FOURTH
INTERNATIONAL
CONFERENCE

CLIB '20

COMPUTATIONAL LINGUISTICS IN BULGARIA CLIB 2020

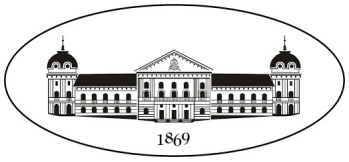
25 – 26 June 2020

Sofia, Bulgaria

PROCEEDINGS

The Special Session on Wordnets and Ontologies at the Fourth International Conference *Computational Linguistics in Bulgaria (CLIB 2020)* is organised with the support of the National Science Fund of the Republic of Bulgaria under the project *Towards a Semantic Network Enriched with a Variety of Relations*, Grant Agreement DN10/3/2016.

CLIB 2020 is organised by:



Department of Computational Linguistics
Institute for Bulgarian Language

Institute for Information and Communication Technologies

Bulgarian Academy of Sciences

PUBLICATION AND CATALOGUING INFORMATION

Title:	Proceedings of the Fourth International Conference *Computational Linguistics in Bulgaria* (CLIB 2020)
ISSN:	2367 5675 (online)
Published and distributed:	Bulgarian Academy of Sciences
Editorial address:	Institute for Bulgarian Language Bulgarian Academy of Sciences 52 Shipchenski Prohod Blvd., Bldg. 17 Sofia 1113, Bulgaria +3592/ 872 23 02

Proceedings of the

Fourth International Conference

Computational Linguistics in Bulgaria

25 – 26 June 2020
Sofia, Bulgaria

PROGRAMME COMMITTEE

Chair:

Svetla Koeva – Institute for Bulgarian Language, Bulgarian Academy of Sciences

Co-chair:

Petya Osenova – Institute of Information and Communication Technologies, Department of Linguistic Modelling and Knowledge Processing, Bulgarian Academy of Sciences / Sofia University, Faculty of Slavic Studies

Galia Angelova – Institute of Information and Communication Technologies, Department of Linguistic Modelling and Knowledge Processing, Bulgarian Academy of Sciences

Iana Atanassova – University of Burgundy, Centre for Interdisciplinary and Transcultural Research

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

Svetla Boytcheva – Institute of Information and Communication Technologies, Department of Linguistic Modelling and Knowledge Processing, Bulgarian Academy of Sciences

Khalid Choukri – Evaluations and Language Resources Distribution Agency

Ivan Derzhanski – Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

Tsvetana Dimitrova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences

Mila Dimitrova-Vulchanova – Norwegian University of Science and Technology

Radovan Garabík – Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences

Maria Gavrilidou – Institute for Language and Speech Processing, Natural Language Processing and Knowledge Extraction Department

Stefan Gerdjikov – Sofia University, Faculty of Mathematics and Informatics

Kjetil Rå Hauge – University of Oslo, Department of Literature, Area Studies and European Languages, ILOS

Ivan Koychev – Sofia University, Faculty of Mathematics and Informatics

Zornitsa Kozareva – Google

Cvetana Krstev – University of Belgrade, Faculty of Philology

Eric Laporte – University of Paris-Est Marne-la-Vallée

Bernardo Magnini – Bruno Kessler Center in Information and Communication Technology

Ruslan Mitkov – University of Wolverhampton

Preslav Nakov – Qatar Computing Research Institute, HBKU

Ivelina Nikolova – Institute of Information and Communication Technologies, Department of Linguistic Modelling and Knowledge Processing, Bulgarian Academy of Sciences

Kemal Oflazer – Carnegie Mellon University in Qatar

Maciej Piasecki – Wrocław University of Technology

Vito Pirrelli – Institute for Computational Linguistics, ILC-CNR

Stan Szpakowicz – University of Ottawa

Ivelina Stoyanova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences

Marko Tadić – University of Zagreb, Faculty of Humanities and Social Sciences, Department of Linguistics

Hristo Tanev

Irina Temnikova – Mitra Translations

Tinko Tinchev – Sofia University, Faculty of Mathematics and Informatics

Maria Todorova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences

Dan Tufis – Research Institute for Artificial Intelligence, Romanian Academy

Cristina Vertan – University of Hamburg

Victoria Yaneva – University of Wolverhampton

Katerina Zdravkova – University St Cyril and Methodius in Skopje

ORGANISING COMMITTEE

Chair:

Svetlozara Leseva – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences

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

Zara Kancheva – Institute of Information and Communication Technologies, Department of Linguistic Modelling and Knowledge Processing, Bulgarian Academy of Sciences

Hristina Kukova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences

Ivaylo Radev – Institute of Information and Communication Technologies, Department of Linguistic Modelling and Knowledge Processing, Bulgarian Academy of Sciences

Valentina Stefanova – Institute for Bulgarian Language, Department of Computational Linguistics, Bulgarian Academy of Sciences

Ekaterina Tarpomanova – Sofia University, Faculty of Slavic Studies

Table of Contents

PLENARY TALKS

TAG SENSE DISAMBIGUATION IN LARGE IMAGE COLLECTIONS: IS IT POSSIBLE?

Prof. D.Sc. Galia Angelova (Institute of Information and Communication Technologies, Bulgarian Academy of Sciences)

Automatic identification of intended tag meanings is a challenge in large annotated image collections where human authors assign tags inspired by emotional or professional motivations. This task can be viewed as part of the AI-complete problem to integrate language and vision. Algorithms for automatic Tag Sense Disambiguation (TSD) need "golden" collections of manually created tags to establish baselines for accuracy assessment. In this talk the TSD task will be presented with its background, complexity and possible solutions. An approach to use WordNet senses and Lesk algorithm proves to be successful but the evaluation was done manually for a small number of tags. Another experiment with the MIRFLICKR-25000 image collection will be presented as well. Word embeddings create a specific baseline so the results can be compared. The accuracy achieved in this exercise is 78.6%.

By improving TSD and obtaining high quality synsets for the image tags, we are actually supporting the machine translation of the large annotated image collections to languages other than English.

CLINICAL NATURAL LANGUAGE PROCESSING IN BULGARIAN

Assoc. Prof. Svetla Boytcheva (Institute of Information and Communication Technologies, Bulgarian Academy of Sciences)

Healthcare is a data intense domain. A large amount of patient data is generated daily. However, more than 80% of this information is stored in an unstructured format – as clinical texts. Usually, clinical narratives contain a description with telegraph-style sentences, ambiguous abbreviations, many typographical errors, lack of punctuation, concatenated words, and etc. Especially in the Bulgarian context – medical texts contain terminology both in Bulgarian, Latin and transliterated Latin terminology in Cyrillic, that makes the task for text analytics more challenging. Recently, with the improvement of the quality of natural language processing (NLP), it is increasingly recognized as the most useful tool for extracting clinical information from free text in scientific medical publications and clinical records. Natural language processing (NLP) of non-English clinical text is quite a challenge because of the lack of resources and NLP tools. International medical ontologies such as SNOMED, MeSH (Medical Subject Headings), and the UMLS (Unified Medical Languages System) are not yet available in most languages. This necessitates the development of new methods for processing clinical information and for semi-automatically generating medical language resources. This is not an easy task because of the lack of a sufficiently accessible repositories with medical records, due to the specific nature of the content, which contains a lot of personal data and specific regulations for their access.

In this talk will be discussed the multilingual aspects of automation Extract text from clinical narratives in the Bulgarian language. This is very important task for medical informatics, because it allows the automatic structuring of patient information and the generation of databases that can be further investigated by retrieving data to search for complex relationships. The results can help improve clinical decision support, diagnosis and treatment support systems.

DETECTING THE FAKE NEWS AT ITS SOURCE, MEDIA LITERACY, AND REGULATORY COMPLIANCE

Dr. Preslav Nakov (Qatar Computing Research Institute, Hamad Bin Khalifa University)

Given the recent proliferation of disinformation online, there has been also growing research interest in automatically debunking rumors, false claims, and "fake news". A number of fact-checking initiatives have been launched so far, both manual and automatic, but the whole enterprise remains in a state of crisis: by the time a claim is finally fact-checked, it could have reached millions of users, and the harm caused could hardly be undone. An arguably more promising direction is to focus on fact-checking entire news outlets, which can be done in advance. Then, we could fact-check the news before they were even written: by checking how trustworthy the outlets that published them are.

We will show how we do this in the Tanbih news aggregator (`http://www.tanbih.org/`), which aims to limit the effect of "fake news", propaganda and media bias by making users aware of what they are reading. The project's primary aim is to promote media literacy and critical thinking, which are arguably the best way to address disinformation and "fake news" in the long run. In particular, we develop media profiles that show the general factuality of reporting, the degree of propagandistic content, hyper-partisanship, leading political ideology, general frame of reporting, stance with respect to various claims and topics, as well as audience reach and audience bias in social media. We further offer explainability by automatically detecting and highlighting the instances of use of specific propaganda techniques in the news (`https://www.tanbih.org/propaganda`).

Finally, we will show how this research can support broadcasters and content owners with their regulatory measures and compliance processes. This is a direction we recently explored as part of our TM Forum & IBC 2019 award-winning Media-Telecom Catalyst project on AI Indexing for Regulatory Compliance, which QCRI developed in partnership with Al Jazeera, Associated Press, RTE Ireland, Tech Mahindra, V-Nova, and Metaliquid.

DEMONSTRATION OF THE EUROPEAN LANGUAGE GRID

Dr. Georg Rehm (Speech and Language Technology Lab, German Research Center for Artificial Intelligence

With 24 official EU and many additional languages, multilingualism in Europe and an inclusive Digital Single Market can only be enabled through Language Technologies (LTs). European LT business is dominated by hundreds of SMEs and a few large players. Many are world-class, with technologies that outperform the global players. However, European LT business is also fragmented – by nation states, languages, verticals and sectors, significantly holding back its impact. The European Language Grid (ELG) project addresses this fragmentation by establishing the ELG as the primary platform for LT in Europe. The ELG is a scalable cloud platform, providing, in an easy-to-integrate way, access to hundreds of commercial and non-commercial LTs for all European languages, including running tools and services as well as data sets and resources. Once fully operational, it will enable the commercial and non-commercial European LT community to deposit and upload their technologies and data sets into the ELG, to deploy them through the grid, and to connect with other resources. The ELG will boost the Multilingual Digital Single Market towards a thriving European LT community, creating new jobs and opportunities. Furthermore, the ELG project organises two open calls for up to 20 pilot projects, one of which was recently closed. The presentation will give an overview of the European Language Grid project and it will also contain a demonstration of the emerging ELG technology platform.

MAIN CONFERENCE

A Corpus-Based Study of Derivational Morphology and Its Theoretical Implications

Junya Morita
Kinjo Gakuin University
College of Humanities
morita@kinjo-u.ac.jp

Abstract

The present study investigates the formal and semantic properties of derivational morphology, dealing in particular with *-able* derivatives in English (e.g. the recorder is *pocketable*). Focusing principally on hapax legomena in a large corpus, a reliable indicator of online coinage, *-able* derivatives are extracted from it. Detailed observation of them is carried out and then their theoretical analysis is conducted in the framework of generative morphology. The data analysis elucidates (i) a core aspect of *-able*: it productively attaches to transitive verbs to produce modalized passive adjectives whose external arguments are restricted to Theme arguments and (ii) a peripheral facet: the basic meaning of *-able* as well as its prototypical base category and external argument are extended, on a small scale, to other kinds of meaning and category. Based on these empirical observations, major and minor formation rules are proposed to deal respectively with regular and sub-regular *-able* derivation.

Keywords: *-able* adjectives, hapax legomena, generative morphology, word formation rules, English

1. Introduction

The system of derivational morphology contributes greatly to children's acquisition of vocabulary by enabling them to generate an infinite number of nominal, verbal, and adjectival complex words. The primary task of generative morphology is then to reveal the regularities of word formation processes and provide a principled account of them. As part of this enterprise, the present study attempts to show how the system works in producing *-able* final words in English, as seen in "a skilled and constantly *re-skillable* workforce (BNC FA8: 1682)." The adjective *re-skillable* has the modalized passive sense of 'can be re-skilled' and it is a hapax—token frequency 1—of a large corpus, and hence it is a new coinage, which is constructed online without being stored in the lexicon. The aim of the present work is to demonstrate how new *-able* adjectives are created systematically by using *-able* words detected in a large-scale corpus and provide a generative-theoretic characterization of the process. This article is organized as follows: after outlining some points of previous studies in §2, we inspect them on the basis of our data analysis (§3) and present theoretical implications for the results of our research (§4). A summary of the main arguments is presented in §5.

2. Previous Studies

-Able has been well observed in the literature from a descriptive perspective: Jespersen, 1949; Marchand, 1969; Quirk et al., 1985. There are many treatments of the suffix in the generative

literature, including Chapin, 1967; Aronoff, 1976; Williams, 1981; Di Sciullo, 1997. A review of the literature identifies four attributes that merit special attention: 1 the formation of deverbal, especially transitive-based, *-able* words is very productive (Jespersen, 1949; Quirk et al., 1985); 2 *-able* attaches only to transitive verbs, ergative verbs, and nouns (Di Sciullo, 1997); 3 *-able* prototypically makes an adjective with a mixture of passive and 'potential' senses (Jespersen, 1949; Chapin, 1967); 4 the external argument of *-able* words is restricted to a Theme argument (Williams, 1981). The first and third points are clear and easy to understand. The second point is that *-able* can affix to transitives (*cuttable cost*), ergatives (*burnable box*), and nouns (*knowledgeable staff*), but not unergatives (**runnable old man*) or unaccusatives (**arriv(e)able boy*). The fourth attribute is demonstrated by the contrastive acceptability of (a) *those things are promisable* (Theme), (b) **those people are runnable* (Agent), and (c) **those people are promisable* (Goal) (Williams, 1981: 93).

3. Observation and Generalizations

This section inspects the four points of previous studies by an in-depth observation of *-able* words and presents generalizations based on it. We will begin by pointing out the method of research and the resulting data. By repeatedly using the "wild card" function of a research engine, the frequency of words ending in *-able* is checked to find hapaxes in the British National Corpus (BNC), a 100-million-word corpus.[1] As for ascertaining the total number of types of *-able* words, we make a list of those which are included in Lehnert, 1971 and attested in BNC. A case in which a prefix occurs outside an *-able* adjective (e.g. *un[washable]*) and a compound of the kind *hand-breakable* (synthetic compounds) are left out of consideration. As a result of the research, we have gained 662 word types in *-able* including 209 hapaxes.

3.1. Productivity of *-able* Affixation

Productivity is defined as the extent to which a word formation device can give rise to new words (Lieber, 2010: 59). There have been several approaches to quantifying productivity, but the most reliable is the one which puts great importance to hapax legomena of a large-scale corpus (Baayen and Renouf, 1996; Plag, 1999). This is based on the view that complex forms that have been observed only once in a large corpus are highly likely to be lexical innovations and hence the capacity of a word formation rule to create new forms crucially involves the degree to which the rule produces words with extremely low frequency (Hay, 2003). Baayen and Renouf, 1996: 73 propose a productivity measure: *Productivity* (*P*)=n_1/N, where n_1 is the number of hapaxes and N is the total number of tokens. Here we revise it so as to place the total number of types (but not tokens) in the denominator; thus, $P=n_1/V$ (V: the number of word types). This is derived from the view that the productivity of a particular process is reflected in the type frequency of the process (Goldberg, 1995: 134-139).

According to the proposed measure, we calculate the productivity values of three classes of *-able*: (i) one which attaches to a verb, (ii) one which joins to a noun, and (iii) one which adjoins to a non-word; verb-attaching *-able* is further divided into three subclasses. The results of the research can be provided in tabular form.[2]

[1]For this hapax-detection I am indebted to the research engine of www.english-corpora.org (BNC).
[2]If the base of an *-able* adjective can be a noun or verb (*issuable*), it is counted separately, that is, we have two word types in *-able*. Similarly, if an *-able* base can be a transitive or unergative (*breathable*), the *-able* word is also counted separately.

-able classes	hapaxes (n_1)	types (V)	productivity (P)	examples
verb-attaching	200	595	0.336	
(a) transitive	170	524	0.324	*affirmable*, *buildable*
(b) ergative (tr/intr)	17	46	0.370	*burnable*, *connectable*
(c) ergative (intr)	7	10	0.700	*crackable*, *cloggable*
noun-attaching	6	39	0.154	*inquestable*, *networkable*
stem-attaching	5	31	0.161	*hereditable*, *satiable*

Table 1: Productivity values of three main classes of *-able*.

Table 1 shows that deverbal *-able* affixation (P=0.336) is much more productive than denominal *-able* affixation (P=0.154) and stem-based *-able* affixation (P=0.161). We see that transitive-based *-able* affixation (P=0.324) and *-able* affixation based on the ergatives which are interpreted as transitives and intransitives (cf. *burnable box*) (P=0.370) are as productive as the one whose bases are verbs in general (P=0.336). There is a set of *-able* words which are based on ergative verbs of intransitive use, as in *crackable walls*. Their productivity value would be very high in the present measure (P=0.700). It should be noted that the total number of types of these *-able* words is very low (V=10). In this connection, Baayen and Lieber, 1991: 818-819 suggest "the global productivity P^*": P^* of an affixation rule is defined in terms of its coordinates in the P-V interaction region, with productivity (P) on the horizontal axis and types (V) on the vertical axis; a productive affix occupies a central position in the region. By this definition, a case where the number of word types is very low like the one in question falls outside the domain for productive process.[3] It can thus be concluded that *-able* fruitfully joins to transitive verbs, but not to intransitive verbs, nouns, or non-word stems.

3.2. Syntactic Categories of *-able* Bases

This section inspects the second point of the previous studies: *-able* attaches only to transitive verbs, ergative verbs, and nouns. The 662 *-able* word types obtained are classified in terms of the syntactic (sub-)categories of their bases: transitives, ergatives, unergatives, unaccusatives, nouns, and stems. An ergative verb engages in a construction where the same noun can be used as the subject when the verb is intransitive and as the object when it is transitive, while an unergative (intransitive) verb describes an action performed by a human actor endowed with consciousness and volition; an unaccusative (intransitive) verb denotes a phenomenon that happens spontaneously without the intervention of any causer (cf. Lyons, 1968; Randall, 2010; Kageyama, 2012). Table 2 displays the percentage of each category class of *-able* bases.

base categories	hapaxes (n_1)	types (V)	examples
transitive verb	170 (80.6%)	524 (78.8%)	*delimitable*, *pardonable*
ergative verb	24 (11.4%)	56 (8.4%)	*digestable*, *fermentable (material)*
unergative verb	2 (0.9%)	8 (1.2%)	*swimmable*, *walkable*
unaccusative verb	4 (1.9%)	7 (1.0%)	*perdurable*, *risable*
noun	6 (2.8%)	39 (5.9%)	*exceptionable*, *presidentiable*
stem (non-word)	5 (2.4%)	31 (4.7%)	*dubitable*, *effable*
total:	211 (100%)	665 (100%)	

[3]We have obtained 8 types of unergative-based *-able* words including 2 hapaxes and 7 types of unaccusative-based ones including 4 hapaxes, all of which are not referred to in Table 1. They can be handled in much the same way.

