T. (1999). Pitch Accent in Spoken-word Recognition in Japanese. *Journal of the Acoustical Society of America*, 105:1877–1888.

Dimitrova, S. (1998). Bulgarian Speech Rhythm: Stress-Timed or Syllable-Timed, *Journal of International Phonetic Association*, 27 (1-2):27–33.

Dimitrova, S. (1999). *Ritmichna organizatsiya na angliyskata i balgarskata rech: sapostavitelno fonetichno izsledvane*. PhD Dissertation Abstract, Sofia.

Friedrich, C.K., Alter K. and Kotz S.A. (2001). An Electrophysiological Response to Different Pitch Contours in Words. *Neuroreport*. Oct 29; 12(15):3189–91.

Giegerich, H. (2005). *Metric Phonology,* Cambridge Studies in Linguistics, Cambridge University Press.

Giegerich, H. (2006). Attribution in English and the Distinction between Phrases and Compounds, In: Petr Rosel (ed.) *Englisch in Zeit und Raum – English in Time and Space: Frschungsbericht fur Klau Faiss*. Trier: Wissenschaftlicher Verlag Trier.

Giegerich, H. (2011). *Compounding and Lexicalism, Handbook of Compounding.* Oxford: Oxford University Press.

Halle, M. and Vergnaud J.-R. (1987). An Essay on Stress. *Current Studies in Linguistics 15*). Cambridge, Mass.: MIT Press.

Hulst, H.G. van der (2002). Stress and Accent. In: Nadel, L. (ed.). *Encyclopedia of Cognitive Science.* vol. 4, 4. London: Nature Publishing Group, pages 246–254.

Hulst, H.G. van der (2009). *Brackets and Grid Marks or Theories of Primary Accent and Rhythm.* In: E. Raimy and C. Cairns (eds.). *Contemporary views on architecture and representations in phonological theory.* MIT press.

Koeva, S., Stoyanova, I., Leseva, S. and Todorova, M. Semi-automatic Compilation of the Dictionary of Bulgarian Multiword Expressions. In *Proceedings of the GLOBALEX Workshop*, LREC, 2016, ELRA.

Kurlova, R. (1997). *Fonetika i fonologia*, Blagoevgrad.

Misheva, A. (1991). *Intonatsionna sistema na balgarskiya ezik*. Sofia, BAS.

Patseva, M. (2011). Dinamika na konotatitvnoto znachenie. *Littera et Lingua*.

Patseva, M. (2007). Aktsentni modei pri nyakoi sashtestvitelni imena ot mazhki I zhenski rod. *Godishnik na Fakulteta po slavyanski filologii na SU*, vol. 96–98.

Prince, A. and P. Smolensky 2002 (1993). *Optimality Theory. Constraint Interaction in Generative Grammar*, ROA 537 0802, Rutgers

Savitska, I. and Boyadzhiev, T. (1988). *Bulgaro-polska supostavitelna gramatika*, vol. 1. Bulgarian Academy of Sciences.

Tilkov, D. (1982). Srichka. In *Gramatika na balgarskiya knizhoven ezik. T. 1. Fonetika.* Sofia, BAS, pages 154–158

Tilkov, D. and Boyadzhiev, T. (1978). *Udarenieto v balgarskiya knizhoven ezik.* Sofia, Narodna prosveta Publishing House.

Verbal Multiword Expressions in Croatian

Krešimir Šojat
Faculty of Humanities and Social Sciences
University of Zagreb
ksojat@ffzg.hr

Matea Filko
Faculty of Humanities and Social Sciences
University of Zagreb
msrebaci@ffzg.hr

Daša Farkaš
Faculty of Humanities and Social Sciences
University of Zagreb
dberovic@ffzg.hr

Abstract

The paper deals with verbal multiword expressions in Croatian. We focus on four types of verbal constructions: light verb constructions, i.e. constructions consisting of a light verb and a noun or prepositional phrase, complex predicate constructions, i.e. constructions consisting of a finite and infinitive verb, prepositional verb constructions, i.e. constructions consisting of a verb and a typical preposition, and, finally, verbal idioms, i.e. constructions with completely idiosyncratic meanings. All the constructions are annotated in the Universal Dependency treebank for Croatian. The identification of verbal multiword expressions is an important task in numerous NLP tasks. It is also important to define and delimitate this concept in linguistic theory.

1. Introduction

The identification and annotation of multiword expressions in Croatian corpora and treebanks have so far gained little attention, although these constructions pose a challenge for various NLP tasks.

Multiword expressions (MWEs) refer to various types of constructions consisting of two or more words that act as a single unit at some level of analysis. Sag et al. (2002) define MWEs as "idiosyncratic interpretations that cross word boundaries (or spaces)" and provide an extensive account of various MWEs in English as well as of criteria for their classification. Generally, MWEs are divided into those that are fixed, i.e. the paradigmatic selection of elements and their syntagmatic order is never altered, and those that can be modified to a certain degree, either in morphosyntactic properties of elements and/or their selection. The meaning of MWEs can vary from more or less compositional to completely idiosyncratic. MWEs usually include noun compounds, multiword named entities, different types of complex verb phrases, idioms and others.

Reporting on annotation schemes in 17 dependency and constituency based treebanks for 15 languages, Rosen et al. (2016:179) point out that there is little agreement on how MWEs should be annotated in treebanks. On top of that, they stress that "there is, in fact, not even agreement on what constitutes a MWE in NLP". Baldwin and Kim (2010) and Rosen et al. (2016) divide MWEs into following groups: 1. nominal MWEs; 2. verbal MWEs; 3. prepositional MWEs; 4. adjectival MWEs; 5. MWEs of other categories; 6. proverbs.

In this paper we focus on verbal MWEs in Croatian. We deal with this type of MWEs because a) there is no previous research done on the identification and annotation of verbal MWEs in Croatian language resources, primarily treebanks and b) there is no resource which would enable an extensive research of MWEs in Croatian and refinement of linguistic criteria for their classification. The paper is structured as follows: In section 2 a brief description of verbal MWE in Croatian is presented and criteria for their classification are given. Section 3 describes the procedure for annotating verbal MWEs and the reasons for selecting the Universal Dependency for Croatian for this purpose. In sections 4 and 5 the results obtained by the MWE annotation of Universal Dependency treebank are presented and discussed. The paper ends with concluding remarks and an outline of future work.

2. Verbal MWEs in Croatian

Both Baldwin and Kim (2010) and Rosen et al. (2014, 2016) divide verbal MWEs into subgroups of phrasal verbs, light verb constructions, VP idioms and other verbal MWEs. All these subgroups require a careful examination for Croatian.

The category of phrasal verbs is generally neither recognized nor discussed in Croatian grammars and reference books. However, Katunar et al. (2012) point out that a particular preposition can significantly change the meaning of a verb and argue that such expressions should therefore be treated as a single unit. The meaning of a verb that co-occurs with an object PP can be significantly different from the meaning of the same verb that co-occurs with an adverbial PP. For instance: 1. *zagrijati se pod pokrivačem* 'to warm up under the blanket' vs. 2. *zagrijati se za kuhanje* 'to become interested in cooking', where the meaning of the verb *zagrijati se* is completely different when used with different PPs.

Light verb constructions (e.g. *donijeti odluku* 'to make a decision; to reach a decision') are made up of a verbal and a nominal component. The nominal component consists of a NP or a PP. Noun in NPs are generally derived from verbal stems and they are usually in accusative case. Light verbs have entirely or partially lost their lexical meaning and the meaning of the whole construction is actually expressed by NPs or PPs. Light verb constructions (LVCs) in Croatian are syntactically flexible since light verbs can be inflected, passivized and marked as perfectives or imperfectives. In some LVCs nouns can be used both in singular and plural and/or in different cases. An important feature of LVCs is that they can frequently be substituted with a single "heavy" verb, e.g. *donijeti odluku – odlučiti*, although the meanings of the LVC and their paraphrases, i.e. semantically full verbs, often do not exactly correspond.

VP idioms (or *phrasemes*) are usually categorized into two groups: decomposable and non-decomposable idioms. The division is based on the degree of semantic and syntactic opaqueness of the whole construction in regard to its elements, as well as on the possibility of the word order change within an idiom.