Table 2: Distribution of each syntactic (sub-)category of *-able* bases.

The results of the inquiry indicate the pros and cons of the claim made by Di Sciullo: greater than 90% hapaxes in *-able* are based on transitives (80.6%), ergatives (11.4%), or nouns (2.8%), supporting the generalization that the bases of *-able* are transitives, ergatives, or nouns. (The same argument applies to the results of research on word types in *-able*.) On the other hand, the results disclose that *-able* can be added to unergatives (0.9%) and unaccusatives (1.9%) in a certain limited way. This is well exemplified in "*fishable*, *swimmable* water (BNC B7L: 669)" and "Puzznic is *lastable* (BNC EB6: 2276)," respectively.

The reason why unergatives and unaccusatives may be combined with *-able* has to do with the meanings of *-able* derivatives. The next issue, then, is to classify their meanings into subgroups and show how the submeanings of *-able* words are related to their base categories.

3.3. The Meanings of *-able* Derivatives

Let us now consider the claim advanced by Jespersen, 1949 and Chapin, 1967: deverbal adjectives in *-able* primarily have modalized passive senses. We examine the total of 206 deverbal and denominal *-able* hapaxes extracted from BNC. The reason for targeting *-able* hapaxes is that we focus on observation of what meaning is assigned to a derivative when it is instantly innovated and that a hapax in a large corpus is a significant indicator of this. The results of the research are offered in Table 3, where the meanings of *-able* words are divided into four submeanings and their base categories are divided into six classes.

meaning: / base categories:	(i) 'able to be V-ed'	(ii) 'should be V-ed'	(iii) 'apt to V/ to be V-ed'	(iv) 'suitable for'
a. transitive	159	4	4	3
b. ergative (tr/intr)	17			
c. ergative (intr)			7	
d. unergative				2
e. unaccusative			4	
f. noun				6
total: 206 (100%)	176 (85.4%)	4 (2.0%)	15 (7.3%)	11 (5.3%)

Table 3: Relation between the submeanings of *-able* words and their base categories.

The notable findings of the research lead to three empirical generalizations. To begin with, in agreement with Jespersen and Chapin, the primary meaning of *-able* words is a mixture of passive and potential senses; greater than 80% of new words in *-able* have this sense (e.g. transitives: *affirmable*, *bitable*, *chaseable* and ergatives (tr/intr): *contrastable*, *diminishable*, *filterable*). A pertinent example is given in "… they raise at least the possibility of a belief being *affirmable* (BNC HYB: 1789)." The submeanings (ii)-(iv) of *-able* are therefore judged to be non-central ones. Importantly, there is a clear correlation between these meanings and the classes of its base. The first one is that *-able* which adjoins to an intransitive verb to coin a new word exhibits a strong tendency to bear the reading 'apt to V'; greater than 80% of intransitive-based hapaxes have this reading (e.g. ergatives (intr): *cleavable*, *cloggable*, corrodable, *crackable*, *digestable*, *smellable*, *smudgeable* and unaccusatives: *lastable*, *perdurable*, *risable*, *swayable*). A good example is given in "If diamonds are the hardest of minerals they also among the most *cleavable* (BNC FBA: 1088)."

The second correlation is that noun-incorporating new *-able* adjectives have only the

sense 'suitable for' (e.g. *filmable*, *inquestable*, *microwaveable*, *networkable*, *presidentiable, raceable*). This is exemplified in "… to translate the confusion … into *filmable* dialogue (BNC AP0: 991)." Note that according to a comprehensive dictionary, the meanings of denominal *-able* adjectives are broadly divided into two kinds: 'of the nature or quality of' (cf. *knowledgeable*) and 'suitable for.' That *-able* hapaxes have only the latter meaning provides evidence that the former meaning is not involved in the creation of new *-able* words, which is only found in some well-established *-able* derivatives.[4]

3.4. Restriction on External Argument

Finally, we turn to a restriction on external argument pointed out by Williams, 1981: an *-able* adjectival can be predicated only of a Theme phrase. Denominal and stem-based *-able* words are excluded from our analysis, since (non-derived) nouns and stems are irrelevant to arguments. In total, 595 word types in *-able* including 200 hapaxes are obtained and they are classified in terms of the thematic roles of their external arguments. Table 4 indicates the ratio of *-able* words involving each thematic role which the external argument assumes:

external arg	hapaxes (n_1)	types (V)	examples
Theme	198 (99.0%)	581 (97.7%)	mailable, maintainable, manageable
Location	1 (0.5%)	9 (1.5%)	fishable, fordable, habitable, ridable, swimmable
others	1 (0.5%)	5 (0.8%)	attainable, reachable (Goal), escapable (Source), passable (Path), kickable (Time)
total :	200 (100%)	595 (100%)	

Table 4: Distribution of the thematic roles of external argument.

The condition under scrutiny is almost confirmed by this research. We can see that 99% of *-able* hapaxes and about 98% of *-able* types take Theme as their external arguments. It is worth noting, however, that there exist cases which are inconsistent with this condition; the external argument is satisfied by a non-Theme phrase. For example, in "The Thames at Abingdon was barely *fishable* (BNC A6R: 1594)," the Location argument *the Thames* of the underlying base verb *fish* occupies the external position of *-able* construction. Similarly, the Source argument *plastic boats* in the following example takes place in the position at issue: "… the development of high molecular density polyethylene has made plastic boats much more *escapable* … (BNC G27: 827)."

4. Theoretical Implications

4.1. Core Word Formation Rule

We have shown that (i) *-able* is productively added to transitive verbs to yield modalized adjectival passive words and their external arguments are generally assigned the Theme interpretation and (ii) *-able* is rather peripherally involved in other kinds of bases and external arguments and produces a limited number of adjectival "active" words. The former process constitutes the core domain of *-able* affixation and the latter can be called peripheral *-able* affixation, lying just outside the core. Let us discuss core *-able* affixation first.

As shown in §3.1, a large number of transitive-based *-able* derivatives are coined temporarily by some form of device. The creativity of related *-able* derivation lends support

[4] *-Ic*, *-ous*, and *-ive* are competing productive suffixes with the meaning 'of the nature or quality of,' and so hapaxes such as *dinosauric*, *foamous*, and *defunctive* block the use of corresponding *-able* words in this meaning.

to Antilexicalism, which holds that word formation takes place outside the lexicon so that a creative aspect of sentence and word construction is uniformly captured in syntax (Halle and Marantz, 1994). In a current theory of Antilexicalism, derived words are constructed by inserting an affix in an appropriate syntactic node based on its formalized lexical entries (Harley and Noyer, 2000; Embick, 2010). The relevant information on *-able* word formation can then be formalized into the core lexical entries of *-able*, as demonstrated in (1):

(1) Core lexical entries of *-able* (major rule)

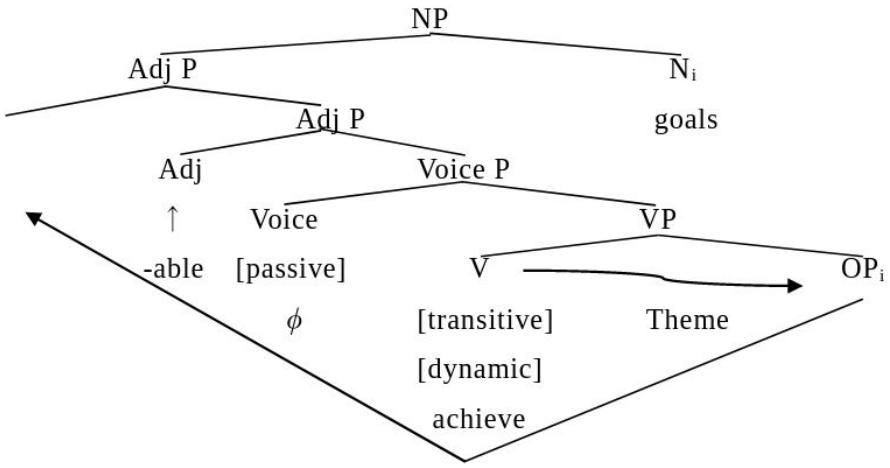
(i) internal features	(ii) meaning	(iii) license environment	(iv) argument
[A][property][modal]	'potential'	+<Voice [pass], [transitive, dynamic]>	+<DP [Theme]>

The definitional features of *-able* are listed in (i); the features [A], [property], and [modal] designate the permanent nature and modality of *-able* adjectives. The meaning of *-able* in (ii) together with the category Voice [passive] in (iii) indicate that the essence of *-able* derivatives is to designate the modalized property of an entity receiving the action of the verb. The license environment of the suffix is put in (iii), according to which *-able* connects to Voice phrase whose lexical head is a dynamic transitive verb and hence unpassivizable stative verbs like *have* are ruled out as the base of *-able* (cf. **hav(e)able*). We here assume "Generalized subcategorization," which enables subcategorization features to include not only the features of the whole category but also those of its lexical head (Emonds, 2000: 286). Thus, *-able* can relate to the features [transitive][dynamic] ascribed to the lexical head within the Voice P, as will be shown in (2) below. It should be emphasized that *-able* freely attaches to dynamic transitive verbs, each item with which it combines being unspecified in the lexical entries.[5] And finally, the external argument of *-able* is specified as in (iv), which allows only Theme argument to occupy the external position of *-able* adjectival.

Let us briefly look at how *-able* is inserted into the terminal node of a syntactic output. Adopting basically the structure of adjectival passive proposed by Bruening, 2014 and assuming that *-able* construction is formed by merging an adjectivizing head with Voice P, the underlying structure of *achievable goals* will be as depicted in (2). The null operator (OP), which occupies the internal argument position, is assigned the Theme role by the head verb (*achieve*). It is then moved to the specifier position of Adj P in passive environments and linked to the noun (*goals*), which is external to the *-able* adjectival. Thus, the external argument (*goals*) fulfills the Theme externalization constraint in (1iv), requiring external argument to bear the Theme role. The adjectivizing head can select Voice as well as its lexical head verb with transitive and dynamic features, whereby the license condition of (1iii) is satisfied. Consequently, *-able* is correctly inserted under the Adj node.

[5]We postulate the distinction between the well-formedness and actuality of a word: possible words may or may not be actual words (cf. Kuiper and Allan, 2004: 35). Thus, transitive-based *-able* words which are unregistered in large dictionaries and do not appear (token frequency 0) in BNC (e.g. *delayable*, *finishable*, *offerable*) are judged as possible but non-occurring words.

(2) *achievable goals*



4.2. Peripheral Word Formation Rules

As argued above, there are a kind of sub-regularities in *-able* affixation that can be characterized as follows: (i) *-able* can atypically be added to ergative, unergative, and unaccusative intransitive verbs as well as nouns (§3.2), (ii) *-able* derivatives can occasionally have non-passive and non-potential senses (§3.3), and (iii) a non-Theme phrase can exceptionally appear in the external position of *-able* adjectivals (§3.4). These kinds of information can be built into the noncentral or peripheral lexical representations of *-able*, as demonstrated in (3). Related rules are called "minor rules" in the sense of Lakoff, 1970:44; there are a set of minor *-able* word formation rules which apply only to exceptional cases.

(3) Peripheral lexical entries of *-able* (minor rules)
(i) internal features (ii) meanings (iii) license environments (iv) argument
[A][property][modal]
(a) 'should be V-ed' +<(Voice), V > +<N > +<DP >
(b) 'apt to V'
(c) 'suitable for'
(d) 'of the nature/quality of'

The specifications in (3ii) indicate that *-able* derivatives may have the meanings of 'should be V-ed,' 'apt to V,' 'suitable for,' and 'of the nature/quality of.' The subclasses of base verbs are unspecified in (3iii), with the result that *-able* may be suffixed to a variety of verbs including ergative, unergative, unaccusative intransitive verbs, and even stative verbs.[6] Likewise, the θ-roles of arguments are left unmarked in (3iv), since the external position may be occupied by a variety of arguments including Location, Goal, Source, Path, and Time arguments.

As indicated in Section 3.3, there is a correlation between the meanings of (3ii) and the license environments of (3iii). Any feature which can be predicted on the basis of other features is said to be redundant. To simplify the form of descriptions, such redundancy should be removed by some kind of redundancy rule. We can then formulate two redundancy rules for minor *-able* affixation: (i) +<V [ergative/unaccusative (intr)]> → meaning (b) and (ii)

[6]Although *-able* generally does not attach to *have* as a stative verb, it may adjoin to this verb in a certain limited context, as in "It kept them apart, kept them foreign to each other, him *unhaveable*, her unhad (BNC A0U: 893)."

+<V [unergative]>/+<N> → meaning (c). Rule (i) signifies that if *-able* attaches to an ergative or unaccusative (intransitive) verb, the *-able* word expresses the reading of 'apt to V' and rule (ii) implies that when *-able* is suffixed to an unergative verb or noun, the derived word bears the reading of 'suitable for.' It may thus be concluded that any morphological phenomena which are not accounted for by core rules will have to be specified as a set of systematic exceptions to the general mechanism in the form of minor rules.

5. Conclusion

On the basis of close analysis of the *-able* coinages discerned in a large corpus, we have identified a number of formal and semantic properties of *-able* derivation. We have then proposed that these properties are formalized into two kinds of formation rules from the perspective of generative morphology; one is central, basic, and productive, while the other is peripheral, derivative, and unproductive. The major rule represents the creative potential of derivational processes that enables us to produce and understand novel coinages, whereas the minor rules explain the observed sub-regularities of *-able* derivation. How best to relate these rules systematically awaits further investigation. Hopefully, the present study will provide a good example of what can be achieved by a corpus-based study of derivational morphology.

Acknowledgement

I would like to express my gratitude to three anonymous reviewers for their valuable comments and suggestions on an earlier draft of this paper. This work is partly supported by a Grant-in-Aid for Scientific Research (C) (No. 17K02697) from the Japan Society for the Promotion of Science.

References

Aronoff, M. (1976). *Word Formation in Generative Grammar*. Cambridge, MA: MIT Press.

Baayen, H. and Lieber, R. (1991). Productivity and English Derivation: A Corpus-Based Study. *Linguistics*, 29:801-843.

Baayen, H. and Renouf, A. (1996). Chronicling *the Times*: Productive Lexical Innovations in an English Newspaper. *Language*, 72:69-96.

Bruening, B. (2014). Word Formation Is Syntactic: Adjectival Passives in English. *Natural Language & Linguistic Theory*, 32: 363-422.

Chapin, P. G. (1967). *On the Syntax of Word-Derivation in English*. Bedford, MA: MITRE.

Di Sciullo, A. (1997). Selection and Derivational Affixes. In Dressler, W. U., Prinzhorn, M., and Rennison, J. R., Eds., *Advances in Morphology*, pages 79-95. Berlin: Mouton de Gruyter.

Embick, D. (2010). *Localism versus Globalism in Morphology and Phonology*. Cambridge, MA: MIT Press.

Emonds, J. E. (2000). *Lexicon and Grammar: The English Syntacticon*. Berlin: Mouton de Gruyter.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Halle, M. and Marantz, A. (1994). Some Key Features of Distributed Morphology. *MIT Working Papers in Linguistics*, 21:275-288.

Harley, H. and Noyer, N. (2000). Formal versus Encyclopedic Properties of Vocabulary: Evidence from Nominalisations. In Peeters, B., Ed., *The Lexicon-Encyclopedia*

Interface, pages 349-374. Amsterdam: Elsevier.
Hay, J. (2003). *Causes and Consequences of Word Structure*. New York: Routledge.
Jespersen, O. (1949). *A Modern English Grammar on Historical Principles* VI. London: George Allen and Unwin.
Kageyama, T. (2012). *Dooshi Imiron* (The Semantics of Verbs). Tokyo: Kurosio.
Kuiper, K. and Allan, W. S. (2004). *An Introduction to English Language: Word, Sound and Sentence*, 2nd ed. New York: Palgrave Macmillan.
Lakoff, G. (1970). *Irregularity in Syntax*. New York: Holt, Rinehart and Winston.
Lehnert, M. (1971). *Reverse Dictionary of Present-Day English*. Leipzip: VEB Verlag Enzyklopädie.
Lieber, R. (2010). *Introducing Morphology*. Cambridge: Cambridge University Press.
Lyons, J. (1968). *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.
Marchand, H. (1969). *The Categories and Types of Present-Day English Word-Formation: A Synchronic-Diachronic Approach*, 2nd ed. München: C. H. Beck.
Plag, I. (1999). *Morphological Productivity: Structural Constraints in English Derivation*. Berlin: Mouton de Gruyter.
Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
Randall, J. H. (2010). *Linking: The Geometry of Argument Structure*. Dordrecht: Springer.
Williams, E. (1981). Argument Structure and Morphology. *The Linguistic Review*, 1:81-114.

Syntactic and morphological features after verbs of perception: Bulgarian in Balkan context

Ekaterina Tarpomanova
Faculty of Slavic Studies
Sofia University Saint Kliment Ohridski
katya@slav.uni-sofia.bg

Abstract

The paper analyses the types of constructions that express a subordinate event after a verb of perception in the languages of the Balkan *Sprachbund*. The subordinate clauses that may follow a verb of perception are a result of common historical processes in Bulgarian, Albanian, Romanian and Greek: the substitution of infinitive by subjunctive and the neutralization of modal and declarative conjunctions after verbs of perception. Additionally, in Albanian and Romanian among the non-finite verbal forms gerund may be found after perception verbs. For the analysed syntactic structures in Bulgarian a corpus approach is further applied in order to support the linguistic analysis with quantitative data.

Keywords: perception verbs, syntactic structure, verb tense and aspect, Balkan languages

1. Introduction

Verbs of perception are a group of verbs whose semantics is related to the experience of one of the senses (traditionally recognized as vision, hearing, taste, smell, and touch, but also internal experiences such as feeling). In a 1984 paper Viberg presents a markedness hierarchy of the perception verbs based on a crosslinguistic study covering 50 languages, concluding that the sense modality hierarchy is the following: sight > hearing > other modalities (Viberg, 1984). As a semantic group, predicates of perception have similar argument structure involving an experiencer who receives the sensory information and a stimulus that prompts the sensory feeling. This study discusses the syntactic and morphological properties of the subordinate clauses or non-finite verb forms that follow the perception predicate expressing a second event the experiencer perceives in the languages of the Balkan linguistic area focusing on Bulgarian.

In many Indo-European languages verbs of perception may add either a non-finite verb form or a clause introduced by a subordinating conjunction. In English non-finite verbal forms that may follow a perception verb are bare infinitive and gerund. In this construction the object of the perception verb is obligatory and it is in fact the logical subject of the non-finite form:

(1a) *I heard him/her **sing***.

(1b) *I heard him/her **singing***.

The two non-finite forms differ in the manner they present the event by the viewpoint of the experiencer: the bare infinitive describes the event as a whole, i.e. the experiencer heard somebody singing from the very beginning to the end; the gerund describes the event in its progress, i.e. the experiencer started hearing the song while somebody has already begun singing.

On the other hand, perception verbs may be followed by a subordinate clause with a finite verb form that allows for temporal marking (depending on the tense of the verb in the main clause and the rules of tense agreement), thus situating the event of the subordinate clause to the temporal axis with respect to the event of the main clause:

(2a) *I saw (that) she **came**.*
(2b) *I saw that he **was bleeding**.*
(2c) *I saw that she **has made** a lot of records.*
(2d) *When I saw that he **had died**, I literally cried myself to sleep.*

In the Balkan languages there are several types of subordinate clauses that may follow verbs of perception, and, additionally, among the non-finite verbal forms gerund may occur in Albanian and Romanian. In what follows, the models that may be found after perception verbs will be analysed and illustrated with examples extracted from corpora available online: the Bulgarian National Corpus (BulNC), the Reference Corpus of Contemporary Romanian (CoRoLa), the Albanian National Corpus (ANC) and the Corpus of Modern Greek (CMG). In addition, for the models found in Bulgarian quantitative data obtained by the BulNC will be presented.

2. The corpora

The abovementioned corpora are used as a source of authentic language examples to confirm the occurrence of the different types of constructions after perception verbs in the languages of the Balkan *Sprachbund*. Additionally, the BulNC is used for the corpus-based approach applied for Bulgarian with the aim to find out how the Balkan feature described here is spread in the language which is the focus of this conference.

The Bulgarian National Corpus is developed at the Institute for Bulgarian Language (Bulgarian Academy of Sciences). It consists of a monolingual (Bulgarian) part and 47 parallel corpora containing altogether 5.2 billion words. The Bulgarian part includes about 1.2 billion words in over 240 000 text samples classified by style, domain and genre and supplied with rich metadata. The monolingual annotation consists in tokenization, sentence splitting, POS tagging, lemmatisation and morphological annotation. The BulNC is dynamic and is constantly enriched with new texts (Koeva et al.,2012).

The Reference Corpus of the Modern Romanian Language was launched in December 2017 by the Research Institute for Artificial Intelligence and the Institute of Computer Science at the Romanian Academy. The CoRoLa contains both written and oral parts. The written texts comprise 1 billion+ tokens and are distributed in an unbalanced way in several language styles (legal, administrative, scientific, journalistic, imaginative, memoirs, blogposts), in four domains (arts and culture, nature, society, science). The written texts are automatically sentence-split, tokenized, part-of-speech tagged, and lemmatized (Barbu Mititelu, Tufiș, Irimia, 2018).

The Albanian National Corpus is developed by a team of linguists from Saint Petersburg (Institute for Linguistic Studies of the Russian Academy of Sciences) and Moscow (the School of Linguistics at HSE). It contains two main subcorpora: Corpus of the modern literary Albanian (main corpus) and Corpus of early Albanian texts. The main corpus contains 31.12 million words, distributed into four styles: press (75.2%), fiction (10.3%), nonfiction (13.8%), poetry (0.7%). The corpus is supplied with a morphological annotation (Morozova and Rusakov, 2015).

The Corpus of Modern Greek is created at the Russian Academy of Sciences using the web interface of the Eastern Armenian National Corpus. The corpus comprises 35.7 million tokens. The main text variety is journalism, additionally there are fiction texts, both Greek and translated. The search engine allows for searching by language variety (dimotiki or katharevousa) and by orthography (monotinic or polytonic) (Kisilier and Arhangel'skij 2018).

3. Substitution of infinitive in the Balkan *Sprachbund*

The loss or avoidance of infinitive is one of the main features of the Balkan morphosyntax (Asenova, 2002: 141). Infinitive has been replaced by subjunctive or subjunctive-like constructions. In Romanian and Albanian subjunctive has a weak morphological marking, thus differing from indicative only in 3 p. sg. and pl. in Romanian and in 2 and 3 p. sg. in Albanian, while in Greek due to phonetical reasons subjunctive has coincided with indicative. Bulgarian as a Slavic language originally has no subjunctive. In the conditions of a weak or missing morphological marking, the main subjunctive marker in the Balkan languages is the conjunction Bulg. *да*, Alb. *të*, Gr. *να*, Rom. *să*, which in Greek and Albanian grammar is considered a particle (Asenova, 2002: 150).

The substitution of infinitive by subjunctive constructions created an opposition between two types of subordinate clauses mentioned in the early studies of the similarities between the Balkan languages (cf. Sandfeld, 1930: 175): modal-voluntative introduced by the conjunction Bulg. *да*, Alb. *të*, Gr. *να*, Rom. *să*, and declarative introduced by the conjunctions Bulg. *че*, Alb. *se*, *që*, Gr. *πως*, *ότι*, *που*, Rom. *că*, *dacă*, *de* (cf. Asenova, 2002: 149). In certain circumstances the opposition between the two types of subordinators may be neutralized and this is the case of the clauses following a perception verb in the main clause:

(3a) Bulg. *Видях го/я да идва*.

(3b) Bulg. *Видях го/я, че идва*.

'I saw him/her coming.'

The use of subjunctive constructions after verbs of perception involve some restrictions in tense and aspect. In all Balkan languages only present is allowed in the subordinate clause introduced by the conjunction Bulg. *да*, Alb. *të*, Gr. *να*, Rom. *să*, exept for Albanian, where imperfect is possible too. The use of perfect is allowed after a negative form of perception verbs, but in this case the modal meaning of the conjunction is preserved, that is why it is not taken into consideration in the study. In Bulgarian and Greek, which have the grammatical category of aspect, subjunctive construction is generally one of the contexts that favor the use of perfective, but despite this fact after perception verbs only imperfective is possible. The exclusive use of the imperfective is motivated by the relation between the events in the main and the subordinate clause, the former being a point on the continuous line of the latter (Bakker, 1970: 81).

4. Constructions after verbs of perception in the Balkan languages

Several models of constructions occuring after verbs of perception may be outlined in the Balkan languages, some of them are due to their common diachronic development, others are bilateral similarities or language-specific peculiarities.

4.1. Subjunctive construction

The subjunctive construction, as mentioned previously, is the substitute of infinitive and an important similarity between the Balkan languages, including the context discussed here. It is introduced by the modal subordinator Bulg. *да*, Alb. *të*, Gr. *να*, Rom. *să* 'to', but after verbs of preception the conjunction has lost its modal functions.

(4a) Bulg. *Видях го да се усмихва.* (BulNC) 'I saw him smiling.'

(4b) Rom. ... *dar nici nu l-am văzut să facă nici un compromis mare.* (CoRoLa) '… but I never saw him making any big compromise.'

(4c) Alb. *Më pëlqen shumë kur e dëgjoj të flasë.* (ANC) 'I like it very much when I listen to her talking.'

(4d) Gr. *Την είδα να πετάει...* (CMG) 'I saw her falling…'

As compared to the infinitive, the finite verb forms in the subjunctive construction are additionally marked for person, number and tense, but due to the temporal and aspectual restrictions, they do not bear any rich grammatical information. The infinitive is preserved in Romanian and in the north dialect of Albanian (Gheg), but it cannot be used after verbs of perception, which proves the limitation of its functions.

4.2. Declarative constructions

The declarative constructions after perception verbs are typically fronted by the subordinator Bulg. *че*, Alb. *se*, Gr. *πως*, *ότι*, Rom. *că* 'that'. Aditionally, in Albanian and Greek in this position may occur the so-called universal relative (the term is originally used by Petya Asenova to denote the invariable pronoun or pronominal adverb in the Balkan languages used in colloquial speech instead of inflected relative pronouns, cf. Asenova, 1983) – Alb. *që*, Gr. *που*. As stated above, declarative constructions allow for different tenses in the subordinate clause. The examples below show some of the possibilities after aorist in the main clause – perfect in (5a), present in (5b), imperfect in (5c), and aorist in (5d):

(5a) Bulg. *Видях, че не е помръднал.* (BulNC) 'I saw that he hasn't moved.'

(5b) Rom. *... am văzut că se mișcă o umbră...* (CoRoLa) 'I saw that a shadow is moving…'

(5c) Alb. *Befas pashë se makina po drejtohej nga bulevardi.* (ANC) 'Suddenly I saw that a car was coming from the boulevard.'

(5d) *Είδα ότι τα χρυσαφικά άρχισαν να τελειώνουν.* (CMG) 'I saw that we were running out of jewels.'

Another option for declarative construction after perception verbs in the Balkan languages is a subordinate clause introduced by the pronominal adverb Bulg. *как*, Alb. *(se) si*, Gr. *πώς*, Rom. *cum* 'how'. In some contexts the adverb preserves the semantics of manner, but it may also be subjected to desemantization and used with a generalized sense just to register a fact without necessarily focusing on the manner the event is performed. This double role is demonstrated with the following examples:

(6a) Bulg. *Видях как загина.* (BulNC) 'I saw how he died.'

(6b) Bulg. *Видях как в тях проблесна облекчение.* (BulNC) 'I saw how they calmed down.'

(7a) Alb. *Pashë si u pushkatua vëllai i këngëtares.* (ANC) 'I saw how the singer's brother was killed.'

(7b) Alb. *... dhe unë pashë se si u largua duke marrë me vete shprehjen enigmatike të syve të saj.* (ANC) '… and I saw her walking away, taking with her the enigmatic expression of her eyes.'

(8a) Gr. *Χαίρομαι που είδα πώς γίνεται.* (CMG) 'I'm glad I saw how it may be done.'

(8b) Gr. *Έτρεξα πίσω τους, μα σαν φθασα στην άκρολιμνιά, είδα πώς ήταν τρεις.* (Πηνελόπη Δέλτα, "Τον καιρό του Βουλγαροκτόνου") 'I ran after them, but when I reached the seashore, I saw they were three.'

(9a) Rom. *... am văzut cum se face vinul în Dobrogea.* (CoRoLa) 'I saw the way wine is made in Dobrogea.'

(9b) Rom. *Gata, mă, l-am văzut cum a plecat pe șosea!* (CoRoLa) 'It's done, I saw him leaving on the road!'

In the sentences given above, examples indexed with (a) indicate the manner of realization, while in the ones indexed with (b) the adverb of manner is synonymous with the declarative conjunction. Disambiguation may only be made by the context and in some contexts both readings are possible. For Greek it should be noticed that one of the declarative conjunctions, *πως*, has derived from the pronominal adverb of manner and in the modern language they can be distinguished only in the written variety by the accent put on the adverb (*πώς* vs. *πως*).

4.3. Gerund

Among the non-finite verb forms, only gerund may occur after perception verbs in Albanian and Romanian. Similarly to other languages that allow for non-finite verb forms after perception verbs, direct object in the main clause is obligatory referring to the logical subject of the event expressed by the gerund.

(10a) Alb. *Unë nuk të pashë duke u nisur, sepse ti kishe marrë udhë në orët e vona të natës dhe nuk doje të ma prishje gjumin.* (ANC) 'I didn't see you leaving, because you left late at night and you didn't want to wake me up.'

(10b) Rom. *... am văzut venind spre mine un bătrân…* (CoRoLa) '… I saw an old man coming to me…'

In Bulgarian and Greek the gerund always refers to the subject, therefore if a gerund us used after a perception verb, it refers to its subject and not to its direct object.

(11a) Bulg. *Видях го, тръгвайки за важно интервю.* 'I saw him when I was going to an important interview.'

(11b) Gr. *Τον είδα γυρίζοντας απ'το φροντιστήριο*. 'I saw him when I was coming back from school.'

5. Quantitative data for Bulgarian

In Bulgarian there are three competing constructions following verbs of perception: subordinate clauses fronted by *да* 'to', *че* 'that', and *как* 'how'. In this section, the frequency of their use and some characteristics of their syntactic structure is examined based on data of the Bulgarian National Corpus.

The rate of occurrence of the three subordinators has been surveyed after three basic perception verbs: 'see', 'hear' and 'feel' in 1 p. sg., aorist. The verb form is chosen as it is representative for the studied type of sentences and the results of the corpus search are focused and less noisy. Verbs for taste, smell and touch do not occur in that type of sentences (**I tasted him coming*). The results are presented in Table 1.

Verb / Conjunction	*че* 'that'	*как* 'how'	*да* 'to'
видях 'I saw'	5762	3191	1638
чух 'I heard'	3317	1347	1721
усетих 'I felt'	2413	1173	68

Table 1. Number of occurrences of the three conjunctions after perception verbs

The ratio between the conjunctions that introduce the different types of subordinate clauses after the three verb forms extracted from BulNC is displayed in Figure 1.

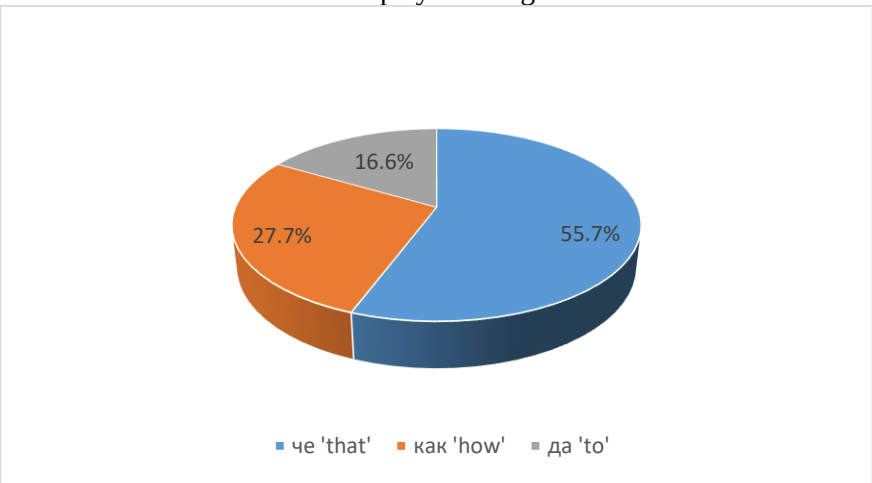


Figure 1. Ratio between the subordinators

The results of the corpus search show a clear preference for subordinate clauses introduced by 'that' in Bulgarian – more than a half as compared with the other two subordinators. The less used are 'to'-clauses – they slightly prevail over 'how'-clauses after the verb for hearing, but they concede considerably in the total result. The restricted use of the 'to'-clauses may be explained by the limited grammatical information that verbs in them may express.

Another issue that may be examined through corpus data is the syntactic structure of the main and the subordinate clause with respect to the argument realization. It is well known that in certain contexts the syntactic position of an argument may remain empty or may be filled by an argument of another predicate (Koeva, 2005: 37). This is the case of the perception verbs in many Indo-European languages whose object is in fact an argument of the predicate in the subordinate clause and its logical subject:

(12a) Eng. *I saw him/her come.*

(12b) Fr. *Je l'ai vu venir.*

In English and French this is the only syntactic structure possible, but in Bulgarian the logical subject of the subordinate predicate in all three constructions may be expressed either as an object of the main predicate (the verb of perception) or as a subject of the subordinate predicate. The argument may remain unexpressed only if the subordinate predicate is impersonal.

(13a) *Видях го/я да идва. / Видях той/тя да идва*.

(13b) *Видях го/я, че идва. / Видях, че той/тя идва.*

(13c) *Видях го/я как идва. / Видях как той/тя идва.*

'I saw him/her come.'

(14) *Видях да/че/как вали.* 'I saw it was raining.'

Nevertheless, the realization of the logical subject of the subordinate predicate as its grammatical subject is more typical for the declarative constructions, while as an object of the main predicate it is most often used in 'to'-constructions. This distribution is visible also by the corpus data presented in Table 2.

Verb / Conjunction	*че* 'that'	*как* 'how'	*да* 'to'
видях го 'I saw him'	226	301	683
чух го 'I heard him'	101	153	740
усетих го 'I felt him'	26	22	20

Table 2. Number of occurrences of the three conjunctions after perception verbs with explicit direct object

The corpus data show that there are no gender-specific differences in the realization of the logical subject of the subordinate predicate as a direct object of the main predicate. The search results with a feminine accusative pronoun in Table 3 display the preference for 'to'-constructions, except for the verb 'feel' whose limited occurrences are not statistically important.

Verb / Conjunction	*че* 'that'	*как* 'how'	*да* 'to'
видях я 'I saw her'	84	152	403
чух я 'I heard her'	33	80	356
усетих я 'I felt her'	15	12	4

Table 3. Number of occurrences of the three conjunctions after perception verbs with explicit direct object in feminine

Provided that the verb in Bulgarian is highly inflected the subject in a clause may be omitted. In sentences with a perception verb, if the argument of the subordinate predicate is realized as its subject, it can be omitted in the declarative 'that'- and 'how'-constructions, but never in the subjunctive 'to'-construction:

(15) *Видях, че идва. / Видях как идва. / *Видях да идва.* 'I saw (somebody) come.'

The fact that the subject cannot be omitted shows that the argument in the subjunctive construction is more naturally interpreted as an object of the main verb.

6. Conclusions and further directions

The study outlines several models of presenting a second (subordinate) event after verbs of perception in the languages of the Balkan *Sprachbund*. A common feature of the Balkan languages is the neutralization of the opposition between modal and declarative subordinators after perception verbs, but despite their semantic equivalence, the respective clauses they introduce differ in terms of possibilities for morphological marking of the verb and have some syntactic peculiarities related to the argument structure. The use of gerund that refers to the object of the main clause is a bilateral similarity between Romanian and Albanian, which is not shared by Bulgarian and Greek. Quantitative data for Bulgarian obtained by corpus search show that the most used model is the subordinate clause introduced by the declarative conjunction 'that' and that the realization of the logical subject of the subordinate predicate as an object of the main predicate is preferred in the 'to'-model. The study may be further enlarged by detecting translation equivalents of the models described here in parallel corpora.

Acknowledgements

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

References

Asenova, P. (1983). A propos des fonctions synatxiques des rélatifs absolus dans les langues balkaniques. In *Die slawische Sprachen 5. Referate des 2. Salzburger Slawistengesprächs "Probleme des Sprachkontakts"*, Teil 2, 5 – 12.