The group of other verbal MWEs in our research refers to multiword predicates consisting of a finite verb and one or more verbs in infinitive form. Verbs in finite form typically belong to modal or phasal verbs (e.g. inchoative verbs).

All verbal MWEs listed above form complex sentence predicates, i.e. multiword units, and therefore need to be identified and annotated in Croatian language resources.

3. Procedure

There are three dependency treebanks available for Croatian. The first one is the Croatian Dependency Treebank (HOBS) that in its latest version encompasses 4,626 sentences of Croatian newspaper. HOBS is freely available for on-line search (hobs.ffzg.hr). The second one – SETIMES.HR dependency treebank (http://nlp.ffzg.hr/resources/corpora/setimes-hr/) – was built on top of the newspaper text from the SETIMES parallel corpus. The treebank contains approximately 9,000 sentences, and it is completely free. These two treebanks are annotated with modified versions of annotation schemes used in the Prague Dependency Treebank project done for Czech. However, we decided to deal with verbal MWEs in the third available treebank, Universal Dependency (UD) Treebank for Croatian.[1] The UD treebank of Croatian was also built from newspaper text originating from SETIMES parallel corpus, but annotated according to UD annotation. The UD treebank version used in this experiment consisted of 3557 sentences.

This treebank was chosen for the task presented in this paper for two reasons: 1) although it is the smallest in size compared to other two treebanks, it is large enough for a preliminary research of identification and annotation of verbal MWEs in Croatian, 2) the UD annotation guidelines account for different types of MWEs and mark the relation between their components on syntactic level. They distinguish between fixed multiword expressions (for example, *in spite of* is marked with *mwe* tag), multiword names (*name*) and foreign phrases (*foreign*). Other types of MWEs are recognized as well. For example, parts of English phrasal verbs are marked as compounds. However, the criteria for the recognition of MWEs are not clearly stated: "Deciding whether an expression in a language should be treated as a MWE is something that has to be decided for each language, and in some cases this will require somewhat arbitrary conventions, because it involves choosing a cut point along a path of

[1] A detailed account of building this treebank and achieved parsing scores is given in Agić and Ljubešić (2015).

grammaticalization."[2] For Croatian, this kind of convention is not established yet, and the following experiment is the first step in this direction.

In the first step of the task we built an initial list of verbal MWEs from available work done in this area for Croatian. A list of approximately 20 phrasal verbs (i.e. combinations consisting of a verb and a preposition) was taken from Katunar et al. (2012), whereas a list of LVCs was compiled from Silić and Pranjković (2005) and Gulić (2015). This list contained 80 LVCs for Croatian. We searched for verbal MWEs from this initial set in the chosen treebank. In this step we wanted to determine which MWEs appear in the treebank and whether they can be automatically annotated. We also wanted to determine whether the light verbs from the list can be used for the detection of other NPs or PPs in new LVCs. Unfortunately, the obtained results were completely unsatisfactory since none of the phrasal verbs was detected in the treebank whereas only 14 LVCs from the initial set were identified. This could suggest that the initial list is too small and narrow (and even not built on the real language data, i.e. data from various corpora) or that only a very limited number of verbal MWEs occurs in the corpus. The other option is not very likely since the corpus consists of newspaper texts and such constructions are very frequent in this genre. This was the reason to manually annotate the selected treebank for verbal MWE types as described in Section 2. In other words, in the second step of the task we manually annotated 3557 sentences from the UD treebank for Croatian for phrasal verbs, LVCs, VP idioms and multiword predicates. The results are presented and discussed in the following section.

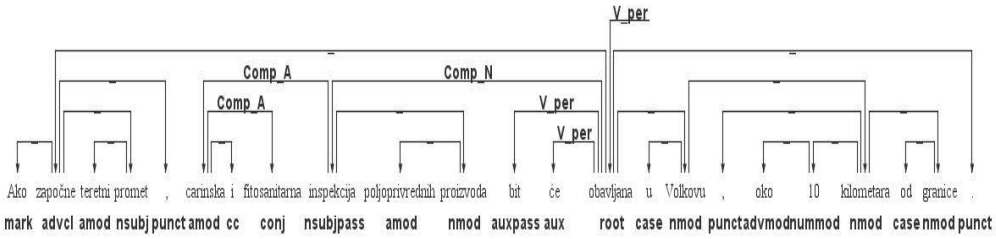

Figure 1: An example of a sentence annotated for a LVC

Verbal MWEs were marked in the corpus on a separate level of annotation in order to enable their explicit differentiation from other verbal phrases with similar or identical morphological and syntactic properties. This is particularly important when dealing with LVCs and verbal idioms. Each member of verbal MWEs was marked in our approach. More details are given in the following section.

4. Results

The total number of verbal MWEs belonging to the group of phrasal verbs (cf. Section 2) is 371. We annotated verbs in such MWEs in the treebank with V_Prep tag and prepositions with Comp_Prep tag. In Table 1 we list the most frequent ten verbal MWEs annotated witih V_Prep tag in the treebank.

[2] UD annotiation guidelines, http://universaldependencies.org/u/overview/syntax.html

verbs	prepositions	frequency of MWEs (verb + preposition)	total frequency of verbs in the corpus (within and outside MWEs)
pozivati 'call'	*na* 'for'	18	30
razgovarati 'talk'	*o / s* 'about / to'	10	29
raditi 'work'	*na / o / za* 'on / for'	9	47
dovesti 'bring'	*do* 'to'	7	16
glasovati 'vote'	*za / o / protiv* 'for / on / against'	7	10
odnositi se 'refer'	*na* 'to'	7	17
ovisiti 'depend'	*o* 'on'	7	14
nastaviti continue'	*s* 'with'	6	32
sastati se 'meet'	*s* 'with'	6	37

Table 1: the most frequent 10 MWEs annotated as prepositional verbs and total frequency of verbs

The second group of verbal MWEs comprises LVCs. The total number of annotated LVCs in the treebank is 847. Verbal parts in these constructions are tagged with V_per tag. NPs and PPs that are elements of these constructions were annotated as Comp_N and Prep_N respectively. In Table 2 we present the most frequent ten light verbs in the selected treebank, NPs and PPs that co-occur in LVCs as well as their frequency in the corpus. The frequency threshold is set at two occurrences.

light verb	frequency in various LVCs	NPs in LVCs	frequency of NPs in LVCs	PPs in LVCs	frequency of PPs in LVCs
imati 'have'	73	*posljedice* 'consequences'	5	*u vidu* 'in sight' (keep in mind)	3
		pravo 'right' (be right)	5	*za cilj* 'as its aim'	3
		utjecaj 'influence'	5		
biti 'be'	56	*domaćin* 'host'	3	*u stanju* 'in position'	6
				u mogućnosti 'able to'	4

dobiti 'get'	30	*nagradu* 'prize' (receive a prize)	5	*na težini* 'on weight' (gain importance)	2
		potporu 'support'	5		
izraziti 'express'	25	*nadu* 'hope'	6		
		potporu 'support'	6		
		zabrinutost 'concern'	6		
osvojiti 'win'	18	*nagradu* 'an award'	7		
		odličje 'a medal'	3		
dati 'give'	17	*izjavu* 'statement' (make a statement)	2		
		potporu 'support'	2		
podnijeti 'submit'	16	*ostavku* 'resignation'	8		
		tužbu 'a complaint'	3		
postati 'become'	14	*članicom* 'member'	5		
predstavljati 'be'	13	*zapreku* 'obstacle'	4		
poduzeti 'take'	12	*korake* 'steps'	8		

Table 2: the most frequent 10 light verbs annotated as V_Per and their nominal components

The third group of verbal MWEs encompasses VP idioms. We have detected and marked 18 verbal MWEs as VP idioms. However, there are only 7 different VP idioms in the selected corpus. They are listed in Table 3, along with their overall frequency. Each member of verbal idioms was marked with V_idiom tag.

VP idiom	frequency
biti na čelu 'be at the head'	4
biti na klimavim nogama 'be without a firm foundation'	3
biti u punom zamahu 'be in full swing'	3
hvatati se / uhvatiti se u koštac 'take the bull by the horns'	3

ne odustati ni pedlja 'not to retreat a single inch'	2
staviti točku na 'put an end to'	2
zatražiti zeleno svjetlo 'request approval'	1

Table 3: 7 VP idioms and their overall frequency

Finally, in Table 4 we present the results for multiword predicates consisting of a verb in finite form and a verb in infinitive form. The elements of these MWEs are marked as V_fin and V_inf respectively.

verb in finite form	frequency
moći 'can'	142
trebati 'should'	120
morati 'must'	95
željeti 'want'	32
biti 'be'	14
planirati 'plan'	14
pokušati 'try$_{pf}$'	12
pokušavati 'try$_{ipf}$'	10
odlučiti 'decide'	9
kaniti 'plan'	9
uspjeti 'succeed'	8
nastaviti 'continue'	7
početi 'begin'	5
htjeti 'will'	5
odbiti 'refuse'	5

Table 4: the most frequent 15 verbs annotated as V_fin and their overall frequency

5. Discussion

The first group contains 371 verbs that form so called phrasal or prepositional verbs in Croatian. PPs in this group should be differentiated from PPs that denote adverbials. The PPs in this group denote objects. Semantically similar prepositional objects can be introduced with different prepositions. In some cases the meaning of the verb is not affected by the selection of a preposition, e.g. *misliti na* 'to think of' and *misliti o* 'to think about'. In other cases the meaning of the verb alters under the influence of the preposition introducing the object, e.g. *odnositi se na* 'to refer to' and *odnositi se prema* 'to treat

somebody in a particular way'. In future work these information will be used for the creation of verb valency frames and distinguishing of senses in the large database of Croatian verbs CroDeriV.[3]

The second group encompasses light verb constructions with 252 unique light verbs. The light verbs from these semi-compositional constructions always have their counterparts that are not impoverished in their lexical meaning. The light verbs retain only a portion of the full lexical meaning of their homonymic counterparts. As the obtained results reveal, this group can be further divided into several subgroups. The first division is based on the ability to be paraphrased with a single verbs (e.g. *dati doprinos* 'to give a contribution' – *doprinijeti* 'to contribute'). However, in numerous cases such paraphrases are not possible, e.g. *dobiti zadatak* 'to get an assignment'. On top of that, some LVCs that can be paraphrased with a single verb in certain contexts, in other contexts acquire additional semantic components and paraphrases are not possible. E.g. the construction *donijeti zaključak* 'to make a conclusion' can be paraphrased with the verb *zaključiti* 'to conclude'. They are not completely interchangeable in all contexts, since the LVC *donijeti zaključak* can in some contexts mean 'to agree that'. This LVC can in some cases imply that the conclusion(s) are presented or given in a written form, whereas the verb *zaključiti* almost never appears in this context. The group of detected LVC is, as far as we know, the biggest list of such constructions available for Croatian.[4]

The third group contains VP idioms. For several reasons this is the most problematic group. Firstly, the inter-annotator agreement was extremely low when dealing with this category. There was a significant overlapping of VP idioms and LVCs. Secondly, the lack of clear criteria for distinguishing LVCs and VP idioms in Croatian literature made the whole procedure even more complicated. Finally, the results show that the division of VP idioms into decomposable and non-decomposable VP idioms discussed in Section 2 seems to have no relevance for such constructions in Croatian since all detected and annotated VP idioms belong to the group of decomposable VP idioms.

The fourth group contains multiword predicates consisting of a verb in finite form followed by one or more verbs in infinitive forms. Verbs that appear in finite forms predominantly belong to modal verbs (e.g. must, should etc.) or phasal verbs (begin, start etc.) However, other detected verbs are those that are usually not classified as modal or phasal in Croatian (e.g. *planirati* 'to plan', *pokušati* 'to try', *uspjeti* 'to succeed', *odlučiti* 'to decide' etc.). These results address the issue of redefinition verbal groups that are followed by infinitive forms as well as the treatment of such constructions as complex predicates. In numerous cases infinitive VPs appear to be morphosyntactic realization of objects. Finally, infinitives often follow nominal predicates (e.g. in constructions as *biti voljan učiniti* 'to be willing to do') or LVCs (e.g. *biti u mogućnosti doći* 'be able to come'). These constructions raise additional questions regarding the status of complex predicates and the traditional notion of object. However, this topic is beyond the scope of this contribution.

6. Conclusion and Future Work

It is clear that the results obtained for all four verbal MWEs in Croatian are valuable in several respects: They were obtained from the first research of such constructions for Croatian that is based on corpus data and therefore more truly indicate the productivity of particular prepositional and light verbs in combinations with various PPs and NPs than the data presented in existing literature. Secondly, the results enable further investigation of possibilities for automatic detection and recognition of MWEs both from monolingual and parallel corpora of the Croatian language. Thirdly, the presented results raise several theoretical questions and provide possibilities for their in-depth analysis. Finally, the obtained results will enable the creation of a language resource that would encompass various types of verbal MWEs and enable queries according to various parameters. The outline of this database is given in Figure 2 below.

[3] CroDeriV in its present shape contains data on derivational relatedness of Croatian verbs. It is available at http://croderiv.ffzg.hr/. The next phase of the development is aimed at valency and meaning description of verbs.

[4] All the results discussed here are available upon request. The complete database will be public and downloadable.

Light verb	AUX	AUX	V_per	AUX	AUX	REF	AUX	PREP	A	A	N	N - lema	PREP	N	Example
dobiti			dobio	je							nagradu	nagrada			dobio V_per je V_per nagradu Comp_N
dobiti			dobio	je						posebnu	nagradu	nagrada			dobio V_per je V_per posebnu Comp_A nagradu Comp_N
dobiti			dobio	je							nagradu	nagrada	za		Nagradu Comp_N je V_per dobio V_per za Comp_Prep
dobiti			dobio								naknadu	naknada			dobio V_per naknadu Comp_N
dobiti			dobili							punu	neovisnost	neovisnost	od		dobili V_per punu Comp_A neovisnost Comp_N od Comp_Prep
dobiti			dobio	je							odobrenje	odobrenje			dobio V_per je V_per odobrenje Comp_N
dobiti			dobiti								odobrenje	odobrenje			treba V_comp_fin dobiti V_per odobrenje Comp_N
dobiti			dobiti								posao	posao			dobiti V_per posao Comp_N
dobiti			dobiti							snažnu	potporu	potpora			dobiti V_per snažnu Comp_A potporu Comp_N

Figure 2: An excerpt from the database of verbal MWEs

Acknowledgement

This paper is partially supported by European Union, European Social Fund, under the project grant HR.3.2.01.-0037 Mrežni portal za online učenje hrvatskoga jezika HR4EU.

References

Agić, Željko and Ljubešić, Nikola (2015). Universal Dependencies for Croatian (that Work for Serbian, too). Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing, pp. 1–8, Hissar, Bulgaria

Baldwin, Timothy and Su Nam Kim (2010). Multiword Expressions. In: Nitin Indurkhya and Damerau, Fred J. (Eds.) *Handbook of Natural Language Processing, Second Edition*. Boca Raton, USA: CRC Press, pp. 267-292.

Gulić, Anamarija. (2015). *Klasifikacija perifraznih glagola u hrvatskom jeziku.* MA thesis. Faculty of Humanities and Social Sciences, University of Zagreb.

Katunar, D., Srebačić, M., Raffaelli, I., and Šojat, K. (2012). Arguments for Phrasal Verbs in Croatian and Their Influence on Semantic Relations in Croatian WordNet. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Available at: https://bib.irb.hr/datoteka/582794.Croatian_phrasal_verbs_KSRS.pdf

Rosén, V. et al. (2014). A Survey of Multiword Expressions in Treebanks. In: *Proceedings of the 14th International Workshop on Treebanks and Linguistic Theories (TLT14)*, 11–12 December 2015, Warsaw, Poland.

Rosén, V. et al. (2016). MWEs in Treebanks: From Survey to Guidelines. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16),* 23-28 May 2016, Portorož, Slovenia

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In: *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, Mexico City, Mexico, pp. 1-15.

Silić, J. and Pranjković, I. (2005). *Gramatika hrvatskoga jezika – za gimnazije i visoka učilišta.* Zagreb: Školska knjiga.