Asenova, P. (2002). *Balkansko ezikoznanie*. Veliko Tarnovo: Faber.

Bakker, W. F. (1970). The aspectual differences between the present and aorist subjunctives in Modern Greek. *Ελληνικά*, 23, 78 – 108.

Barbu Mititelu, V., D. Tufiș, E. Irimia. (2018). The Reference Corpus of Contemporary Romanian Language (CoRoLa). *Proceedings of the 11th Language Resources and Evaluation Conference – LREC'18*, Miyazaki, Japan, European Language Resources Association (ELRA), 1178–1185. Available at http://www.lrec-conf.org/proceedings/lrec2018/index.html.

Kisilier, M. L., T. A. Arhangel'skij. (2018). Korpusa grecheskogo iazyka: dostizheniia, celi i zadachi. *Indoevropeiskoe iazykoznanie i klassicheskaia filologiia*, XXII(1), 2018. P. 50 – 59.

Koeva, S. (2005). Argumenti – semantichni otnosheniya i sintaktichna realizatsiya. In Koeva, S., Ed., *Argumentna struktura. Problemi na prostoto i slozhnoto izrechenie.* Sofia: Semarsh, 25 – 42.

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

Morozova, M., A. Rusakov. (2015). Albanian National Corpus: Composition, Text Processing and Corpus-Oriented Grammar Development. *Sprache und Kultur der Albaner. Zeitliche und räumliche Dimensionen. Akten der 5. Deutsch-albanischen kulturwissenschaftlichen Tagung (5.–8. Juni 2014, Buçimas bei Pogradec, Albanien)* / Hrsg. von B. Demiraj. Wiesbaden: Harrassowitz Verlag, 2015. (Albanische Forschungen, 37), 270 – 308.

Sandfeld, K. (1930). *Linguistique balkanique*. Problèmes et résultats. Paris.

Viberg, Å. (1984). The verbs of perception: a typological study. In Butterworth, B., Ed., *Explanations for language universals.* Berlin: Mouton, 123 – 162.

Corpora

ANC: Maria Morozova, Alexander Rusakov, Timofey Arkhangelskiy. Albanian National Corpus. (Available online at: albanian.web-corpora.net, accessed on 02.06.2020.)

BulNC: Bulgarian National Corpus, available at: http://search.dcl.bas.bg/, accessed on 02.06.2020

CoRoLa: Reference Corpus of Contemporary Romanian, available at: http://corola.racai.ro/, accessed on 02.06.2020

CMG: Corpus of Modern Greek, availab;e at: http://web-corpora.net/GreekCorpus/search/, accessed on 02.06.2020

On the Valency Frames of type Subject-Predicate in Bulgarian

Petya Osenova
Division of Bulgarian Language
Sofia University "St. Kl. Ohridski"
petya@bultreebank.org

Abstract

The paper presents some observations on the semantic constraints of the intransitive subjects with respect to the predicates they combine with. For these observations a valency dictionary of Bulgarian was used. Here two clarifications are to be made. First, the intransitive predicates are viewed in a broader perspective. They combine true intransitives as well as intransitive usages of transitive verbs. The complexity comes from the modeling of these verbs in the morphological dictionary. Second, the semantic constraints that are considered here, are limited to a set of semantic roles and build on the lexicographic classes of verbs in WordNet.

Keywords: intransitive verbs, semantic constraints, subject, lexicographic classes, valency dictionary

1. Introduction

The aim of this paper is to describe some of the syntactic and semantic varieties within the valency frames of type subject-predicate in Bulgarian with the help of a data-driven valency dictionary. The valency dictionary that was used here is the one built over the syntactically annotated corpus BulTreeBank (Simov et al., 2005). Also, some general semantic constraints were available over the grammatical role 'subject'. These semantic constraints include a set of basic semantic roles and general concepts. My aim is to exploit them for the construction of a more formalized and more detailed set of semantic roles in the future.

Let me first briefly introduce the valency dictionary for Bulgarian as described in (Osenova et al., 2012). The data-driven valency lexicon covers the verbs in the syntactically analyzed corpus of Bulgarian — BulTreeBank. It adopts a representation of the surface syntactic structure, and consists of constraints in the form of coarse ontological labels and semantic roles. The process of valency lexicon creation underwent several steps. First, all the verbs were extracted together with the sentences they have been used in. Then they were lemmatized and sorted by the lemma marker. A default valence frame was inserted that presents an example predicate with its core arguments: a subject (SUBJ), a direct object (DIROBJ) and an indirect object (INDOBJ). Since the default valence frame obviously cannot match all the real frames, a manual checking was performed afterwards for the purposes of frame repair and validation.

Here I am interested in frames that have only one grammatical role — Subject. The other roles might by anything but direct object, because it is the well-known marker of transitivity. In principle, the verbs of interest should be the intransitive ones only, i.e. verbs that do not have a direct object. However, since the used valency dictionary followed the surface realizations of the verbal arguments in the corpus, the intransitive verb group is actually wider.

Note that verbs with clausal objects are considered intransitive. The intransitive group includes also transitive verbs that underwent de-tranzitivization under various circumstances (for example, reflexivization, de-causativization, lexical shifts, etc.) and thus can be used as intransitives. In this paper I will not

dwell into the specifics of all these processes that result in intransitive verb usages. I will just mention that different frameworks view these processes in various ways.

It should be noted that the dictionary presentation of verbs, especially the ones with the de-tranzitivizing particles се 'se' and си 'si' as well as the ones with optional arguments, is not trivial. On the one hand, this is due to the fact that grammar and dictionary have complex common interfaces that cannot be fully represented neither in the grammar, nor in the dictionary only. Thus, such a representation needs intermediate levels. On the other hand, there is no ideal way to deal with optionality of the arguments in discourse. Hence, far from trivial is also the relation between the dictionary representation and the text realization of these cases. Not surprisingly, there is vast literature on the specifics of се se- and си si-verbs together with the related phenomena on morphological, syntactic and semantic levels. See (Nitzolova, 2017), (Koeva, 1998), (Petrova, 2014) among others.

The paper is structured as follows: in the next section some more details about the valency dictionary are given. Section 3 outlines my observations on the distribution of certain types of subjects per semantic roles/types of predicates. Section 4 concludes the paper.

2. Valency Lexicon in Brief

The principles behind the valency lexicon are as follows: as mentioned above, the valence frames were kept to the surface syntax. However, the verb usage has been encoded only in active voice[1]. The verbs in perfective and imperfective aspects were encoded as separate lemmas following one of the two linguistic views within the Bulgarian grammar literature. The other one considers them as forms of the same lemma.

The frame includes only the inner participants (semantically obligatory for the event or situation, presented by the predicate, but might be unexpressed on the surface level) (Pustejovsky, 1998). According to Pustejovsky there are three types of arguments:

- true arguments (obligatory for the predicate on the syntactic level like in 'devour a sandwich')
- default arguments (optional on the syntactic level like in the sentence 'I like reading a book' and 'I like reading.')
- shadow arguments (expressed internally in the lexical semantics of the predicate like in 'I kicked the football [with my leg]'). The prepositional phrase 'with my leg' is presupposed by the verb 'kick', so its explicit realisation is possible only if some additional information is added like in 'I kicked the football with my left leg'.

All these argument types can have also intransitive usages. Note that the Bulgarian subject is considered a default argument in this analysis, i.e. it can be omitted on a regular basis but under certain circumstances. Thus, its explicit or implicit realization, although grammatically possible due to the rich verbal inflection, often depends on specific discourse-related conditions.

Based on the statistics from BultreeBank — (Osenova et al., 2012), the type with an explicit nominal subject that is of interest to me 'Subject (NP) - Predicate' comes third by frequency after the types 'Predicate - Direct Object (NP)' and 'Subject (NP) - Predicate - Direct Object (NP)'.

The construction of the valency frames included also the following steps: extracting examples from the treebank for the corresponding verb; classifying the verb with respect to one of the 15 lexicographic classes in WordNet through the BTB-WN (Osenova and Simov, 2018b); making semantic abstractions over the examples with respect to a general ontology and the transferred typical semantic roles based on VerbNet[2]. Note that the semantic abstractions are still very general and that the set of semantic roles is not exhaustive. It includes the following roles that vary across classes: Agent, Patient, Experiencer, Theme, Goal, Locative, Cause. Also, it much be taken into account that the semantic roles were assigned automatically to the verb arguments and then manually fixed. So, the data is still not completely refined.

[1] With the exception of cases where the predicates do not have active voice.

[2] https://verbs.colorado.edu/verbnet/

The frequencies extracted from the valency dictionary are as follows: from 1928 verbs in the valency dictionary, 520 verbs are intransitive by type or by usage which makes approximately one-fourth of the cases. From them 342 are true intransitives (including intransitive usages) and 178 are de-transivised with the reflexive particle се 'se'.

3. Observations

As already mentioned above, in order to get oriented within the predicate types, the lexicographic classes of verbs from the Wordnet were used. These 15 classes are listed below. Their occurrences in BulTreeBank (215 000 tokens) are given in the brackets according to the information reported in (Osenova and Simov, 2018a):

- verb.communication (283)
- verb.social (222)
- verb.stative (219)
- verb.motion (204)
- verb.cognition (203)
- verb.change (184)
- verb.possession (130)
- verb.contact (97)
- verb.creation (95)
- verb.perception (86)
- verb.competition (63)
- verb.emotion (53)
- verb.body (41)
- verb.weather (14)
- verb.consumption (13)

The total number of the annotated classes is 1907.

The initial semantic restrictions on the nominal groups were based on the SIMPLE lexicon ontology[3]. Below a very small part from it is shown in a simplified flat manner.

```
Person
Organization
Animal
Plant
Physical Object
Artefact (social/cognitive)
    Clothing
Event
Activity
Location
```

[3]http://webilc.ilc.cnr.it/clips/Ontology.htm

From the list of labels, observations were made on the following ones only: Person, Animal, Plant, Artefact and Event. It should be noted that at this stage Organization was subsumed by Person and Activity by Event.

The truly intransitive verbs as well as intransitive verb usages, show the following distribution of the respective nominal subject types:

- 234 subjects with the label Person
- 38 subjects with the label Event
- 34 subjects with the label Artefact
- 14 subjects with the label Animal
- 9 subjects with the label Plant

It can be seen that the most frequent type is Person, then almost equally often come Event and Artefact. Finally, with the fewest occurrences are Animal and Plant. Again, it should be taken into account that the corpus is mainly news media and partly literature. This fact influences the distribution of the semantic constraints over subjects. However, apart from the fact that Person subjects prevail over the Event and Artefact ones, this observation is not very informative per se. For that reason I focus on the semantic roles of subjects of intransitive/de-tranzitivized verbs within the most frequent lexicographic classes: verb.communication, verb.social, verb.stative, verb.motion and verb.cognition. I will briefly introduce each group according to (Miller et al., 1990).

3.1. Verb.communication Subjects

Verbs of communication are considered as: "verbs of verbal and nonverbal communication (gesturing); the former are further divided into verbs of speaking and verbs of writing [...] verbs referring to animal noises (neigh, moo, etc.) and verbs of noise production and uttering that have an inanimate source and lack a communicative function (creak, screech)." (p. 58). This class is expectedly the most frequent one in our news media corpus.

From 283 verbs 60 are with intransitive usages. This is around one-fifth of the cases. Here come verbs like броя (count), бягам от (avoid, escape), изпитвам (exam), наричам (name), говоря (speak), договарям се (negotiate), etc. Most of the subjects are AGENTS with a constraint persons. This cluster includes also the role of EFFECTOR and other ones that can cause an event, but are not persons. Rarely there occur other types. For example, animals (the verb вия (howl) with a subject wolves); events (the verb гръмна (disclose) where the subject is a scandal, a secret, etc).

Let us look into some of the typical verbs. For example, the verb говоря (speak) has an intransitive usage in one of its senses, namely: make a speech. A person can speak in front of an organization, audience; for some time; from a certain place. A variant of this verb is the perfective one заговоря (start speaking). However, more frequent is its subjectless impersonal usage in se-passive with an indirect object: В града се заговори за нея 'In town-the se.REFL spoke about her' (In the town they spoke about her).

The verb потека (spread, circulate) has as its subject an artefact (THEME): После потекоха компроматите 'Then leaked compromising-material-the' (Then the compromising material was disclosed).

Figure 1 shows an example from the valency dictionary visualized in XML in the CLaRK System[4].

The screenshot shows the verb гръмна (disclose) with a subject скандал 'scandal'. The notations are as follows: 'FD' stands for a Frame in the Dictionary; 'l' encodes the lemma; 'def' gives the definition; 'F' presents the general semantic constraint over the subject which says 'event discloses'; 'FSRL' encodes the semantic role AGENT; 'en' gives the link to this meaning of the verb in Princeton WordNet; 'senses' outlines the Bulgarian definition; 'tok' provides examples from the treebank.

[4]http://bultreebank.org/en/clark/

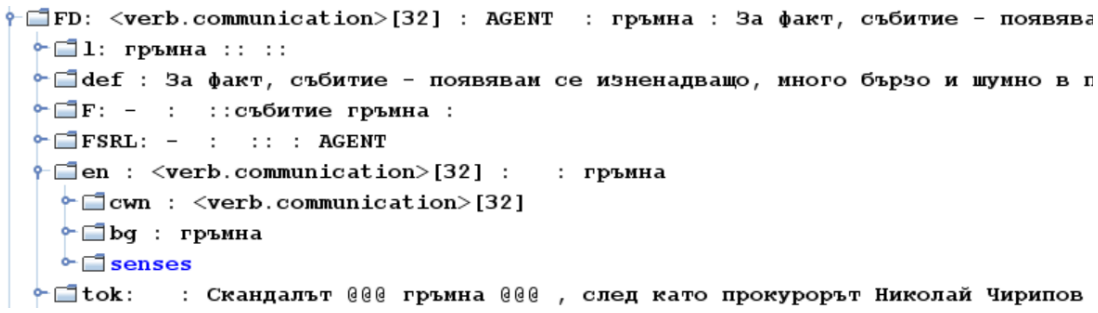


Figure 1: Verb.communication Subjects

3.2. Verb.social Subjects

This group refers to "verbs from different areas of social life: law, politics, economy, education, family, religion, etc. Many have a specialized meaning, restricted to a particular domain of social life, and they tend to be monosemous". (pp. 60–61)

This is the second most frequent type in the corpus. From 222 verbs 40 are with intransitive usages. This also makes approximately one-fifth of the cases.

One of the typical verbs here is действам (act, perform an action). It has three main occurrences: a) as it is: За да действа, човек трябва да говори 'In order to act, person must speak'; b) with a se-particle: Трябва да се действа 'Must to se.REFL act' (One has to act), and c) as an attributive present participle: действаща военна структура, 'acting military structure' (an active military unit). Another typical verb is работя (work). People mostly work at some position, or at some organization, or in some place, for some time, with some device. Here come also verbs as сгреша (sin), служа (serve, do military service), сътруднича (collaborate), etc. The verb справям се (cope, manage) presents either frames without any participants apart from the subject (manage), or with an indirect object (cope with something) and with adjuncts (typically adverbs of manner).

Among the usages there are a number of idioms, such as потъна (fall through, collapse). This holds also for the other verb classes.

In spite of the predominance of the person constraint within the AGENT role there occur also some social verbs whose subject is different. For example, спомагам/спомогна (help). In the following example the subject is an event and the semantic role is not AGENT but a kind of EFFECTOR: Физическите натоварвания ще спомогнат за повишаване на тонуса 'Physical-the exercises will help for increasing of tonus' (Physical exercises will make one fit).

3.3. Verb.stative Subjects

This group includes "for the most part verbs of being and having. Many stative verbs also have non-stative senses that have been placed into other files." (p. 60)

It is the third largest type in the corpus. From 219 verbs approximately one-half exhibits intransitive usages (i.e. around 100). Moreover, in this group the AGENT subject role more often is alternated by the roles PATIENT and THEME.

Concerning the AGENT subjects, person is typical for verbs like гостувам (visit as a guest at some place, organization, event); присъствам (attend), etc.

As for the roles other than AGENT, there is a big variety of semantic constraints, mostly of type THEME. For example, the verb действам (apply, hold) has as its subject some artefact (legal text, legal document, project, contract, etc.): Редът е такъв, откакто действа новият Закон за държавната собственост 'Order-the is such, from-where applies new-the Act for state ownership' (This has been the case since the new Act for the State Property came into force).

Some event can also take the subject role with verbs like бавя се (prolong). For example: Ремонтът на летището се бави 'Renovation-the of airport-the se.REFL late' (The renovation of the airport is delayed). Thus, the semantic role is a THEME. Another example of a THEME role subject is the verb

водя (lead, go) with a subject that is a street, path, road. See: Пътят води към върха 'Road-the leads towards peak-the' (The road leads to the peak). More examples refer to verbs, among which: предстоя (impend), приключвам/приключа (end, stop).

There are cases in which the verb can take as subjects both - AGENT (person, organization, state) and THEME (event, artefact, object, etc.). For example, the verb идвам (follow): На следващо място идва този аргумент 'To next place comes this evidence' (Next comes this argument). Another verb is липсвам (be absent): Липсва добрата алтернатива 'Lacks good-the alternative' (There is a lack of a good alternative). More verbs are: оставам (endure, persist), преобладавам (predominate, loom), принадлежа (belong), служа (serve), etc.

3.4. Verb.motion Subjects

The motion verbs "derive from two roots: move, make a movement, and move, travel". (p. 59)

This is the fourth most frequent group of verbs in the corpus. From 206 verbs 150 are with an intransitive usage. Thus, within this group of typical verbs of moving and acting the intransitives do prevail as expected.

The AGENT role with a person constraint but allowing also other ontological concepts like animal is typical for verbs like бягам (leave, exit), вървя (walk), идвам/дойда (arrive). The generalized AGENT role can combine various constraints: persons/vehicles (plane)/celestial bodies like обикалям (circle); person/artefact/vehicle like потъвам/потъна (sink); person/vehicle/bird like пътувам (travel) or person/event/activity like стигам/стигна (reach): Докъде стигна работата по случая? 'To where reached work-the on case-the?' (What is the status of the work on this case?). Such cases have to be refined with respect to the specific semantic roles. Here come also verbs with restricted subjects other than person AGENT like бия (heart beats), изминавам/измина (time elapses).

3.5. Verb.cognition Subjects

This group includes "verbs denoting various cognitive actions and states, such as reasoning, judging, learning, memorizing, understanding, and concluding". (p. 59)

This is the fifth most frequent group in the corpus. From 203 verbs only 50 are with an intransitive usage which makes one-fourth of the cases. Here the subject roles are labeled exclusively EXPERIENCER. A typical EXPERINCER person subject belongs to verbs like: знам (know, cognize), надниквам/надникна в нещо (get through, sink in), научавам/науча за нещо (learn, hear), мисля (think, judge): Тя го мисли за глупав човек 'She thinks him.ACC for stupid person' (She thinks that he is a fool).