A Simple Approach to Unifying Ambiguously Encoded Kurdish Characters

Sardar Jaf
The University of Durham
sardar.jaf@durham.ac.uk

Abstract

In this study we outline a potential problem in the normalisation stage of processing texts that are based on a modified version of the Arabic alphabet. The main source of resources available for processing resource-scarce languages is raw text. We have identified an interesting challenge that must be addressed when normalising certain natural language texts. Many less-resourced languages, such as Kurdish, Farsi, Urdu, Pashtu, etc., use a modified version of the Arabic writing system. Many characters in harvested data from the Internet may have exactly the same form but encoded with different Unicode values (ambiguous characters). It is important to identify ambiguous characters during the normalisation stage of most text processing tasks. We will demonstrate cases related to ambiguous Kurdish and Farsi characters and propose a semi-automatic approach to identifying and unifying ambiguously encoded characters.

1. Introduction

One of the main challenges in processing less-resourced languages is the lack of natural language processing (NLP) tools and resources.

Large numbers of languages use a modified version of the Arabic writing system, such as Kurdish, Farsi, Urdu, Pasthtu, etc. We have observed a situation where characters of these languages have exactly the same form but encoded differently. The problem with the inconsistent encoding of some characters (ambiguous characters) is they are treated as different characters. This makes large numbers of similar words, which are similar in meaning and form, to be treated as completely different words. This situation is evident in most languages that use a modified version of Arabic script. In this paper, we attempt to shed light on ambiguous characters, which results in generating multiple instances of words of similar forms but different encodings. Moreover, we will show an approach for identifying ambiguous characters and an approach for correcting them by appropriately unifying their Unicode values. We will mainly focus on Kurdish but we show the applicability of our work to other related languages such as Farsi.

This paper is organised as follow: in section 2 we present a brief overview of Kurdish and in Section 3 we highlight some of the general challenges in processing Kurdish. While in Section 4 we outline situations where inconsistencies in character encoding could generate multiple words unnecessarily, in section 5 we describe our approach to identifying and unifying Unicode values of characters of equal forms. In section 6 we briefly describe the applicability of our approach to processing other related languages and we conclude our paper in Section 7.

2. A Brief Overview of Kurdish

Kurdish belongs to the Indo-European language family and it is closely related to Farsi. Kurdish is spoken by approximately 25 to 30 million people, but the exact number varies depending on the source. Despite the fact that large numbers of people speak Kurdish, the language is considered as a resource-scarce language.

There are several dialects in Kurdish, such as Sorani, Kurmanji, Zazaki, Hawrami, Gorani, etc. However, the main two dialects are Sorani and Kurmanji. These two dialects differ in many ways, one of the main differences is the writing style. Sorani uses a modified version of Arabic scripts while Kurmanji uses a modified version of Latin scripts. The use of a modified version of Arabic alphabet poses an interesting challenge in processing Sorani text, where this challenge is applicable to other related languages that use Arabic alphabet. Although our focus is to unify the encoding of Kurdish Sorani we will show that our approach could be applied to other related languages, such as Farsi. In the following section we will highlight some of the major challenges in processing Kurdish text.

3. The Challenges of Processing Kurdish

Two of the challenges in processing Kurdish are the dialect diversity and script diversity. However, the main challenge that we address in this paper is about unifying the Unicode values of some characters that are similar in form (ambiguous characters). Before describing our approach it is worth highlighting some aspects of the dialect and script diversity of Kurdish so that we can demonstrate the character ambiguity with examples.

3.1. Dialect Diversity

In Kurdish, there are several dialects. Two of the main dialects are Sorani and Kurmanji. These dialects differ in a number of ways. The effect of the differences is that developing an NLP tool for one dialect is not easily applicable to another dialect, hence the tasks for developing any NLP applications for Kurdish is twice more compared to working on other languages, therefore we state that our solution has been applied to Kurdish Sorani dialect only. Below is a short list of some of the main differences between Sorani and Kurmanji dialects (Esmaili, 2012; MacKenzie, 1961; McCarus, 1958):

- Gender distinction: both gender (feminine:masculine) is retained in Kurmanji while it is ignored in Sorani.
- Case Opposition: Kurmanji uses case opposition (absolute:oblique) for nouns and pronouns while Sorani ignores them but uses the pronominal suffixes to take over the function of the case.
- For past tense transitive verbs, Sorani uses pronominal enclitics, because of the absence of oblique pronoun, while Kurmanji uses the full ergative alignment.
- Sorani verb morphology is used for constructing passive and causative while in Kurmanji the verb هاتن (*hatin* "to come") and دان (*dan* "to give") are used respectively.
- The definite suffix ەکە- (*-eke*, "the") is used only in Sorani.

3.2. Script Diversity

One of the major differences between Sorani and Kurmanji is the writing system. Kurmanji uses a modified version of Latin alphabet while Sorani uses a modified version of Arabic alphabet. The script diversity makes it difficult to bijectively construct a mapping between Sorani and Kurmanji in many cases (Gautier, 1998). Some of the one-to-many mappings between the two writing systems are demonstrated in Table 1.

As can be noted from Table 1 (a) multiple Latin letters could be mapped to one Arabic letter. Similarity in Table 1 (b) the mapping from the Arabic-based letter {ه} to the Latin-based letters is a one-

to-many mapping, where mapping to any of H, h, E, or e, is not a trivial process. The same situation applies to mapping letter {و } to U, u, W, or w and letter {ی} to Î, î, Y, or y. The mapping shows that multiple Latin letters may potentially be mapped to an Arabic letter. This paper is to identify multiple Unicode values that are assigned to Arabic-based letters that have the same form and identify a way to unify them. Table 1 shows a list of identified characters with different Unicode values, however we anticipate that there may be other characters that did not appear in our dataset.

Unicode Value	Latin-based letters	Arabic-based letters	Unicode Value
u0048	H	ه	u06BE u06D5 or u0647
u0068	h		
u0049	I	-	-
u0069	i		
u0055	U	و	u0648
u0075	u		
u0057	W	و	u0648
u0077	w		
u0059	Y	ی	u06CC
u0079	y		

a) Mapping from Latin-based to Arabic-based

Unicode Value	Arabic-based letters	Latin-based letters	Unicode Value
u06BE u06D5 or u0647	ه	H	u0048
		h	u0068
		E	u0057
		e	u0077
u0648	و	U	u0055
		u	u0075
		W	u0045
		w	u00EA
u06CC	ی	Î	u00CE
		î	u00EE
		Y	u0059
		y	u0079

b) Mapping from Arabic-based to Latin-based

Table 1. Mapping between Arabic-based and Latin-based of Kurdish Alphabets (Esmaili, 2012).

From Table 1 we can see that the letter {ه} constitutes one letter (H, h, E, e) which is pronounced as either */ha/* or */a/* depending on its location in the word. If it appears at the start of a word it forms {هـ} or in some cases if it appears in the middle it forms {ـهـ}, and in both cases it is pronounced as */ha/*, however it may be assigned different Unicode values. If it appears at the end of a word it forms {ـه}or in isolation it forms {ه} and in both cases it constitutes /E/ or /e/. In addition to these two cases, in most electronic texts, it may appear as a *zero-width non-joiner (zwnj)* character, which prevents joining a character from its follower (Esmaili, 2012). For example, in the word بارهەڵگرەکە (*barHelgreke, "*The goods carrier*"*) it constitutes */H/* in the fourth position, it constitutes */e/* in the fifth position, and it constitutes a *zwnj* character in position nine in the word. For the same letter (i.e., the letter {ه}) different Unicode values are often used. For example when it appeared in position four in the word its Unicode value was 06BE but in some cases it is assigned 0647, when it appeared in position five and nine its Unicode value was either 0647, 06BE, or 06D5. This inconsistent encoding makes large numbers of words lose their unique form. Table 2 contains examples of different words that have the same forms but different Unicode values. This kind of ambiguity has also been observed in Urdu (Bajwa et al, 2011).

The problem that we are going to address is related to identifying ambiguous characters (i.e., characters that have the same form but different Unicode values) and unify their Unicode values. The solution to this problem is essential during the normalisation process of Sorani text because ignoring this problem will lead to incorrectly treating large numbers of words as unique words.