The combination of EXPERIENCER subjects that are persons with oblique participants possessing a GOAL role are verbs like отстъпвам/отстъпя от позиция (abondon, give up): гледам на нещо по някакъв начин (consider): Политиците гледат практично на нещата 'Politicians look practically on things-the' (Politicians view everything from a practical point of view).

4. Conclusions

The paper presents some observations on the combination of certain semantic types/roles of subjects in 5 lexicographic classes with intransitive predicates.

Within these most frequent types the verb.communication and verb.social exhibit predominantly AGENT subjects with a person constraint.

Verb.stative type increases the intransitive frames and also the PATIENT/THEME subject roles. Verb.motion keeps the AGENT subjects as majority similarly to verb.communication and verb.social, but like verb.stative it has prevailing numbers of intransitive frames. The only type among the five most frequent ones in the corpus – verb.cognition – imposes the EXPERIENCER subject role within the group of not so many intransitive cases.

Depending on the verb meaning, the frame can have a more specific or a more general set of semantic constraints/roles. Since the valence dictionary presentation of frames is data-driven, it requires more

work on the proper mappings among the lexical meanings, verb valencies and semantic labels of the verb arguments.

Acknowledgements

This work was partially supported by the *Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant number DO01-272/16.12.2019.

References

Koeva, S. (1998). Reflexive, passive, optative, reciprocal and impersonal verbs in Bulgarian. In *Nauchni trudove na Plovdivskiya universitet – Filologiya, 36, 1*, pages 142–157.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. 3:235–244.

Nitzolova, R. (2017). *Bulgarian grammar*. Berlin: Frank and Timme GmbH.

Osenova, P. and Simov, K. (2018a). Enriching valency frames lexicon of Bulgarian with Semantic Roles. In *Slovanská lexikografie počátkem 21. století. Sborník z konference Praha 20. – 22. 4. 201*, pages 239–246.

Osenova, P. and Simov, K. (2018b). The datadriven Bulgarian WordNet: BTBWN . In *Cognitive Studies, Études cognitives, 18*.

Osenova, P., Simov, K., Laskova, L., and Kancheva, S. (2012). A Treebank-driven Creation of an OntoValence Verb lexicon for Bulgarian. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 2636–2640.

Petrova, G. (2014). Medialni glagoli s refleksivna semantika. In *Nauchni trudove na Rusenskiya universitet – volume 53, series 6.3*, pages 36–40.

Pustejovsky, J. (1998). *The Generative Lexicon*. The MIT Press.

Simov, K., Osenova, P., Simov, A., and Kouylekov, M. (2005). Design and Implementation of the Bulgarian HPSG-based Treebank. *Journal of Research on Language and Computation, Special Issue*, pages 495–522.

Analysis of Similes in Serbian Literary Texts (1860-1920) using computational methods

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

Jelena Jaćimović
University of Belgrade
School of Dental Medicine
jelena.jacimovic@stomf.bg.ac.rs

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

Abstract

Similes are rhetorical figures which play an important role in literary texts. This paper presents a finite-state methodology developed for the description of adjectival similes, which enables their retrieval and annotation in Serbian novels written in the mid-19th and early 20th centuries. The results of a textometric analysis reveal the most frequent adjectival similes and the specificity of their usage, with respect to the author, title, or publication date, in a subset of the SrpELTeC corpus.

Keywords: rhetorical figures, literary corpus, simile figure, multi-word units

1. Introduction

In the history of rhetoric simile holds a particular place. Although it is one of the oldest recognized figures of speech, from the very beginning, simile has often been taught and studied in conjunction with metaphor. Ever since ancient times, many researchers have been reconsidering the status of simile and its convergence with other familiar figures, treating it either as a literal comparison, a weaker form of metaphoric expression or as a completely distinct figure of speech. Indeed, simile is essentially a rhetorical figure presented, unlike metaphor, as an explicit form of comparison. On the other hand, in contrast to literal comparison, simile is also essentially figurative, making unexpected connections between literally unlike concepts (Israel et al., 2004).

Similes rely on comparisons, semantic figures which bring two different entities together based on a shared feature (Israel et al., 2004), implying a certain likeness between them. Both literal comparison and simile have the same recognizable formal structure, the surface form consisting of the following elements: the subject of comparison (**tenor**, **target**, or **topic**), the object of comparison (**vehicle** or **source**), a conjunction which signals a comparison (**marker**, usually *as ... as*, *as* or *like* in English, *kao* in Serbian, or *comme* in French), and the basis of the comparison implied by the expression (**ground**, **property**, or **tertium comparationis**) (Example 1.1).

Example 1.1.	[She]	was	[free]	[as]	[a bird].
	tenor		**ground**	**marker**	**vehicle**

However, the subject of comparison (**tenor**) most often does not form part of a simile (Brehmer, 2009). Therefore, similes are multi-word expressions (MWE) that, as introduced in (Beardsley, 1981), can either be *closed* (represented with a three-part structure **ground + marker + vehicle**, as given in Example 1.1) or *open* if the shared attribute is not explicitly stated, but could be derived from the context (**marker + vehicle**), as illustrated in Example 1.2 where the shared property of being free is left implicit.

Example 1.2.	[She]	was	[as]	[a bird].
	tenor		**marker**	**vehicle**

In addition to the aforementioned multi-word structure, another formal characteristic of similes is that they are often quite conventionalized, generally known and accepted phrases used by all members of a linguistic community. Even though their lexical composition is highly stable, consisting of at least two or three components, it is not absolute having numerous variants of the essential simile elements. As far as their semantic features are concerned, similes are characterized as being idiomatic and remarkably expressive, which is the result of a powerful connotation, for instance, positive or negative sentiment toward something, and the picturesqueness of their essential parts.

The widespread presence of similes in everyday language stands to reason since they rely on comparing, a fundamental human cognitive activity, producing a particular image in a person's mind (Mpouli, 2016). In view of their evocative power and descriptive capacity, similes are the most attractive comparative structures to investigate in literary texts. As rhetorical instruments, they can easily be combined with other figures of speech (Israel et al., 2004) and used for stylistic effects. As a part of an author's imagery, the similes used can uncover and define the personality and experiences of the author, the tonality of a particular text, or even a literary period. Hence, identifying all simile varieties in a novel appears vital for stylistic examination.

Similes have a structure that appears fairly amenable to automated processing (Niculae and Yaneva, 2013). Still, in computational linguistics, which is particularly interested in figurative language, similes have been overlooked in favor of metaphor even more than in linguistics. Simile analysis has become a particularly appealing topic of interest in the field of computational linguistics and corpus studies in recent years (Niculae, 2013; Yoshimura et al., 2015; Qadir et al., 2015; Qadir et al., 2016; Hu et al., 2017). One of the main tasks is automatic simile recognition, which can be divided into partial and full simile identification. Partial simile identification principally involves retrieving specific simile patterns, either complete expressions consisting of all simile elements, or only preselected grounds and vehicles. Furthermore, this process depends on heuristics or human reasoning to recognize the difference between similes and literal comparisons. On the other hand, full simile recognition involves extraction and analysis of all sentences containing a simile marker in unstructured texts, and subsequent identification of the separate components of each potential simile. For the simile recognition task, various methods have been proposed. Most of them can be classified as feature-based (Niculae and Danescu-Niculescu-Mizil, 2014), pattern-based (Niculae and Yaneva, 2013; Niculae, 2013), or neural network-based (Liu et al., 2018; Zhang et al., 2019).

The research related to rhetorical figures and their automatic processing for the Serbian language started with building the Ontology of Rhetorical Figures (Mladenović and Mitrović, 2013). A method of automatic recognition and classification of rhetorical figures, including similes, that uses ontological inference rules in an ontology based on Serbian WordNet (SWN), was also developed (Mladenović, 2016). In (Mitrović et al., 2019), the authors applied a corpus-driven crowdsourcing method for enrichment of lexical resources with Serbian and Greek similes. A corpus of similes used in modern Serbian language was produced based on a methodology for semi-automated collection of similes from the World Wide Web using text mining techniques (Milošević and Nenadić, 2016; Milošević and Nenadić, 2018).

The main goal of this paper is to provide an analysis of adjectival similes in Serbian novels written in the mid-19th and early 20th centuries, retrieved through automatic recognition and annotation. Moreover, it aims to identify the most frequent similes and their components (such as grounds and vehicles), using the textometric method for analysis and visual presentation of results.

2. About the Corpus

One of the main objectives of the *Distant Reading for European Literary History* (COST Action CA16204) project[1] is compilation of a multilingual European Literary Text Collection (ELTeC). This work is still in progress, but before it ends, the project is expected to comprise around 2,500 full-text novels in at least 10 different languages. All texts from this corpus have to fulfill the same criteria: they should be originally written in a language of the subcollection to which they belong, their first publication date should fall between 1840-1920 (preferably appearing as a book and not published in installments) and

[1]Distant reading `https://www.distant-reading.net/`

they should be at least 10,000 word tokens long. Each language subcollection will eventually comprise up to 100 novels that fulfill certain balancing criteria:[2]

- each of the four twenty-year periods should be represented by approximately the same number of novels;
- at least 10%-50% of the works featured should be written by female authors;
- 9 to 11 authors should be represented by exactly three novels (other authors should be represented by one novel);
- at least 20% should be short novels (10-50k word tokens), at least 20% should be long novels (>100k word tokens);
- at least 30% should be highly popular novels and at least 30% should be novels that are not known to the general public.

For this research 41 novels that are candidates for the Serbian subcollection of the ELTeC corpus were used.[3] The characteristics of the sample corpus having a size of 1,471,141 word tokens are represented in Table 1. It becomes apparent that the Serbian subcollection will not be able to meet all the balancing criteria: presently, there are neither novels from the 1840–1859 time period, nor those that exceed 100,000 word tokens.

Period	**Number**	**Length**	**Number**	**Sex**	**Number**
1840-1859	0	short	30	Male	34
1860-1879	3	medium	11	Female	7
1880-1899	16	long	0		
1900-1920	22				

Table 1: Corpus distribution

This particular collection contains novels of exceptional value for the history of Serbian literature. Besides well-known novels, which introduce a modern narrative structure, this corpus contains novels by forgotten authors, like Dragomir Šišković and Stevan Mamuzić, as well. Moreover, the first Serbian science-fiction novel *Jedna ugašena zvezda* (*An Extinguished Star*) by Lazar Komarčić is part of the srpELTeC corpus too, and so is the novel *Babadevojka* (*Old Maid*) by Draga Gavrilović, the first female author who wrote a novel in the Serbian patriarchal society of the time. A complete list of the novels used in this research can be found in Appendix A. The dimensions of the srpELTeC corpus used on this occasion and partitioned based on authorship are presented in Figure 1.

3. Simile retrieval and annotation

The first step of our research consisted of an attempt to retrieve as many similes as possible from our corpus. Two approaches have been adopted for this purpose: first, we looked for simile figures in the electronic morphological dictionary of Serbian (SMD), and then we applied a simple regular pattern to spot simile occurrences. In both cases, we used the Unitex system and the incorporated SMD (Krstev, 2008).[4]

At present, SMD contains 68 multi-word expressions that represent similes. In our corpus, we retrieved 98 occurrences of these already identified simile figures, or 33 different forms among which *bled kao smrt* 'pale as death' (n=14) was the most frequent one. Based on a regular expression `<A> (<jesam.V>+<E>) (kao+ko+k('+')o)` – an adjective followed by the conjunction *kao*, or some

[2]Encoding Guidelines for the ELTeC: level 1 `https://distantreading.github.io/Schema/eltec-1.html`

[3]The Serbian subcollection is still under construction, and some of the prepared novels might not become part of the final collection due to the balancing criteria that have to be met `https://distantreading.github.io/ELTeC/`.

[4]Unitex/Gramlab, the multilingual corpus processing suite `https://unitexgramlab.org/`.

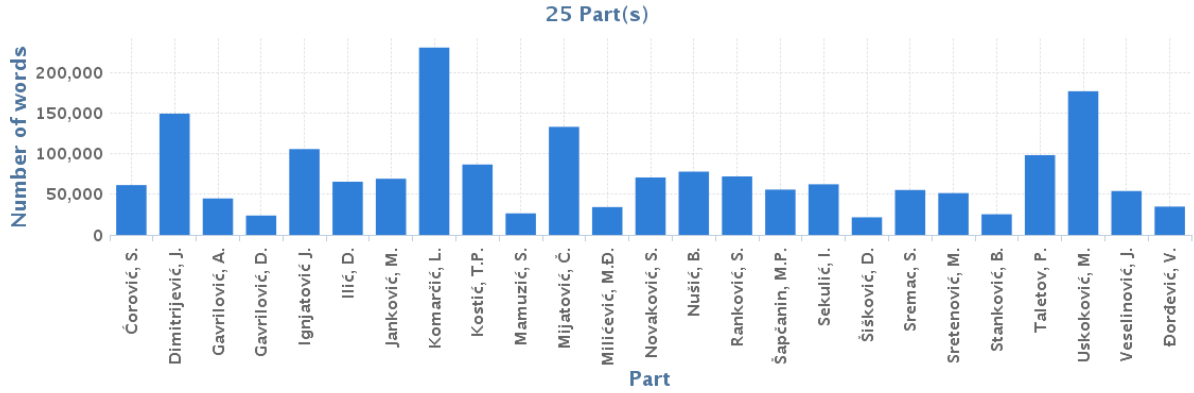


Figure 1: Dimensions of the srpELTeC corpus parts created based on authorship

of its irregular variants, with a possible auxiliary *jesam* 'to be' in between, we obtained a list of possibilities from which we extracted 267 simile occurrences, or 225 distinct forms where *žut kao vosak* 'yellow as wax' (n=5) was the most frequent case. In 4 novels out of 41 in the corpus, no simile was retrieved (all of them were "short" novels).

As mentioned above, some similes have already been recorded in the Serbian morphological dictionary of MWUs (Krstev et al., 2013). This format is consistent with the morphological dictionary of simple words and it allows a description of the various properties of an MWU, besides its morphological features. In the case of simile figures, a dictionary description can specify:

- morphological behavior of an adjective – it does not inflect in degree, it is always used in the positive form;
- morphological behavior of a noun (vehicle) – whether it changes in number to agree with a noun (tenor) or not;
- the order of constituents which can be *A kao N* or *kao N A*.

The example 3.1 illustrates this by way of the entry *gladan kao vuk* 'hungry as a wolf'. DELAC entry is used to produce all inflected and variant forms of an MWU (Savary, 2009).

Example 3.1.

```
DELAC:  gladan(gladan.A18:akms1g) kao vuk(vuk.N128:ms1v)
DELACF: gladnog kao vuk,gladan kao vuk.A:adms4v
        gladnome kao vuk,gladan kao vuk.A:adms7g
        gladni kao vuk,gladan kao vuk.A:aemp1g
        gladni kao vuci,gladan kao vuk.A:aemp1g
        gladni kao vukovi,gladan kao vuk.A:aemp1g
        kao vuk gladnog,gladan kao vuk.A:adms4v
        kao vukovi gladni,gladan kao vuk.A:aemp1g
        ...
```

However, with this representation, a number of deviations occurring in the use of similes cannot be described, such as:

1. variations that may occur in all constituents of similes:
   (a) variation in the ground: for instance, three different forms (near synonyms) *mek*/*mehak*/ *mekan* in the figure *mek kao pero* 'soft as a feather';
   (b) variation in the vehicle: for instance, three different forms (near synonyms) *pero*/*perce*/ *paperje* in the figure *mek kao pero* 'soft as a feather';

(c) variation in the marker: the conjunction *kao* can also be written as *k'o*, *ko*, *ka'*, etc. Although only the first form is sanctioned by the Serbian orthography, other forms occur frequently in literary texts.

2. the vehicle can be modified:
   (a) with an adjective, for instance, *crven kao rak* and *crven kao pečen rak* 'red as a (fried) crayfish'. One should note that the choice of an adjective is not free, in this case it can be just *pečen/kuvan* 'fried/cooked';
   (b) with an adjunct, for instance, *slobodan kao ptica*, *slobodan kao ptica na grani*, *slobodan kao ptica u gori* 'free as a bird (on a bough/in a wood)'. The possible adjuncts are also limited;
   (c) with a determiner (adjectival pronoun), for instance, *mudar kao kakav pop* 'wise as some priest'.

3. variations due to the free word order:
   (a) insertion of an auxiliary, for instance, *crvena si kao zreli nar* 'you are red as a ripe pomegranate';
   (b) insertion of a subject (tenor), for instance, *vreo dah kao plamen* 'breath hot as a flame';
   (c) insertion of a pronoun (clitic), for instance, *privržen mu kao pašče* 'attached to him as a dog';

4. variations resulting from rephrasing, for instance, *žut kao što je slama* 'yellow as straw is'.

The presented variations can be described by local grammars in the form of finite-state automata. One such automaton that recognizes the figure *beo/bjel kao sneg/snijeg [u planini]* 'white as snow [in the mountain]' is presented in Table 2a).[5] The production of such graphs for each individual figure would be impractical. For that reason, generic automata were constructed (Table 2b) in which the information in certain nodes is filled with specific information stored in the table describing all similes (Table 2c). For instance, the upper left node containing `%A,B,C%` in the generic automaton, is replaced by the content of the cells A, B, and C from the simile table to obtain, for the first table line, the corresponding node in the specific automaton: `<beo.A:a>+<bjel.A:a>`, meaning adjective *beo* or *bjel* in the positive form.[6]

The upper path in the generic graph recognizes similes with a regular word order, while the lower part recognizes figures with a reverse order. Variations in the ground and the vehicle (see items 1(a) and 1(b) from the list above) are figure specific and the information filling the appropriate graph node is obtained from the table, while marker variations (item 1(c) are common to all figures and they are coded in the graph. Most optional additions are also figure specific (items 2(a) and 2(b)) and for them, the information is transferred from the table; some are common (item 2(c)), including word order and other insertions (items 3 and 4) and they are coded in the generic graph. In this way, 243 graphs are produced for simple similes (one adjective) and 44 for complex figures (two adjectives).

Even though most researchers tend to mark only phrases representing similes (Mpouli, 2017), we have decided to annotate all recognized similes both at phrase and word levels. Each identified simile has been enclosed within the tag `<simile>`, specifying the type of the rhetorical figure in question and its range. Furthermore, each simile basic element, namely ground, marker, and vehicle, has also been marked with the corresponding tag. Example 3.2 illustrates the annotation of one simile. The annotation process at the moment includes neither annotation of entire sentences containing similes nor additional information regarding simile semantic features.

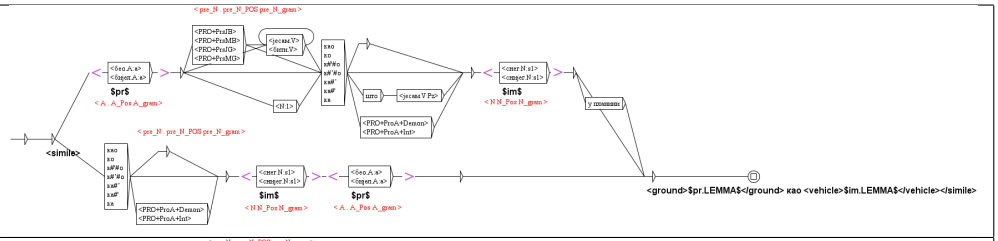
Example 3.2. *zdrav i rumen kao jabuka* 'healthy and ruddy as an apple'

```
<simile><ground>zdrav</ground> i <ground>rumen</ground>
<marker>kao</marker> <vehicle>jabuka</vehicle></simile>
```

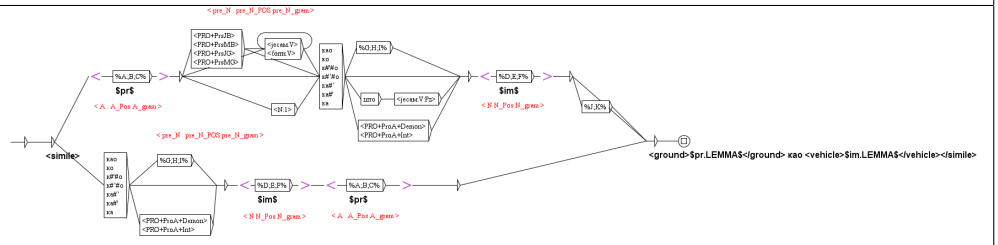
[5] These graphs are implemented in Unitex/Gramlab and they use SMD information implemented in the same environment.

[6] All examples in this paper are given in the Latin script. Most of the novels were published in the Cyrillic script, with just a few in the Latin script. Specific automata were constructed automatically for both scripts.

a



b



c

A	B	C	D	E	F	G	H	I	J	K
beo,bjel	A	:a	sneg,snijeg	N	:s1				u planini	X
blažen	A	:a	dete,djete	N	:s1	mali,malen				
crn	A	:a	gavran	N	:s	zlokoban				
lep,lijep	A	:a	upisan	A	:as1					

Table 2: a) specific finite-state automata; b) generic finite-state automata; c) data describing specific similes.

Finally, the annotated corpus was imported into the TXM program environment (Heiden et al., 2010; Heiden, 2010) for a quantitative and qualitative analysis of the recognized similes. Based on the total number of simile occurrences in the whole corpus (F), the total number of simile occurrences in the texts in a particular part of the corpus (f) and the *specificity score* (S), significantly common or significantly rare occurrences of adjectival similes in distinct parts compared to the whole corpus were identified, as well as the specific use cases of simile adjectival and nominal elements.

4. Analysis of results

The results of the study show that in the corpus consisting of 41 Serbian novels written between 1860 and 1920, 404 occurrences of 251 distinct adjectival similes are found. Among them there are 392 similes with one adjectival ground and 12 represent examples with two adjectives used as simile ground (as shown in Example 3.2). This figure of speech appears most frequently in the texts penned by Uskoković M. (f=58), Komarčić L. (f=56), Dimitrijević J. (f=54), and Taletov P. (f=36). Nevertheless, the results reveal that adjectival similes are extremely specific to the part of the corpus written by Sretenović M. (S=4.2), despite lower absolute frequency (f=26) compared to the previously mentioned authors. The specificity distribution of the recognized adjectival similes in the corpus partitioned based on authorship is presented in Figure 2. On the other hand, if the total number of simile occurrences in the whole corpus is taken into account, for the part dedicated to the works by Komarčić L, where a high simile frequency is recorded, the observed *specificity score* is 0.7, which indicates common rather than specific uses of adjectival similes. Adjectival similes are less represented lexical units in the part of the corpus written by Ignjatović J. (f=7, S=-4). These figures are also significantly rare in corpus parts authored by Gavrilović A. (f=1), Sremac S. (f=3) and Milićević M.Đ. (f=1), with the *specificity score* of -3.2, -2.6 and -2.3, respectively. The novels characterized by a significantly high frequency of use of adjectival similes are *Radetića Mara* (*Mara of the Radetic's*) (S=4.2), *Jedna ugašena zvezda* (*An Extinguished Star*) (S=4.1), *Došljaci* (*Newcomers*) (S=3.2), *Novac* (*Money*) (S=2.9) and *Nove* (*New Women*) (S=2.6). Moreover, if we look at the adjectival simile use in the corpus partitioned by decades, significantly common use of similes occurs in the novels published in the first decade of the 20th century (S=3.8).

The syntactic pattern Adjective + Conjunction + Noun is the most frequent (86.9%) of all adjectival simile occurrences in the corpus (F=404). We have also recorded other syntactic variants and examples

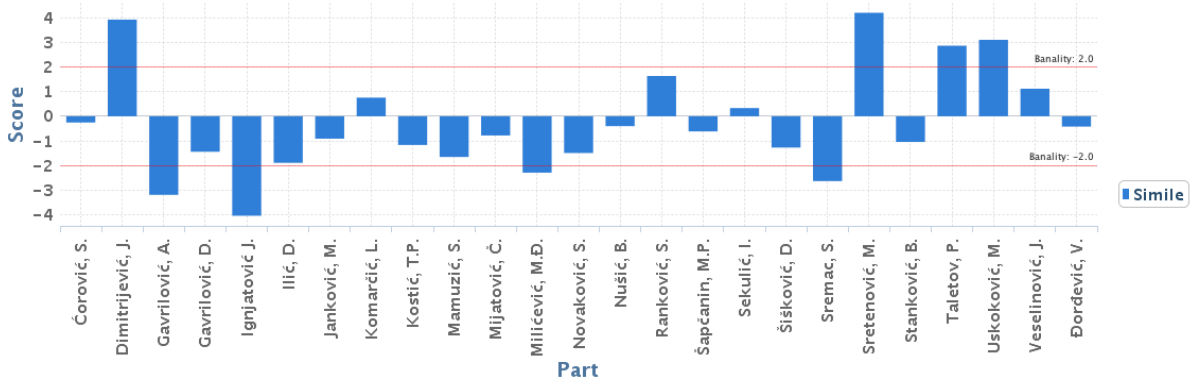


Figure 2: The specificity of adjectival simile use in the srpELTeC corpus by authors

where nominal (10.6%) and adjectival phrases (2.5%) are used instead of a noun as a vehicle of the recognized simile. Besides closed similes, we have also found cases of open simile syntactic patterns such as Conjunction + Noun. In these situations, the connection between the adjective and the noun is very strong making it possible to omit the adjective and still retain the meaning, for instance, *Arhimandrit beše (ljut) kao ris* 'Archimandrite was (angry) like a lynx' where the adjective in the parentheses is omitted. An open simile occurred in two more cases: *(mali) kao makovo zrno* '(small) as a poppy seed' and *(vredan) kao pčela* '(hard working) as a bee'.

The most frequently used adjectival simile in the sample corpus of Serbian novels is *beo kao sneg* 'white as snow' (F=28), followed by the simile *bled kao smrt* 'pale as death' (F=15), which also turns out to be the most frequent simile in the British and French corpora consisting of novels written between the mid-19th and early 20th century (Mpouli and Ganascia, 2015). The adjectives *beo* 'white' and its synonym *bled* 'pale' often appear in the most frequent similes in Serbian novels, and they occur in British and French literary texts from a similar period as well.

There are 416 occurrences of 111 distinct adjectives used as simile ground in this corpus. Adjectives occurring in the retrieved simile figures are frequently connected to different nouns and vice versa. The adjectives and nouns that show the widest variety in connections are represented in Table 3. In some cases, two adjectives are explained by the same noun as in *dobar i miran kao jagnje* 'good and quiet as lamb'. There were 12 such cases.

***beo* 'white' (45)**	***bled* 'pale' (33)**	***stena* 'rock' (10)**	***jagnje* 'lamb' (9)**
sneg (28) 'snow'	*smrt* (15) 'death'	*hladan* (6) 'cold'	*miran* (3) 'quiet'
mleko (7) 'milk'	*krpa* (9) 'cloth'	*nem* (1) 'mute'	*dobar* (2) 'good'
krin/ljiljan (3) 'lily'	*vosak* (4) 'vax'	*neosetljiv* (1) 'impassive'	*blag* (1) 'mild'
alabaster (1)	*senka* (2) 'shadow'	*nepomičan* (1) 'motionless'	*poslušan* (1) 'docile'
list hartije (1) 'sheet (of paper)'	*mrtvački* (1) 'deathly'	*silan* (1) 'strong'	*smiren* (1) 'serene'
sir (1) 'cheese'	*sveća* (1) 'candle'		*nevin* (1) 'innocent'
ovca (1) 'sheep'	*zemlja* (1) 'ground'		
ruža (1) 'rose'			
šećer (1) 'sugar'			
srebro (1) 'silver'			

Table 3: The most frequent adjectives and nouns and their connections

Besides, the same adjective can be used in two different similes, either to describe a physical characteristic (*čist kao sneg* 'clean as snow') or a person's character trait (*čist kao suza* 'pure as a tear'). Moreover, one and the same simile can be used in both senses: *mek kao pamuk* 'soft as cotton'.

A simile can undergo numerous variations. An adjective or a noun can be used in either Ekavian or

Iekavian form,[7] such as, for instance, Ekavian variant *beo kao sir* vs. Iekavian variant *bijel kao sir* 'white as cheese'. Some other variations can be observed as well: the use of diminutive forms of nouns (*lak kao pero* vs. *lak kao perce* 'light as feather') or collective nouns for plural forms (*nevin kao jagnje* vs. *nevini kao jagnjad* 'innocent as a lamb/innocent as lambs). In some cases, a non-literary form of either adjectives or nouns is used: *hladan kao led* vs. *ladan kao led* 'cold as ice' and *slobodan kao ptica* vs. *slobodan kao tica* 'free as a bird'. Finally, (near) synonyms are used as well: *oštar kao zmija* vs. *oštar kao guja* 'sharp as a snake' and *velik kao jaje* vs. *golem kao jaje* 'big as an egg'.

The similes retrieved from the srpELTeC corpus can be classified into the following groups based on the ground:

- figures referring to physical characteristics of objects or people (F=184): *zdrav kao jabuka* 'healthy as an apple', *čist kao sneg* 'clean as snow', *okrugao kao pun mesec* 'round as the full moon';
- figures referring to colors (F=133): *crn kao gavran* 'black as a raven', *crven kao krv* 'red as blood', *plav kao more* 'blue as the sea';
- figures used for describing a person's character or abilities (F=69): *ljut kao ris* 'angry as a lynx', *pljašljiv kao srna* 'timid as a roe deer' *čist kao suza* 'pure as a tear';
- figures representing tastes (F=2): sladak kao šećer 'sweet as sugar';
- other figures (F=28): *skup kao šafran* 'expensive as saffron', slobodan kao ptica 'free as a bird'.

Among the most commonly used adjectival grounds in similes are lexemes denoting color concepts, which can designate not only the color of an object, but also someone's emotional, mental, or physical state. Such frequency of use is expected since colors evoke the vividness of visual images, having a wide range of connotative meanings culturally associated with them (Mpouli, 2016; Filipović Kovačević, 2019). With respect to the whole period covered by the corpus based on the *specificity score* values, novels published in the 20$^{\text{th}}$ century are distinguished by a significantly positive use of colors as adjectival grounds, especially yellow (S=1.4), white (S=1.3), red (S=1), grey (S=0.8), black (S=0.6) and blue (S=0.4).

The nouns most commonly used for the description of physical characteristics of objects or people are *smrt* 'death' (*bled kao smrt* 'pale as death'), *led* 'ice' (*hladan kao led* 'cold as ice'), and *krpa* 'cloth' (*bled kao krpa* 'pale as cloth'), as well as an adjective *upisan* 'inscribed' (*lep kao upisan* 'beautiful as inscribed' ('pretty as a picture')), while a person's character or abilities are most frequently compared to the nouns *jagnje* 'lamb' (*nevin kao jagnje* 'innocent as a lamb'), *stena* 'rock' (*hladan kao stena* 'cold as a rock'), *anđeo* 'angel' (*čist kao anđeo* 'pure as an angel') or *devojka* 'girl' (*stidan kao devojka* 'bashful as a girl'). The nouns that name animals, used as vehicles in adjectival similes, are especially interesting because of their expressiveness, connotations and picturesqueness. The animals that are the most frequently featured in similes are *jagnje* 'lamb', *ovca* 'sheep', *srna* 'roe deer', or *detlić* 'woodpecker'.

5. Conclusion

This paper presents the use of the current version of the SrpELTeC corpus, consisting of Serbian prose works published between 1860 and 1920, in order to retrieve and annotate the instances of rhetoric figures, namely, similes and analyze their usage. As a result, we developed a method for the description of these figures, based on finite-state transducers that makes their retrieval and annotation in Serbian texts possible. The annotated texts were used to study their specific use with respect to the author, title, or publication date. In the future, we will collect other types of simile figures, for instance, those that use prepositional phrases instead of nouns, e.g. *težak kao od olova* 'heavy as if it were made out of lead', as well as verbal similes, e.g. *rikati kao vo* 'roar like a bull'. Besides, we plan to enrich the current annotation scheme with the attributes indicating semantic characteristics of the recognized similes. Our ultimate goal is to publish a database of simile figures used in Serbian novels written between 1860 and 1920.

[7] Two different pronunciations in Serbian.

Acknowledgements

This research was made possible through the support of COST Action CA 16204 *Distant Reading for European Literary History*. We would like to thank numerous volunteers from the Society for Language Resources and Technologies *Jerteh*[8] who helped the production of the SrpELTeC corpus by correcting and annotating the novels.

6. Appendix A. List of the novels from the srpELTeC corpus used in the research

Author	Title	Publication Year
Ćorović, Svetozar	Ženidba Pere Karantana	1905
	Brđani	1919
Dimitrijević, Jelena	Fati-Sultan	1907
	Nove	1912
Đorđević, Vladan	U front	1913
Gavrilović, Andra	Prve žrtve	1893
Gavrilović, Draga	Babadevojka	1887
Ignjatović, Jakov	Jedna ženidba	1862
	Vasa Rešpekt	1875
	Pojeta i advokat	1882
Ilić, Dragutin	Hadži Đera	1904
Janković, Milica	Pre sreće	1918
	Kaluđer iz Rusije	1919
	Neznani junaci	1919
Komarčić, Lazar	Dragocena ogrlica	1880
	Moj kočijaš	1887
	Jedna ugašena zvezda	1902
	Prosioci	1905
Kostić, Tadija	Gospoda seljaci	1896
	Prvo veselje	1903
Mamuzić, Stevan	Nejednaka braća	1896
Mijatović, Čedomilj	Ikonija, vezirova majka	1891
	Rajko od Rasine	1892
	Knez Gradoje od Orlova grada	1899
Milićević, Milan	Jurumusa i Fatima	1879
	Deset para	1881
Novaković, Stojan	Kaluđer i hajduk	1913
Nušić, Branislav	Opštinsko dete	1902
Popović Šapčanin, Milorad	Sanjalo	1888
Ranković, Svetolik	Porušeni ideali	1900
Sekulić, Isidora	Đakon Bogorodičine crkve	1919
Šišković, Dragomir	Jedan od mnogih - roman iz prestoničkog života	1920
Sremac, Stevan	Ivkova slava	1895
Sretenović, Mihailo	Radetića Mara – pripovetka iz seoskog života	1894
Stanković, Borisav	Uvela ruža	1899
	Pokojnikova žena	1902
Taletov, Pera	Novac - roman iz beogradskog života	1906
Uskoković, Milutin	Došljaci	1910
	Potrošene reči	1911
	Čedomir Ilić	1914
Veselinović, Janko	Seljanka	1893

References

Beardsley, M. C. (1981). *Aesthetics: Problems in the Philosophy of Criticism*. Indianapolis: Hackett Publishing.

Brehmer, B. (2009). Äquivalenzbeziehungen zwischen komparativen Phraseologismen im Serbischen und Deutschen. *Südslavistik online*, 1:141–164.

Filipović Kovačević, S. (2019). Metonymy-based Colour Metaphors Expressing Mind and Body States: Evidence from English and Serbian. *Godišnjak Filozofskog fakulteta u Novom Sadu*, 44(1):75–92.

[8] http://jerteh.rs/

Heiden, S., Magué, J.-P., and Pincemin, B. (2010). Txm: Une plateforme logicielle open-source pour la textométrie-conception et développement. In *10th International Conference on the Statistical Analysis of Textual Data-JADT 2010*, volume 2, pages 1021–1032. Edizioni Universitarie di Lettere Economia Diritto.

Heiden, S. (2010). The txm platform: Building open-source textual analysis software compatible with the tei encoding scheme. In *24th Pacific Asia conference on language, information and computation*, pages 389–398. Institute for Digital Enhancement of Cognitive Development, Waseda University.

Hu, X., Song, W., Liu, L., Zhao, X., and Du, C. (2017). Automatic recognition of simile based on sequential model. In *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*, volume 2, pages 410–413. IEEE.

Israel, M., Harding, J. R., and Tobin, V. (2004). On simile. In Achard, M. and Kemmer, S., Eds., *Language, Culture, and Mind*, pages 123–135. Stanford: Center for the Study of Language and Information.

Krstev, C., Obradović, I., Stanković, R., and Vitas, D. (2013). An approach to efficient processing of multi-word units. In *Computational Linguistics*, pages 109–129. Springer.

Krstev, C. (2008). *Processing of Serbian – Automata, Texts and Electronic dictionaries*. Faculty of Philology, University of Belgrade.

Liu, L., Hu, X., Song, W., Fu, R., Liu, T., and Hu, G. (2018). Neural multitask learning for simile recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553, Brussels, Belgium, October-November. Association for Computational Linguistics.

Milošević, N. and Nenadić, G. (2016). As Cool as a Cucumber: Towards a Corpus of Contemporary Similes in Serbian. *arXiv preprint arXiv:1605.06319*.

Milošević, N. and Nenadić, G. (2018). Creating a contemporary corpus of similes in Serbian by using natural language processing. *arXiv preprint arXiv:1811.10422*.