3.2.1. Other challenges of Processing Kurdish

Another challenge in processing Kurdish stems from the lack of NLP tools and resources. The unavailability of data, resources and tools for Kurdish text processing restrict developers in their approaches to processing the language. Another noticeable challenge is in segmentation and tokenisation is the identification of sentence, phrase, and word boundary. It is not possible to use spaces as a boundary sign because they may appear within a word or between words, or they may be absent between some sequential words (Esmaili, 2012; Shamsfard, 2011; Bajwa et al, 2011). Additionally, because of the absence of capitalisation the task of segmentation, tokenisation, and recognising Named Entities are further complicated.

The difficulty in processing Kurdish is further aggravated by the lack of gold-standard dataset. Although there are several dictionaries available for Kurdish annotated corpus and large datasets are still unavailable (Walther and Sagot, 2010). It is possible to use the large data available on the Internet for developing a corpus of raw text, which could be used in Information Retrieval application for example. However the harvested data from the Internet may pose a number of problems and solving the ambiguity of characters must be performed during the normalisation stage of raw text processing.

4. Dataset Collection

In this section we will describe the database that we have used for validating the appropriateness of our approach to identify ambiguous characters. Fortunately, there is a large number of Kurdish news websites, where we can harvest the required data. We have collected various data from several website.[1]

The collected dataset contains about 2,000,000 words which constitute just over 21,000 articles, which is large enough to capture a large variety of word forms. The dataset is also diverse, which includes topics covering sport, economy, politics, art, culture, health, multimedia and lifestyle, science and technology.

5. Collecting, Identifying and unifying ambiguously encoded characters

In this section we will describe the steps that we have used in identifying ambiguous words and characters in terms of their forms.

5.1. Collecting and parsing web pages

The first step of the process involved collecting over 21,000 news articles from a large number of websites. Then, we parsed each web page and removed various unwanted data, such as mark-up text, numbers, punctuations, foreign words, etc. A small challenge in this step is that although it is easy to identify Latin-based scripts in the pages, detecting Arabic or Farsi words is hard because they share the same writing system as Kurdish Sorani. A simple way to tackle this issue is to extract all unique words from the data with a specific frequency threshold. We have intentionally removed words that have occurred less than 0.001% in the data. These words were either Arabic or Farsi names, which are occasionally used in Kurdish news articles; words with incorrect spelling; and words that are accidentally merged with some other words during the parsing process of the web pages.

5.2. Identifying unique characters

Once a set of clean text is retrieved we have processed all the data and generated a lexicon, which contained unique words, and manually inspected the most frequently occurring words. At this stage, a

[1]The data are collected from the following websites: www.asoyroj.com, www.chawigal.com, www.aweza.co, www.dengekan.info, www.gulan-media.com, www.hawlati.co,, www.hawpshti.com, www.helwist.com, www.lvinpress.com, www.malmokurd.com, www.nnsroj.com, www.radionawxo.org, www.regaykurdistan.com, www.rojnews.net,www.rozhnamawany.com, www.serbexo.com, www.shaqam.net, www.shrova.org, and www.xendan.org

large number of words were treated as unique words even though they had similar forms with some other words. For example, as we have mentioned in Section 3.2. some Arabic-based characters are ambiguous. These ambiguous characters may appear in many words that are exactly the same in terms of meaning and form. Table 2 contains examples of some of the most ambiguously occurring words in the lexicon. Also, we can note from Table 2 the frequency of most ambiguous words is high.

Total words	Frequency	Unicode Value
لە (*le*, “on”)	135059	u0644 u0647
لە (*le*, “on”)	122063	u0644 u06D5
کە (*ke*, “as”)	125881	u0643 u06D5
کە (*ke*, “as”)	92812	u0643 u0647
کە (*ke*, “as”)	39747	u06A9 u0647
کە (*ke*, “as”)	11312	u0643 u06D5
کوردستان (*kurdistan*, “Kurdistan”)	16081	u0643 u0648 u0631 u062F u0633 u062A u0627 0646
کوردستان (*kurdistan*, “Kurdistan”)	13196	u06A9 u0648 u0631 u062F u0633 u062A u0627 0646
ئێمە (*aeme*, “us”)	4252	u0626 u06CE u0645 u0647
ئێمە (*aeme*, “us”)	4050	u0626 u06CE u0645 u06D5
هێزی (*hyz*, “its power”)	2472	u0647 u06CE u0632 u06CC
هێزی (*hyz*, “its power”)	2042	u06BE u06CE u0632 u06CC
هێزی (*hyz*, “its power”)	1674	u06BE u06CE u0632 u0649

Table 2. Ambiguous words with their frequency and Unicode values (letter 'u' is used in front of each Unicode value to distinguish different character's encoding value)

The identification of ambiguous characters in words is performed by manually inspecting the encoding values of characters in many frequently occurring words. Using the identification of unique words is time consuming and does not give an accurate account of the level of character ambiguity in the data, and it is neither efficient nor easy to locate ambiguous characters in large numbers of words.

An efficiency improvement is achieved by processing every character in every word in the lexicon and record all the unique characters along with their Unicode values. Then manually inspect the encoding of the recorded characters. However, the inefficiency aspect of this approach is it requires processing very large number of characters. For example, our dataset contained 2,983,579 words and the average word length was 6 characters, which yielded approximately 18 million characters to process. The time taken to process all the words was 56 seconds.

This approach can be improved using a very simple technique. That is, recording all the unique words in a second lexicon which does not contain duplicate words[2]. Then process the characters of the recorded unique words. The total number of unique words was dramatically reduced to 52,987 words and the processing time was reduced to 29 seconds.

Once we have identified all the unique characters and their Unicode values, which gave us a total of 37 unique characters (excluding punctuations), we have then identified three ambiguous characters, i.e., characters with the same form but different Unicode values. Those characters are shown in Table 3.

[2]Ambiguous words are treated as duplicated but are treated as unique because their Unicode values are unique even thought their forms are similar.

Characters	Unicode Values	Frequency
ہ (pronounced as /ha/, /a/ and used as zero-width non-joiner character)	u06D5	2361391
	u0647	1961352
	u06BE	51442
ی (pronounced as /ye/)	u06CC	81987
	u0649	69363
	u06BE	61961
ک (pronounced as /k/)	u06A9	585728
	u0643	537621

Table 3. Ambiguous Characters

5.3. Unifying Different Unicode Values of Similar Characters

Once we identified all the unique words and manually identified the ambiguous characters, we generated a mapping dictionary that mapped each Unicode value to different Unicode value, which is shown in Table 4. The content of the mapping dictionary is simple and can be formatted in any styles.

Unicode Value	Mapped Unicode Value
u0647	u06BE or u06D5
u0643	u06A9
u0649	u06CC
u06BE	u06CC

Table 4. Mapping from one Unicode Value to Another Unicode Value

Generally, if we find a specific character with a specific Unicode value in a word then we replace it with a given Unicode value. However, as it can be noted from Table 4 the Unicode value u0647, which represents ہ (a, "a"), (h, "ha"), or *zwnj* should remain as it is or be mapped to u06BE or u06D5. The location in which the character appears dictates its form. If the character was followed by a character with the same Unicode value then it's changed to u06BE. Otherwise, there are exceptional cases for correctly mapping u0647 to u06BE or u06D5: (i) if the character is final then we replace it with u06D5. (ii) if a specific vowel (with the Unicode value u06CE, u06CC, u0627, or u06c6) follows the character then it should be mapped to u06D5. (iii) If the previous two cases do not apply then it should be mapped to u06BE. The steps for finding the Unicode values of ambiguous characters are given in Figure 1.

The mapping dictionary that we have compiled contains one entry per line. In each entry there are two comma separated Unicode values (parameters), where the first parameter represents an ambiguous character in a word and the second parameter represents the Unicode value that is used for replacing the ambiguous character. In order to deal with the exceptional cases for handling u0647 Unicode value, the format of the dictionary entry for characters with u0647 Unicode value is in the following: the first parameter is u0647 Unicode value; the second parameter is the value that replaces u0647 if the character with u0647 Unicode is a final character; the third parameter is a list of *n* number of Unicode values, where n is a positive number; the last parameter is the value that is used for replacing the character with u0647 Unicode value if and only if the immediate following character is similar to Unicode values in the list of *n* Unicode values.

1. read words from a file and add them to a lexicon (*L*).
2. count the frequency of each word in *L* and create a create a new lexicon containing words and their frequency (*LF*).
  2.1. optional: sort content of *LF* in ascending order.
  2.2. for each word in *LF*, write the word and its Unicode values to a file for manual inspection.
  2.3. retrieve the characters of each word in *LF*.
    2.3.1. add the characters to a list (*CL*) if it is not in *CL*.
  2.3.2. write the character and their Unicode values to file if it is not in *CL*.
3. Inspect the characters that were written to the file in step 2.3.2 and identify n duplicate characters, where n is a predetermine number with the same form but with different Unicode values.

Figure 1. Finding the Unicode values of ambiguous characters

From the list of characters that we have identified by following the steps in Figure 1. we have manually inspected the characters that were of the same form but with different Unicode values. This way we have identified the characters that had the same form but different Unicode values, which resulted in duplicating a large number of words; some examples are shown in Table 2. Once we have identified all the ambiguous characters, we have compiled a mapping dictionary for replacing the Unicode values, which is shown in Table 4. The algorithm for mapping/unifying ambiguous characters is given in Figure 2. The evaluation of the solution is conducted by extracting all the unique characters and their Unicode values from the lexicon and manually inspecting them to identify a character that is similar in form to one or more character(s) but with different Unicode value. The absence of an ambiguous character indicated that all characters in the lexicon were encoded correctly.

```
For each word w is in the lexicon L do:
  For each entry uv in the mapping Unicode value dictionary do:
    Find uv in w
    If uv is in w:
      If there are more characters in w after the identified uv do:
        If the character that follows uv is the same as uv do:
          Replace uv with the second parameter in the entry
        Else do:
            If the uv is at the start of w do:
              If the length of the entry is more than 2 parameters
                If the character that immediately follows uv is in the list of special characters:
                  Replace uv with the second parameter in the entry
                Else:
                  Replace uv with last parameter in the entry
              Else:
                Replace uv with second parameter in the entry
            Else:
              If the length of the entry is more than 2 parameters
                If the character that immediately follows uv is in the list of special characters:
                  Replace uv with the second parameter in the entry
                Else:
                  Replace uv with last parameter in the entry
              Else:
                Replace uv with second parameter in the entry
```

Figure 2. Unifying ambiguous characters

6. Applying Our Approach to Related Languages

We also used the same approach on Farsi, which is closely related to Kurdish. From our experiment on Farsi we identified that in Farsi the number of ambiguous characters is fewer than those in Kurdish. For example, from Table 5 we can see that the final and medial character ي (*y,* "y") appears with different Unicode values. It is noticeable that the final ي (*y,* "y") has u06CC assignment more frequently than u06BE while a medial ـيـ (*y,* "y") is assigned u06CC Unicode value more than u06BE. Unlike in Kurdish, the character ه (a, "a") have not been assigned the Unicode value u06BE. The u0647 Unicode value is assigned to the initial, medial and final character ه (a, "a") more than u06D5 Unicode value. The third ambiguous character in Farsi was the character {ک} (*k,* "K") which was often assigned the Unicode values u06A9 instead of u0643 Unicode value. In conclusion, after applying the same technique to related languages we can identify ambiguous characters and semi-automatically correct them.

Total words	Frequency	Unicode Value
آئین	118	u0622 u0626 u06CC u0646
آئین	10	u0622 u0626 u06BE u0646
آزادی	112	u0622 u0632 u0627 u062F u06CC
آزادی	11	u0622 u0632 u0627 u062F u0649
جامعه	197	u062C u0627 u0645 u0639 u0647
جامعه	22	u062C u0627 u0645 u0639 u06D5
حاکم	183	u062D u0627 u06A9 u0645
حاکم	14	u062D u0627 u0643 u0645

Table 5. Farsi ambiguous words

7. Conclusion

The normalisation of text often involves removing unwanted texts (noise) such as foreign words, numbers, punctuations, etc. This stage of text processing is one of the main stages in processing less-resourced languages because in most cases raw data is collected from the Internet, which may contain various noise. In addition to noise removal process of online text we have identified an interesting case in processing Kurdish, and other related languages such as Farsi, where some characters of similar forms are assigned different Unicode values (ambiguous characters). We anticipate that the reason is that for languages that use a modified version of Arabic script for writing may interchangeably use different Unicode values, which could be the Unicode value of the original Arabic character or a specific code for the modified character. Another possibility is it may be due to the type of Operating Systems or the data entry devices that are used in compiling web pages, where they have different Unicode values for characters with similar forms.

Unifying ambiguous characters is an important step in the text normalisation stage because ambiguous characters, which are used for constructing words, lead to ambiguous words. In many inductive NLP processing tasks it is not plausible to induce information from noisy data. Therefore, unifying Unicode values of ambiguous characters is an essential step towards removing noise.

In this paper, we have presented a semi-automatic approach to unifying Unicode values of Kurdish text. Furthermore, we have used the same approach on Farsi and we have identified the same issue. Our experiment on Farsi shows that our approach could be applicable to other related languages, such as Urdu and Pashtu, which we aim to apply it to them in the near future.

References

Bajwa, U. I., Rehman, Z., and Anwar, W. (2011). Challenges in Urdu Text Tokenization and Sentence Boundary Disambiguation. *In Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP 2011)*.

Esmaili, K. Shykh. (2012). Challenges in Kurdish Processing. *In Proceedings of the 9th Information Retrieval Society Conference (AIRS 2013), Singapore.*

Gautier, G. (1998). Building a Kurdish Language Corpus: An Overview of the Technical Problems. *In Proceedings of ICEMCO 1998.*

MacKenzie, D. N. 1961. *Kurdish Dialect Studies,* V*olume 1 of London Oriental Series*. Oxford University Press.

McCarus, E. N. (1958). *A Kurdish Grammar Descriptive Analysis of the Kurdish of Sulaimaniya Iraq*. PhD thesis. New York, USA: American Council of Learned Societies.

Shamsfard M. (2011). Challenges and Open Problems in Persian Text Processing. *In Proceedings of the 5th Language and Technology Conference (LTC 2011), Poland.*

Thackston. W. M. (1960). *Sorani Kurdish: A Reference Grammar with Selected Reading*s. Oxford University Press, UK.

Walther, G., B. and Sagot, B. (2010). Developing a Large-Scale Lexicon for a Less-Resourced Language: General Methodology and Preliminary Experiments on Sorani Kurdish. *In Proceedings of the 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages (LREC 2010 Workshop), Malta*.

A Possible Solution to the Problem of Machine Translation of Verb Forms from Bulgarian to English

Todor Lazarov
Department of computational linguistics, IBL-BAS
todorlazarov91@abv.bg

Abstract

The paper`s main subject is concerned with the problems related to machine translation of verb forms from Bulgarian to English. In separate sections of this article we discuss the problems related to differences between word formation in both languages and differences in the information that the verb forms grammaticalize. We also introduce the idea of implementing the statistical method of machine translation altogether with the rule-based method as a proposal for future research and the possible practical and theoretical outcomes.

1. Introduction

The verb is a part of speech, which denotes changeable in time actions and states of objects, i.e. dynamic properties. From what has been said, it follows that one of the essential features of the verbs is to give information about the relation of these properties with other elements on the temporal axis, which has an absolute referential point – the act of speaking (now). The grammatical tense serves for expressing the different types of correlation of events and actions in time. Languages differ in the number and the types of tenses that they can express. The two languages, which this article discusses, share several common features, but main object of interest for us are the numerous distinctive features of the Bulgarian verb system, which make it interesting and difficult for formal description for the purposes of machine translation.

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. In this paper we propose a hypothesis that implementing statistical language modelling with rule-based machine translation can improve our knowledge not only about the relation of the verb systems of Bulgarian and English, but also about the structural dependencies between these two languages. In future we can use the results of this research for the purposes of achieving higher quality of translation and better understanding of both languages.