Mitrović, J., Markantonatou, S., and Krstev, C. (2019). A cross-linguistic study on Greek and Serbian fixed similes and enrichment of lexical resources via crowdsourcing. In *Multiword expressions: drawing on data from Modern Greek and other languages. Bulletin of Scientific Terminology and Neologisms*, pages 1–17. Academy of Athens.

Mladenović, M. and Mitrović, J. (2013). Ontology of Rhetorical Figures for Serbian. In Habernal, I. and Matoušek, V., Eds., *Text, Speech, and Dialogue*, pages 386–393, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mladenović, M. (2016). Ontology-based rhetorical figures recognition. *Infotheca - Journal for Digital Humanities*, 16(1-2):24–47.

Mpouli, S. and Ganascia, J.-G. (2015). "Pale as death" or " pâle comme la mort": Frozen similes used as literary clichés. In *EUROPHRAS2015:Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, Malaga, Spain, June.

Mpouli, S. (2016). *Automatic annotation of similes in literary texts*. Ph.D. thesis, Université Pierre et Marie Curie.

Niculae, V. and Danescu-Niculescu-Mizil, C. (2014). Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2008–2018, Doha, Qatar, October. Association for Computational Linguistics.

Niculae, V. and Yaneva, V. (2013). Computational considerations of comparisons and similes. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 89–95, Sofia, Bulgaria, August. Association for Computational Linguistics.

Niculae, V. (2013). Comparison pattern matching and creative simile recognition. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, pages 110–114, Trento, Italy, November.

Qadir, A., Riloff, E., and Walker, M. A. (2015). Learning to recognize affective polarity in similes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 190–200, Lisbon, Portugal. Association for Computational Linguistics.

Qadir, A., Riloff, E., and Walker, M. A. (2016). Automatically Inferring Implicit Properties in Similes. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1223–1232, San Diego, California, June. Association for Computational Linguistics.

Savary, A. (2009). Multiflex: a multilingual finite-state tool for multi-word units. In *International Conference on Implementation and Application of Automata*, pages 237–240. Springer.

Yoshimura, E., Imono, M., Tsuchiya, S., and Watabe, H. (2015). A simile recognition system using a common-sense sensory association method. *Procedia Computer Science*, 60:55–62.

Zhang, P., Cai, Y., Chen, J., Chen, W., and Song, H. (2019). Combining part-of-speech tags and self-attention mechanism for simile recognition. *IEEE Access*, 7:163864–163876.

Classification of L2 Thesis Statement Writing Performance Using Syntactic Complexity Indices

Kutay Uzun
Trakya University, Turkey
Department of English Language Teaching
kutayuzun@trakya.edu.tr

Abstract

This study primarily aimed to find out if machine learning classification algorithms could accurately classify L2 thesis statement writing performance as high or low using syntactic complexity indices. Secondarily, the study aimed to reveal how the syntactic complexity indices from which classification algorithms gained the largest amount of information interacted with L2 thesis statement writing performance. The data set of the study consisted of 137 high-performing and 69 low-performing thesis statements written by undergraduate learners of English in a foreign language context. Experiments revealed that the Locally Weighted Learning algorithm could classify L2 thesis statement writing performance with 75.61% accuracy, 20.01% above the baseline. Balancing the data set via Synthetic Minority Oversampling produced the same accuracy percentage with the Stochastic Gradient Descent algorithm, resulting in a slight increase in Kappa Statistic. In both imbalanced and balanced data sets, it was seen that the number of coordinate phrases, coordinate phrase per t-unit, coordinate phrase per clause and verb phrase per t-unit were the variables from which the classification algorithms gained the largest amount of information. Mann-Whitney U tests showed that the high-performing thesis statements had a larger amount of coordinate phrases and higher ratios of coordinate phrase per t-unit and coordinate phrase per clause. The verb phrase per t-unit ratio was seen to be lower in high-performing thesis statements than their low-performing counterparts.

Keywords: L2 Writing Performance, Machine Learning, Syntactic Complexity, Thesis Statement, Performance Classification

1. Introduction

Writing in L2 is one of the difficult skills within the higher education context where most assignments and exams need to be performed and evaluated in written form. This difficulty comes from the fact that writing in L2 requires a variety of motor skills and memory resources for the successful completion of the task (Burdick et al., 2013). Even though there is a large body of research investigating the factors that have effects on L2 writing performance in general, much of the previous literature on the construct seems to fall behind the advances in computational linguistics which offer numerous opportunities to second language acquisition/learning researchers by allowing them to analyse large chunks of learner texts by means of natural language processing and corpus analysis methods (Meurers, 2012).

Many of the texts written in the higher education context are in the form of essays, which typically have a main idea expressed as the thesis statement. Borrowing from Systemic Functional Linguistics the concept of macro-theme (Halliday and Mathiessen, 2004; Martin, 1992), Miller and

Pessoa (2016) define a thesis statement as a generalized main idea, located typically at the end of an introduction paragraph, which serves to predict the overall development of a text by stating the topic and making suggestions regarding how a particular point of view would be supported. Burstein et al. (2001) define the concept in a similar way, indicating that a thesis statement is an explicitly stated sentence that includes the main idea and the purpose of a text. From these definitions, it is seen that a thesis statement is mainly a summary of the core of a text, stating the central claim and the argumentative structure explicitly.

The importance of the thesis statement in writing stems from the fact that it does not only carry the main idea of a text, but it is also a sufficiently powerful part of a text that distinguishes a high-quality text from a low quality one. For instance, Coffin (2006) states that a successful essay in history writing contains a macro-theme which suggests the development of the text. Similarly, Oliveira's (2011) study reveals that history essays written by 11th-grade students were distinguishable in terms of success by having a macro-theme or not. In an English as a Foreign Language (EFL) context, Uzun (2019: 31) discovers that the thesis statement is the strongest rhetorical move in a literary analysis essay in terms of predicting total performance with the following equation for prediction intervals:

$$\text{Essay Score} = 18.377 + (\text{Thesis Statement Score} \times 3.748) \pm (1.96 \times 9.595)$$

Even though the literature indicates that the thesis statement is the most vital part of a text, it is seen that this particular part is yet an underresearched one. For this reason, it is argued in this study that the linguistic variables which contribute to a good thesis statement should be identified using corpus analysis and/or natural language processing methods.

Being an increasingly researched area by means of the mentioned corpus analysis and natural language processing methods, syntactic complexity appears to be an integral part of L2 writing quality. In general, syntactic complexity measures attempt to produce frequency counts of interconnected components within the structures of a language (Pallotti, 2014). Neary-Sundquist (2017) briefly describe those components as the length of certain phrases, their frequency per clause and the frequency of clauses per unit. According to Wolfe-Quintero, Inagaki and Kim (1998), the ratio of dependent clauses to clauses and clauses to t-units as well as the lengths of t-units and clauses are among the measures that can give clues regarding L2 writing performance. In addition, Ai and Lu (2013) suggest that the frequencies of subordination and coordination in addition to the length of production units are also among the syntactic complexity measures. Ortega (2003) suggests that the mean lengths of clause, t-unit and sentence are syntactic complexity measures, too. Casanave (1994) states that the amount of complex t-unit per t-unit is also a measure of the construct. Lu (2011) adds to the others by suggesting coordinate clauses per clause, coordinate phrase per t-unit, complex nominal per clause and complex nominal per t-unit as the measures positively correlated with syntactic complexity and dependent clause per t-unit and per clause as the negatively correlated measures.

Studies of Biber, Gray and Staples (2016), Staples and Reppen (2016), Yang, Lu and Weigle (2015) and Casal and Lee (2019) reveal that syntactic complexity and L2 writing quality are related constructs with higher levels of complexity indicating higher quality and lower levels indicating lower quality in L2 writing according to the findings. The exception to this is Crossley and McNamara's (2014) study, in which they reveal that there is no statistically significant correlation between phrasal syntactic complexity measures L2 writing quality. However, to the researcher's knowledge, none of these studies have a particular focus on the thesis statement, which is the strongest predictor of writing quality as mentioned above.

Considering the significance of both the thesis statement and syntactic complexity in L2 writing performance, it can be said that searching for the syntactic complexity measures that signal L2 thesis statement writing performance seems to be a worthy endeavour. For this reason, this study aims to fill in a gap in the literature by identifying the syntactic complexity measures which can be utilized to identify L2 thesis statement writing performance. In line with the aims of the study, the following research questions have been formulated:

1. Can syntactic complexity measures accurately classify L2 thesis statement writing performance?
2. Can the accuracy of L2 thesis statement writing performance classification using syntactic complexity indices be increased by balancing the data set?

3. How do the syntactic complexity indices from which classification algorithms gain the largest amount of information interact with L2 thesis statement writing performance?

2. Methodology

The study employed a machine learning (ML) approach to solve the classification problem. ML is a subfield of artificial intelligence that is utilized to discover relationships, patterns or rules using statistical methods to solve prediction or classification problems (Hastie et al., 2009; Murphy, 2012; Witten and Frank, 2005). Since this study aimed to classify L2 thesis statement writing performance using syntactic complexity indices, ML was considered suitable for the purposes of the study.

2.1. Context

The research context was a compulsory English Literature course in the English Language Teaching department of a public university in Turkey. Aiming to teach students how to analyze and interpret literary texts written in English, which is their L2, the English Literature course requires an extensive use of essay writing skills. The essays that the students write in this course are literary analysis essays, in which they write their personal interpretations of how a theme or character is presented in a text or how a particular concept is functionally used to form the plot structure.
The literary analysis essays within the context of the study are typically in the expository or argumentative style, 400-600 words in length and include an introduction (stating the background to the work and the thesis of the essay), main body (presenting, supporting/proving and concluding arguments) and conclusion (consolidating the thesis and stating personal opinion).

2.2. The Corpus

For the creation of a corpus relevant to the research aims, 206 literary analysis essays were chosen by the researcher. These were reliably scored in previous studies using the Genre-Based Literary Analysis Essay Scoring Rubric (Uzun, 2019; Uzun, In Press; Uzun, Unpublished Manuscript, Uzun & Zehir Topkaya, 2019), which is an analytical scoring rubric that is used to score each rhetorical move in a literary analysis essay and produce a total score between 0 and 100. The rubric allows for the scoring of the thesis statement separately between 0 and 15 where 15 is suitable for a thesis statement which provides a direct response to the essay question with at least two points that can be developed and justified in the main body, using appropriate grammar and lexis.

Within the research context, the thesis statement of a literary analysis essay is typically located at the end of the introduction paragraph and it can be in the form of a single sentence or a few related sentences (Uzun, 2019). Considering this description, the thesis statements of the essays were manually extracted along with their thesis statement scores by the researcher.

As a result, a corpus of 206 thesis statements with a sum of 3946 words (M = 19.16, SD = 8.93) were obtained. The thesis statements within the corpus had a minimum of 5 and a maximum of 69 words. In accordance with the scoring weights of the rubric, all thesis statements had scores between 0 and 15 (M = 10.97, SD = 3.47).

2.3. The Dataset

Each thesis statement was analysed using the web-based L2 Syntactic Complexity Analyzer (L2SCA) developed by Lu (2010), Lu (2011), Ai and Lu (2013) and Lu and Ai (2015). L2SCA (available for public use on https://aihaiyang.com/software/) is a web-based piece of software which was written in Python and generates syntactic complexity indices by means of Natural Language Processing methods, part-of-speech tagging and morphological analyses. The following variables, all of which were continuous, were obtained in this study as a result of the analyses:

- Word count (W)
- Sentence count (S)
- Verb phrase count (VP)
- Clause count (C)
- Clause per sentence (C/S)
- Verb phrase per t-unit (VP/T)
- Clause per t-unit (C/T)
- Dependent clause per clause (DC/C)

- T-unit count (T)
- Dependent clause count (DC)
- Complex T-unit count (CT)
- Coordinate phrase count (CP)
- Complex nominal count (CN)
- Mean length of sentence (MLS)
- Mean length of t-unit (MLT)
- Mean length of clause (MLC)
- Dependent clause per t-unit (DC/T)
- T-unit per sentence (T/S)
- Complex t-unit ratio (CT/T)
- Coordinate phrase per t-unit (CP/T)
- Coordinate phrase per clause (CP/C)
- Complex nominal per t-unit (CN/T)
- Complex nominal per clause (CN/C)

The operational definitions of the key terms related to the variables are presented below in Table 1.

Term	**Definition**	**Source**
Sentence	Group of words ending with a sentence-final punctuation mark	Lu (2011)
Clause	Group of words with a subject, finite verb but no nonfinite verbs	Lu (2011)
Dependent Clause	A finite nominal, adjective or adverbial clause	Lu (2011)
T-unit	A main clause + any subordinate clause or nonclausal structure	Hunt (1970)
Complex T-unit	A t-unit which contains at least one dependent clause	Lu (2011)
Coordinate Phrase	A coordinating verb, noun, adverb or adjective phrase	Lu (2011)
Verb Phrase	Finite or nonfinite verb phrases	Lu (2011)
Complex Nominal	1. Noun + participle, appositive, prepositional, possessive, adjective phrase or clause 2. A nominal clause 3. Gerund or infinitive as subject	Lu (2011)

Table 1. Operational Definitions of Key Terms

Following the computation of the mentioned variables, the thesis statement scores in the corpus were grouped as Low (n = 69, M = 6.74, SD = 2.00) and High (n = 137, M = 13.10, SD = 1.59) by means of a cluster analysis which produced a good fit with two clusters.

As an example of a high-scoring thesis statement, the following thesis statement, written as a response to the question "How is the concept of reputation presented in Beowulf?", can be seen:

> The concept of reputation in Beowulf is represented in through the main character in two aspects: victories of Beowulf and his loyalty. (Essay 28)

As seen above, the thesis statement provides a direct answer to the essay question and includes two arguable points (i.e. victories and loyalty) that can be further explained in the main body paragraphs of the essay. Also having a clear language appropriate for academic writing despite negligible errors, the thesis statement of Essay 28 has a score of 15/15.

The high-scoring thesis statement is a single clause and a single t-unit which has 22 words, one verb phrase, one coordinate phrase and three complex nominals. It does not have any dependent clause or complex t-unit.

An example of a low-scoring thesis statement in the corpus for the same question is given below:

> What given Beowulf by the poet as character are huge power and beautiful, faithful attitude. (Essay 5)

In this example, the thesis cannot be directly linked to the essay question unless the rater makes inferences which may or may not have been considered by the learner-writer. Moreover, erroneous grammar and low-level lexis is visible in the text. Therefore, it has a score of 1/15.

The low-scoring thesis statement in the example is also a single clause and single t-unit with 15 words, two verb phrases, one coordinate phrase and two complex nominals. The statement does not have any dependent clause or complex t-unit.

2.4. Experiment and Data Analysis

Weka 3.8.2 (Eibe, Mark & Witten, 2016) was used for the experiments. The data set of 137 high and 69 low-performing thesis statements was initially divided by 80:20 as training (n_{low} = 51, n_{high} = 114) and test (n_{low} = 18, n_{high} = 23) data to avoid overfitting. To get the baseline classification accuracy, the ZeroR algorithm was run using both sets of data, outputting 56.10% (*KS* = .00) classification accuracy. Following the computation of the baseline accuracy, Naïve-Bayes (NB), Logistic Regression (LR), Sequential Minimal Optimization (SMO), Stochastic Gradient Descent (SGD), KStar (K*), Instance-based Learning with Parameter K (Ibk), J48, Random Forest (RF), Locally Weighted Learning (LWL) and Random Tree (RT) algorithms were tested in terms of their classification accuracy.

Synthetic Minority Oversampling was used to balance the training data set due to its superiority over random resampling methods (Akbani et al., 2004). As a result, a balanced data set of 114 high-performing and 112 low-performing thesis statements was generated. The same algorithms were tested with the balanced data set.

A confusion matrix was produced for the most successful algorithm following the tests with the original and balanced data sets. To find out the variables which provided the largest amount of information to the classifiers, InfoGainAttributeEval algorithm was used.

Along with classification accuracy, the Kappa Statistic was also reported to control for the chance factor in the classification (Ben-David, 2008).

Since none of the variables which provided the largest amount of information to the classifiers was distributed normally, Mann-Whitney U tests were run to see how those variables interacted with high and low L2 thesis statement writing performance groups.

3. Results

The results of the experiments to find the best algorithm that would classify L2 thesis statement writing performance using syntactic complexity indices are presented below.

Algorithm	**Accuracy (%)**	***KS***
LWL	75.61	.50
LR	70.73	.41
SGD	70.73	.40
Ibk	68.29	.34
RF	65.85	.28
K*	65.85	.27
RT	65.85	.29
J48	63.42	.23
NB	60.98	.16
SMO	58.53	.10

Table 2. Classification Performance of Different Algorithms

As seen in Table 2, the best-performing algorithms to classify L2 thesis statement writing performance accurately were LWL, LR and SGD, which outputted classification accuracy percentages of 75.61 (*KS* = .50), 70.73 (*KS* = .41) and 70.73 (*KS* = .40) respectively. The values were seen to be 15-20% above the baseline accuracy. On the other hand, J48 ($\%_{accuracy}$ = 63.42, *KS* = .23), NB ($\%_{accuracy}$ = 60.98, *KS* = .16) and SMO ($\%_{accuracy}$ = 58.53, *KS* = .10) were seen to be the least successful algorithms, exceeding the baseline accuracy only by a few percents.
The confusion matrix for the LWL algorithm can be seen below in Table 3.

High/Low	**High**	**Low**
High	19	4
Low	6	12

Table 3. LWL Confusion Matrix

According to the matrix, the LWL algorithm classified 19 of 23 (82.61%) high-performing thesis statements and 12 of 18 low-performing thesis statements (66.67%) accurately using syntactic complexity indices. The precision, recall and F-measure values for this classification were .76, .76 and .75 respectively on average. For the high-performing thesis statements, the same values were .76, .83 and .79. They were seen to be slightly lower for the low-performing thesis statements, being .75, .67 and .71 in the same order.

Information gain ranking list for the LWL algorithm, obtained by means of the InfoGainAttributeEval algorithm is tabulated below in Table 4.