2. Main differences and similarities between the verb systems of both languages.

A well-known fact is that one of the most distinguishing properties of the Bulgarian language is its well-developed verb system. On one hand, regarding the semantic of the verb forms, the Bulgarian verb can have over 2000 forms with different grammatical meaning. The literature on Bulgarian tense system consists of many disagreements, mainly about the grammatical categories that it contains and the differential properties of these categories. In this paper we acknowledge the view of the Bulgarian tense category as a hyper-category, in which the meaning is formed by the relationship between the individual members of different categories, as it is possible to include additional elements that can

modify the meaning. (Gerdzhikov, 2000) The English temporal system share the same feature – it can be considered as a hyper-category, but a main difference regarding the Bulgarian temporal system is the much smaller number of grammaticalized meanings in English. Other distinctive feature of the Bulgarian verb is the potential to grammaticalize the indication of the nature of evidence for a given statement – the category of evidentiality (Nitsolova, 2008: 261). These characteristics contribute for the difficulties in the process of semantic transfer during translation.

On the other hand, the word formation of the verbs in both languages is very similar. Both languages have synthetic and analytic forms. While the synthetic forms can carry all the information in one single lexical unit, their number is significantly small on account of the analytic forms. Word formation of analytic forms in both languages uses main verb and different forms of auxiliary verbs. Main difference in Bulgarian is that only the verb "съм" ("*sam"* be)[1] in its different forms is used to form all the synthetic forms, thus this auxiliary verb carries most of the grammatical information, while in English several other auxiliary verbs combine with each other to form different meanings. Other distinctive feature of the word formation of the verb forms in Bulgarian is that both the main verb and the auxiliary verb can carry grammaticalized information about the grammatical gender of the doer of the action denoted by the verb. These differences contribute to the complicated lexical transfer between Bulgarian and English.

Of course the differences and the similarities of both languages` verb systems are much more numerous and complicated, in our introduction we try to outline the most essential ones, which lead to several difficulties of conducting lexical and semantic transfer regarding machine translation.

3. Differences and similarities of the semantics of the temporal systems of both languages.

As we said, both languages temporal systems share a common feature – they consist of categories within the hyper-category. Both Bulgarian and English have category that expresses a completed action in relation to a referential point – the perfect tenses. 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 semantic of the lexical unit and expresses the action as finished or unfinished related to the action`s own completion (Kucarov 2007:551). These two grammatical categories contribute to one of the great difficulties when translating from Bulgarian to English – altering the semantic information of one lexical grammatical category into morphological grammatical. While sometimes changes in meaning are not perceptible, most of the times we have two different meanings: *Чел съм романа/Прочел съм романа- I have read the novel*.

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 the source language needs to be reduced or unevenly distributed between different grammatical categories in the target language. Nevertheless, as it has been pointed out before (Lazarov, 2016), the characteristics of grammaticalized information in Bulgarian and English verb forms share numerous similarities. That is why we have similar grammatical meaning in most of the verb forms. We have to point out again that most of the grammaticalized information is lost during the semantic transfer between the grammatical categories of both languages. The grammatical number and person are grammaticalized by every form in Bulgarian and most of the verb forms carry information about the grammatical gender of the doer of the action, whereas in English most of the times we have tenses with only one form. In Table 1 we present as example the formal accordance in meaning between Bulgarian and English tenses and the ratio of the forms.

Bulgarian	**English**	Number of forms
Praesens	**Present simple/Present continuous tense**	
Person, number	3rd person, sg. num./1st person and 3rd person, sg.num	6:2/6:3

[1] The particles from Old Bulgarian language ща (*schta,* will) is also used the word formation of Futurum.

Aorist	**Past simple tense**	
Person, number	-	5:1
Imperfekt	**Past continuous time**	
Person, number	1st and 3rd person, sg. .umn.	5:2
Perfekt	**Present perfect tense /Present perfect continuous tense**	
person, number, gender	3rd person, sg. num.	12 :2
Plusquamperfekt	**Past perfect/Past perfect continuous tense**	
person, number, gender	-	9 :1
Futurum	**Future simple/Future simple continuous tense**	
Person, number	-	6:1
Futurum exactum	**Future perfect/future perfect continuous tense**	
person, number, gender	-	12:1
Futurum praeteriti	**Future simple tense in the past (*going to*)**	
Person, number	1st and 3rd person, sg. .num.	6: 2
Futurum exactum praeteriti	**Future perfect in the past /Future perfect continuous tense in the past**	
person, number, gender	-	12:1

Table 1: Accordance in meaning of tenses between Bulgarian and English.

The huge diversity of verb forms in Bulgarian leads to several problems when translating in English. For the purposes of transfer-based machine translation developing rules for all possible variations, although more reliable, can be time-consuming and hard. On one hand, the much smaller number of forms in English can be a great advantage, because a large number of forms in Bulgarian are transferred into a much smaller in English, thus the possible outcomes in the target language are equal to the number of the transfer rules (Lazarov, 2016). On the other hand, most of the forms, which grammaticalize meaning for evidentiality, voice and mood, have very low frequency and incomprehensible usage. Of course we have to point out that for the purposes of rule-based machine translation this problem can be resolved by providing more precise contextual rules. Our point of view is that these problems can be better studied and resolved by the method of statistical language modeling.

4. Towards the statistical method in machine translation.

As we said before, the rule-based method in machine translation is reliable, as it depends on language models, which are constructed by people – thus the knowledge of language is exterior – it is still the human competence of language. Essential for the rule-based method is the presence of large and accurate grammars and dictionaries, which must take into account all possible language variations. Needless to say, there is no such grammar that can describe human language in such depth and detail in all of its possible manifestations. Therefore we need to gather information about the language not by prescribing it, but by describing its actual usage – we need a grammar that prescribes probable language models, rather than describing theoretical ones.

In the short history of computational linguistics and machine translation we have achieved more than the fathers of this scientific field ever imagined and predicted. Starting from the basic understanding of language as a set of rules, nowadays we have opportunity to discover more and new inner dependencies throughout all natural languages. We are able not only to build grammatical models of languages, but also statistical ones.

The goal of statistical language modeling 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 (Song and Croft, 1999: 317). By expressing various language usages and deviations in terms of simple parameters in a statistical model, it can provide an easy way to deal with complex natural

language phenomena. Statistical language modeling (SLM) originated in the late 1980`s for the purposes of speech recognition, but it has also played a vital role in various other natural language applications like machine translation, part-of-speech tagging, intelligent input method, etc. It has passed through two periods of its development – word based SLM (1992) and phrase based SLM (2003). Main principle of statistical language modeling is more data is better data. For the purposes of statistical machine translation (SMT) we need enormous corpora with enough variable data in order to provide enough linguistic material. As we said, both Bulgarian and English are resource rich languages and both can provide sufficient data for research on their own and between them. Figure 1 shows the process of analyzing data in SMT.

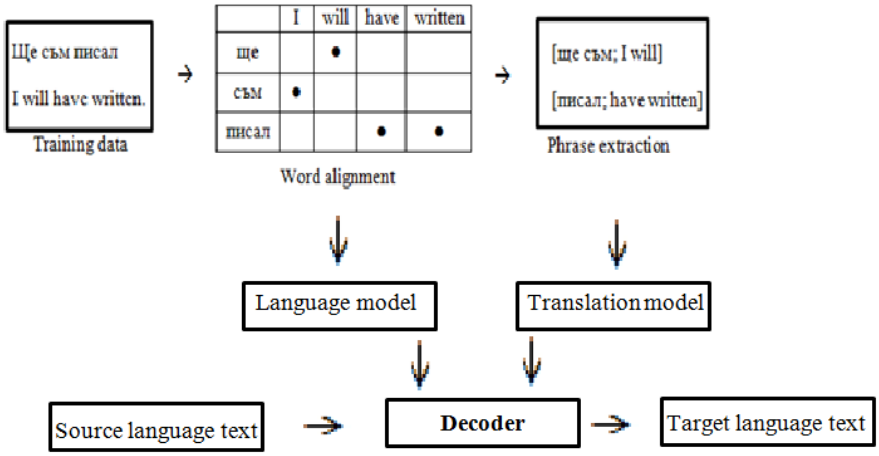

Figure 1: Simplified process of analyzing data from bilingual corpora.

For the purposes of statistical language modeling we need to know whether a given string of words is the right string of words in given language – we need to know what the probability of the string is. That is why we need to decompose this probability to the product of the probabilities of each word appearing in context of other words. Nowadays the n-gram model is the most widely used for language modeling and SMT. In a n-gram model, the probability $P(w_1,w_2,\ldots w_n)$ of observing the sentence $w_1,w_2,\ldots w_n$ is calculated as: $\prod_{i=1}^{m} P(w_i|w_1,w_2,\ldots w_{i-1}) \approx \prod_{i=1}^{m} P(w_i|w_{i-(n-1)}\ldots w_{i-1})$ or it is assumed that the probability of observing the i^{th} word w_i in the context history of the preceding $i-1$ words can be approximated by the probability of observing it in the shortened context history of the preceding $n-1$ words.

Of course statistical language modeling directly depends on quantity and quality of the avilable linguistic resources. Main principle of it 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. We need to specify that we can calculate the probability not only of words, but of any linguistic unit – phonemes, morphemes, words, phrases etc. In this paper we introduce the idea that we can build statistical language model of the verb systems of Bulgarian and English adding up two methods of machine translation.

5. A possible approach of applying SMT for the purposes of translating verb forms.

As it has been pointed out, there are formal similarities of the semantics of the temporal systems of Bulgarian and English. Nevertheless, transfer-based rules are not reliable enough to translate the majority of Bulgarian verb forms in English. This is why we propose the hypothesis that a possible collaboration between these two methods (rule-based MT and SMT) can be a solution to this problem.

We already pointed that transfer-based rules can provide information about the exact semantic and lexical transfer between the two languages of interest for us, nevertheless, in the case of translating from Bulgarian to English they cannot be prescribed with 100% certainty due to the huge amount of grammatical information that is lost during the process of translation, thus we need to construct a statistical language model of the transfer-based rules on their own. After that we could generate translation model, which is going to rely on the transfer-based rules. Our view is that using PoS-annotated corpora we can construct language model of the verb systems of both languages, which

language models will be able to present statistical information about the usage of different morphological categories and the frequency of some of the verb forms that have uncertain usage and vague meaning. After constructing these single language statistical models, we will be able to construct the statistical translation model of the transfer-based rules. By comparing the data from extracted verb forms from parallel corpora, we can see how frequently the data we have supports the usage of given transfer-based rules. In that way we will be able even to derive new rules, if the present ones do not get approval by the available linguistic data. The final stage of constructing the translation model is going to include verification of the gathered data by attempting to translate different types of other textual resources.

As we know the phrase based translation gives us better results, so as the verbal phrase in sentences constructs a whole syntactical unit, we can try to analyze the whole VP and build a language model based on its behavior. As we said before the asymmetric grammatical categories in English and Bulgarian contribute to the fact that large number of forms in Bulgarian have to be translated with much smaller number of forms in English. We propose that creating a language model of the verbal systems of both languages for the purposes of machine translation can be achieved by implementing the transfer-based rules with statistical language models. Our hypothesis is that a statistical language model of the verbal systems of both languages can complete the language model that the rule-based method composes. As we said, the rule-based method gives us strict information about the semantic and lexical transfer from one language to another, but in our case we have more coinciding verb forms in English for the Bulgarian forms. Taking into account what is the probability of certain verb forms to occur in English when we have a given forms in Bulgarian, we can prescribe our transfer-based rules with this certain probability. In this way we can have information based on actual data of language usage combined with exterior knowledge of language. Combining these two methods, we can relate verb forms with certain probability between languages. Also in cases where two or more verbal forms in English correspond to one in Bulgarian we will have statistical data of the probability of each corresponding form to occur in our target language and the context in which it can occur. This can help us theoretically establish any correlation between the lexical aspect of the verb in Bulgarian and the category of aspect in English. Another aspect in which implementing statistical and rule based machine translation can help us is to establish, based on various data, what is the statistical probability of certain verbal form to occur – as we know the verbal forms for evidentiality, mood and voice in Bulgarian tend to peter out at the expense of other more frequent forms which carry less grammatical information, but are more recognizable for the users of the language. The lost grammatical information is retrieved within the frames of the sentence. Thus if we get low probability for given verb form, we need better transfer-based rules.

By applying statistical language modeling to the rule based method, we can extract information about given language on its own. We can gain statistical information about the frequency of a certain verb forms in different kinds of texts, thus prescribing a probability of some verb forms to be in this kind of text. Based on the size and quality of the corpora we have, we can make conclusions about what type and what size of grammatical information is lost during the translation process, so in future we can try to figure out ways to prevent that by providing more contextual rules. This way we can also gather information about the cases in which we lose grammaticalized information because of dissimilarities in the working languages. A simplified chart of our linguistic model is presented in Figure 2.

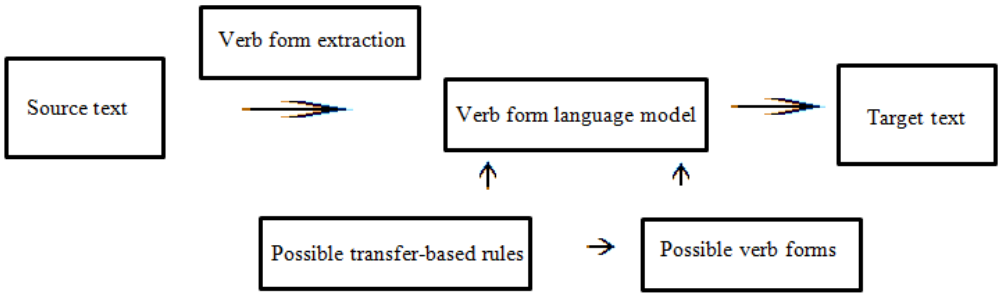

Figure 2: Chart of the stages of analyzing the verb forms in two languages.

6. Conclusion and future research

Of course we have to point out that our idea, although restricted to analyzing only the verb forms in Bulgarian and English, carries the possibility to give us new look at the linguistic data that we have of both languages. It is dependable on the quality and quantity of the corpora we have, it will also depend on the transfer rules we already have. The perfect corpora of Bulgarian and English for the purposes of our research must have aligned sentences with several types of annotations – morphological, syntactical, semantical and also information about the possible transfer rules and their variations. The available Bulgarian National Corpus, the Bulgarian PoS annotated corpus and the British National Corpus are suitable for constructing the preliminary single language models, in order to gather data about the frequency of usage of the verb forms. For the purposes of constructing the translation model we will need parallel corpora with PoS annotation such as the Bul-X-Cor and also other parallel language corpora will be suitable after careful PoS-annotation. Our future research will include, first of all, gathering and analyzing the available corpora. Extracting all the verb phrases and the context in which they appear. After we analyze this information, we can continue with constructing our verb language model, which must also include information about all possible derivations from the available data we have. The final stage of our work will include comparison of the two fundamental methods we use – transfer-based and statistical, in order to find out what kind and what number of mistakes each of them makes and how they piece out. In that way combining the two main methods of machine translation – rule based and statistical, we will be able to study English and Bulgarian verb systems on their own and also to find the deep inner dependencies between both languages that are in the middle of our linguistic competence and performance.

References

Gerdzhikov, G. (2000). Kategoriyata vreme kato hiperkategoriya. *Bulgarian language and literature – educational journal*, Ministry of Education and Science. http://liternet.bg/publish/ggerdzhikov/hyper.htm

Hutchins, W. J. (1986). *Machine Translation: Past, Present, Future*. Ellis Horwood, Chichester, UK. Halstead Press, New York.

Kucarov, I. (2007). *Teoretitshna gramatika na balgarskiya ezik. Morfologiya.* University Press "Paisii Hilendarski".

Lazarov, T. (2016). Osobenosti na glagolnite sistemi i natshinite za izrazyavane na vremeto v balgarski i anglijski. Semantitshen transfer pri prevod ot balgarski na anglijski. *Littera et Lingua – electronic journal*, Sofia University. http://slav.uni-sofia.bg/naum/en/lilijournal/2015/12/1-2/tlazarov

Nitsolova, R. (2008). *Balgarska gramatika. Morfologiya*. University Press "St. Kiment Ohridski".

Song, F. and W. B. Croft (1999). A General Language Model for Information Retrieval. In *Proceedings of the Eight International Conference on Information and Knowledge Management.* ACM, New York, USA.

DEPARTMENT OF
COMPUTATIONAL
LINGUISTICS

http://dcl.bas.bg/clib/

ISSN: 2367-5675