Average Merit	**Average Rank**	**Attribute**	
0.157 +- 0.013	1.4 +- 0.49	21	CP/C
0.153 +- 0.017	1.9 +- 0.83	8	CP
0.145 +- 0.012	2.7 +- 0.46	20	CP/T
0.089 +- 0.012	4.5 +- 0.81	14	VP/T
0.083 +- 0.014	5.5 +- 1.02	15	CT
0.079 +- 0.015	6.3 +- 1.42	13	C/S
0.069 +- 0.024	8.5 +- 4.01	17	DC/T
0.053 +- 0.028	10.4 +- 2.01	3	VP
0.056 +- 0.021	11.2 +- 3.43	16	DC/C
0.057 +- 0.021	11.2 +- 2.82	19	CT/T
0.046 +- 0.038	11.4 +- 5.39	12	MLC
0.048 +- 0.018	12.3 +- 1.9	7	C/T
0.044 +- 0.023	12.6 +- 1.85	6	DC
0.026 +- 0.033	13.6 +- 2.06	4	C
0.029 +- 0.036	14.5 +- 5.45	10	MLS
0 +- 0	15.2 +- 2.18	2	S
0 +- 0	15.3 +- 2.57	5	T
0 +- 0	17.4 +- 2.06	23	CN/C
0 +- 0	18.1 +- 2.26	9	CN
0 +- 0	19.8 +- 1.17	18	T/S
0.007 +- 0.021	19.9 +- 4.25	11	MLT
0 +- 0	20.8 +- 0.98	22	CN/T
0.009 +- 0.027	21.5 +- 4.5	1	W

Table 4. Information Gain Ranking List for LWL

As seen in the table, coordinate phrases per clause, the number of coordinate phrases, coordinate phrases per t-unit, verb phrases per t-unit and complex t-units were the attributes from which the largest amount of information was gained in the classification of L2 thesis statement writing performance using syntactic complexity indices. On the other hand, the mean length of sentences, the mean length of t-units and the number of words were the attributes from which the smallest amount of information was gained. No information was gained from the number of sentences, the number of t-units, complex nominal per clause, the number of complex nominals, t-units per sentence and complex nominal per t-units.

The results obtained with balanced data by means of SMOTE are presented below in Table 5.

Algorithm	**Accuracy (%)**	***KS***
SGD	75.61	0.51
LWL	75.61	0.50
LR	73.17	0.45
SMO	73.17	0.45
Ibk	68.29	0.34
RF	68.29	0.34
RT	65.85	0.28
J48	65.85	0.28
K*	65.85	0.27
NB	60.98	0.16

Table 5. Classification Performance of Different Algorithms with Balanced Data

As seen in Table 5, balancing the data set did not cause a significant change in the performance of the algorithms except for SGD, LR and SMO whose performance increased to some extent. In this dataset, SGD, LWL and LR were the most successful classifiers producing 75.61 (*KS* = .51), 75.61 (*KS* = .50) and 73.17 (*KS* = .0.45) percent classification accuracy respectively. The accuracy values obtained were seen to be 18-20% above the baseline accuracy. In this data set, J48, K* and NB were seen to be the least accurate classifiers, producing accuracy values 5-10% above the baseline.
The confusion matrix for the RF algorithm can be seen below in Table 6.

High/Low	**High**	**Low**
High	18	5
Low	5	13

Table 6. SGD Confusion Matrix for Balanced Data

The SGD algorithm could classify 18 of 23 high-performing thesis statements (78.26%) and 13 of 18 (72.22%) low-performing thesis statements (90.51%) accurately using syntactic complexity indices. The weighted average precision, recall and F-measure values for this classification were .76, .76 and .76 respectively. For the high-performing thesis statements, the same values were .78, .78 and .78. They were seen to be slightly lower for the low-performing thesis statements, being .72 for each of the values.
Information gain ranking list for the SGD algorithm, obtained by means of the InfoGainAttributeEval algorithm is tabulated below in Table 7.

Average Merit	**Average Rank**	**Attribute**	
0.325 +- 0.023	1.1 +- 0.3	8	CP
0.231 +- 0.047	2.5 +- 0.81	20	CP/T
0.227 +- 0.014	2.6 +- 0.66	21	CP/C
0.166 +- 0.026	4.6 +- 0.66	14	VP/T
0.158 +- 0.046	6.2 +- 4.07	23	CN/C
0.119 +- 0.011	7 +- 1.48	17	DC/T
0.119 +- 0.012	7.3 +- 0.9	3	VP
0.114 +- 0.01	7.7 +- 0.78	15	CT
0.105 +- 0.021	11 +- 3.69	12	MLC
0.101 +- 0.009	11.2 +- 1.47	6	DC
0.101 +- 0.009	11.5 +- 0.5	16	DC/C
0.104 +- 0.01	11.5 +- 2.54	13	CS
0.099 +- 0.033	12 +- 5.16	1	W
0.101 +- 0.009	12.5 +- 1.02	19	CT/T
0.101 +- 0.009	12.8 +- 1.25	7	C/T
0.076 +- 0.012	15.3 +- 2	10	MLS
0.071 +- 0.009	16.3 +- 0.9	4	C
0.048 +- 0.024	18.8 +- 1.66	11	MLT
0.036 +- 0.024	19.4 +- 1.43	22	CNT
0 +- 0	20.6 +- 1.2	2	S
0.01 +- 0.019	21.2 +- 1.6	9	CN
0 +- 0	21.3 +- 0.78	18	T/S
0 +- 0	21.6 +- 1.5	5	T

Table 7. Information Gain Ranking List for SGD

According to the results, the SGD algorithm gained the largest amount of information for the classification of L2 thesis statement writing performance using syntactic complexity indices from the number of coordinate phrases, coordinate phrases per t-unit and coordinate phrases per clause. On the other hand, the smallest amount of information was seen to have been gained by the algorithm for the classification task from the mean length of t-units, complex nominal per t-unit and the number of complex nominals. The number of t-units, t-units per sentence and the number of sentences were seen to have had no contribution to the algorithm for the classification task.

Since both LWL and SGD were found out to have gained the largest amount of information from the number of coordinate phrases, coordinate phrase per t-unit, coordinate phrase per clause and verb phrase per t-unit, how they interacted with high and low L2 thesis statement writing performance was tested by means of multiple t-tests. The findings are presented below in Table 8.

Index	Performance	Mean Rank	*U*	*Z*	*p*	*r*
CP	High	115.50	3082.00	4.885	<.001	.34
	Low	79.67				
CP/T	High	117.38	2824.50	5.289	<.001	.37
	Low	75.93				
CP/C	High	121.12	2312.50	6.495	<.001	.45
	Low	68.51				
VP/T	High	92.69	3245.50	3.997	<.001	.28
	Low	124.96				

Table 8. Mann-Whitney U Test Results for High (n = 137) and Low (n = 69) Score Groups

As seen in the table, L2 thesis statement writing performance differed according to the number of coordinate phrases ($Z = 4.89$, $p < .001$, $r = .34$), coordinate phrase per t-unit ($Z = 5.29$, $p < .001$, $r = .37$), coordinate phrase per clause ($Z = 6.50$, $p < .001$, $r = .45$) and verb phrase per t-unit ($Z = 4.00$, $p < .001$, $r = .28$), indicating small effects. The results indicated that the high-performing group had a higher number of coordinate phrases, coordinate phrase per t-unit and coordinate phrase per clause. On the other hand, verb phrase per t-unit ratio was higher in the low-performing group.

4. Discussion and Conclusion

This study mainly aimed to find out if L2 thesis statement writing performance could be successfully classified using syntactic complexity indices. The results showed that an identical classification accuracy percentage of 75.61, which exceeded the baseline accuracy of 55.60% by 20.01% could be obtained using the Locally Weighted Learning algorithm with the original imbalanced data set and the Stochastic Gradient Descent algorithm with the data set balanced by means of Synthetic Minority Oversampling. Even though the classification accuracy percentages were the same in both imbalanced and balanced data, it was seen that the balanced data set produced negligibly more successful results by classifying one more low-performing thesis statement and one fewer high-performing thesis statement accurately with a Kappa Statistic 1% higher than the imbalanced data set.

Being able to classify high and low performance in L2 thesis statement writing, syntactic complexity, indeed, seems to be an integral part of writing quality as suggested by Biber et al. (2016), Staples and Reppen (2016), Yang et al. (2015) and Casal and Lee (2019). Confirming the findings of those studies, the findings of this study revealed that L2 thesis statement writing performance could be classified in a way that exceeded the baseline accuracy to a considerable extent by means of a model solely based on syntactic complexity.

However, it was seen that 75.61% classification accuracy could not be increased in either imbalanced or balanced data. Even though this result exceeded the baseline accuracy percentage to a considerable extent, it appears that other features of L2 writing performance should also be included in classification models for increased classification accuracy. In this respect, a combination of lexical and syntactic complexity indices may result in a higher level of accuracy in the classification of L2 thesis statement writing performance.

An interesting finding was that the number of coordinate phrases, coordinate phrase per t-unit, coordinate phrase per clause and verb phrase per t-unit provided the largest amount of information to the classifiers in both imbalanced and balanced data sets. Further analyses showed that a higher number of coordinate phrases and higher ratios of coordinate phrase per t-unit and coordinate phrase per clause were present in the high-performing thesis statements. On the contrary, a lower ratio of verb phrase per t-unit was present in the high-performing group in comparison to the low-performing one. Apparently, high-performing L2 writers that produced the thesis statements in the data set resorted to coordination more often than their low-performing peers to join multiple concepts and ideas, which may have increased their performance in writing thesis statements in L2 by allowing them to express their textual interpretations from multiple perspectives. In the same vein, a lower ratio of verb phrases

per t-unit in those essays may be indicating that high-performing L2 writers made more extensive use of nominalization to express their ideas, avoiding narration through verb phrases, the overuse of which can be indicative of low performance in literary analysis essays (Uzun, 2016).

For further classification studies regarding L2 thesis statement writing performance, both lexical and syntactic complexity indices can be tested in a similar model to see if higher classification accuracy can be obtained. Moreover, which form of coordination, syndetic, asyndetic or polysyndetic, contributes better to L2 thesis statement writing performance was not investigated in this study. For this reason, further studies can be conducted to find out if a particular type of coordination contributes better to the construct. A higher percentage of classification accuracy in terms of L2 thesis statement writing performance can be used to develop automated feedback provision systems to scaffold learners into higher levels of L2 writing performance. Finally, the thesis statements investigated in this study were extracted from essays manually. An algorithm which tokenizes the sentences in an essay and detects the thesis statement automatically may allow for the analysis of larger data in a shorter amount of time, producing more precise findings.

References

Ai, H. and Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In A. Diaz-Negrillo, N. Ballier, P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data* (pp. 249-264). Amsterdam: John Benjamins Publishing Company.

Ai, Haiyang & Lu, Xiaofei (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson (eds.), Automatic Treatment and Analysis of Learner Corpus Data, pp. 249-264. Amsterdam/Philadelphia: John Benjamins.

Akbani R., Kwek S., Japkowicz N. (2004). Applying Support Vector Machines to Imbalanced Datasets. In J. F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), Machine Learning: ECML 2004. Lecture Notes in Computer Science, 3201. Springer, Berlin, Heidelberg.

Ben-David, A. (2008) Comparison of Classification Accuracy Using Cohen′s Weighted Kappa. *Expert Systems with Applications*, 34(2), 825-832. DOI: 10.1016/j.eswa.2006.10.022

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, 37(5) 639-669.

Briscoe, T. (2006). *An Introduction to Tag Sequence Grammars and the RASP System Parser*. Technical report, University of Cambridge, Computer Laboratory Technical Report.

Burdick, H., Swartz, C., Stenner, J., Fitzgerald, J., Burdick, D., and Hanlon, S. (2013). Measuring students′ writing ability on a computer-analytic developmental scale: An exploratory validity study. *Literacy Research & Instruction,* 52:255–280. doi:10.1080/19388071.2013.812162

Burnard, L. (2005). *Developing Linguistic Corpora: a Guide to Good Practice*, chapter Metadata for corpus work. Oxford: Oxbow Books. `http://ota.ahds.ac.uk/documents/creating/dlc/index.htm`.

Burstein, J., D. Marcu, S. Andreyev, and M. Chodorow (2001). Towards automatic classification of discourse elements in essays. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France (pp. 98−105). Association for Computational Linguistics.

Casal, J. E., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. Journal of Second Language Writing, 44, 51–62. doi:10.1016/j.jslw.2019.03.005

Casanave, C. P. (1994). Language development in students′ journals. *Journal of Second Language Writing*, 3(3): 179–201. http://doi.org/10.1016/1060-3743(94)90016-7

Christ, O. and Schulze, B. M. (1994). *The IMS Corpus Workbench: Corpus Query Processor (CQP) User's Manual*. University of Stuttgart, Germany.

Clear, J. (1992). Corpus Sampling. In Leitner, G., Ed., *New Directions in English Language Corpora*. Berlin: Mouton de Gruyter.

Coffin, C. (2006). *Historical discourse: The language of time, cause and evaluation*. London, England: Continuum.

Crossley, S. A., and McNamara, D. S. (2014). Does writing development equal writing quality?. A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26*, 66–79.

de Oliveira, L. C. (2011). Knowing and writing school history: The language of students' expository writing and teachers' expectations. Charlotte, NC: Information Age.

EAGLES. (1996). *EAGLES: Preliminary Recommendations on Corpus Typology. EAGLES Document EAG|TCWG|CTYP/P*. `http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html`.

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

Halliday, M. A. K., and Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar.* London, England: Hodder.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer, New York, NY

Hunt, K. W. (1970). Do sentences in the second language grow like those in the first? *TESOL Quarterly, 4*, 195–202. https://doi.org/10.2307/3585720.

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

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

Lu, X. (2011). A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers′ Language Development. *TESOL Quarterly*, 45(1): 36–62.

Lu, Xiaofei & Ai, Haiyang. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. Journal of Second Language Writing, 29, 16-27.

Lu, Xiaofei (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15(4):474-496.

Lu, Xiaofei (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers's language development. TESOL Quarterly, 45(1):36-62.

Martin, J. R. (1992). *English texts: System and structure*. Amsterdam, the Netherlands: John Benjamins.

Meurers, D. (2012). Natural language processing and language learning. In Carol A. Chapelle (Ed.), Encyclopedia of Applied Linguistics (pp. 4193–4205). Oxford, UK: Blackwell.

Miller, R. T. and Pessoa, S. (2016). Where's your thesis statement and what happened to your topic sentences? Identifying organizational challenges in undergraduate student argumentative writing. *TESOL Quarterly*, 7(4):847-873.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. Cambridge, MA: MIT Press.

Neary-Sundquist, C. (2017). Syntactic complexity at multiple proficiency levels of L2 German speech. *International Journal of Applied Linguistics*, 27(1):242-262.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4): 492–518.

Palotti, G. (2014). Revisiting the readability of management information systems journals again. *Research in Higher Education Journal, 15*:77-84

Staples, S., & Reppen, R. (2016). Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. Journal of Second Language Writing, 32, 17-35. https://doi.org/10.1016/j.jslw.2016.02.002

Uzun, K. (2016). Developing EAP writing skills through genre-based instruction: An action research. International journal of educational researchers, 7(2), 25-38.

Uzun, K. (2019). Using Regression to Reduce L2 Teachers′ Scoring Workload: Predicting Essay Quality from Several Rhetorical Moves. i-manager′s Journal on English Language Teaching, 9(3), 24-31.

Uzun, K. (in press). Future prediction of L2 writing performance: A machine learning approach. Journal of Educational Technology.

Uzun, K. (Unpublished Manuscript). Using rhetorical writing frames to enhance negotiated independent construction in L2 writing.

Uzun, K., and Zehir Topkaya, E. (2019). The Effects of Genre-Based Instruction and Genre-Focused Feedback on L2 Writing Performance. *Reading & Writing Quarterly: Overcoming Learning Difficulties.* https://doi.org/10.1080/10573569.2019.1661317

Witten, I. H. and Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). Elsevier, San Francisco, CA

Wolfe-Quintero, K., Inagaki, S. and Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity (No. 17)*. Honolulu, HI: University of Hawaii Press.

Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. Journal of Second Language Writing, 28, 53-67. https://doi.org/10.1016/j.jslw.2015.02.002

Categorisation of Bulgarian Legislative Documents

Nikola Obreshkov **Martin Yalamov** **Svetla Koeva**
Institute for Bulgarian Language "Prof. Lyubomir Andreychin"
Bulgarian Academy of Sciences
{nikola,martin,svetla}@dcl.bas.bg

Abstract

The paper presents the categorisation of Bulgarian MARCELL corpus in top-level EuroVoc domains. The Bulgarian MARCELL corpus is part of a recently developed multilingual corpus representing the national legislation in seven European countries. We performed several experiments with JEX Indexer, with neural networks and with a basic method measuring the domain-specific terms in documents annotated in advance with IATE terms and EuroVoc descriptors (combined with grouping of a primary document and its satellites, term extraction and parsing of the titles of the documents). The evaluation shows slight overweight of the basic method, which makes it appropriate as the categorisation should be a module of a NLP Pipeline for Bulgarian that is continuously feeding and annotating the Bulgarian MARCELL corpus with newly issued legislative documents.

Keywords: document categorisation, document classification, legislative domain

1. Introduction[1]

The paper presents the categorisation of Bulgarian MARCELL corpus in top-level EuroVoc domains[2]. The Bulgarian MARCELL corpus is part of a recently developed multilingual corpus representing the national legislation in seven European countries. The presented work is an outcome of the CEF Telecom project Multilingual Resources for CEF.AT in the Legal Domain1 (MARCELL) aiming to enhance the eTranslation system of the European Commission by supplying large-scale domain specific data in seven languages (Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovenian).

The Bulgarian MARCELL corpus consists of 27,283 documents (at the beginning of April 2020), which are classified into fifteen types: Administrative Court, Agreements, Amendments (Legislative acts), Compacts, Conventions, Decrees, Decrees of the Council of Ministers, Decisions of the Central Election Commission, Decisions of the Constitutional Court, Decisions of the Council of Ministries, Guidelines, Instructions, Laws (Acts), Memorandums and Resolutions.

Classifying the national legislation documents into EuroVoc classes serves the purpose of compiling multilingual domain-specific corpora corresponding to top-level EuroVoc domains. Only a few of the national legislations in the seven countries have been (manually) classified so far according to the EuroVoc Thesaurus (Croatian and Slovenian). The initial task was to categorise national legislation documents with the JRC EuroVoc indexer software – JEX (Steinberger et al., 2012). The high number of categories used by JEX Indexer, combined with a very unevenly balanced training set, is a big challenge for a multi-label categorisation task and even bigger for a one-label classification task. We show that

[1] Sections 1, 2, 3, 4, 7 are written by Sv. Koeva.
[2] https://op.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoc

taking into consideration the specific properties of legislative documents (namely, the specific terminology and the structure of the titles used in legislative documents) can be exploited for the document classification task.

The paper is organised as follows: in Section 2, we present in brief the related work in the field of categorisation of legislative documents. Sections 3 and 4 describe the specific tasks we are going to solve and the preliminary processing of the documents. The methods we have used for categorisation of legislative documents are presented in Section 5, and the evaluation of the results is presented in Section 6. Finally, Section 7 presents conclusions for our results and explains how the categorisation of legislative documents will be further enriched. The target result is a large-scale monolingual corpus of Bulgarian national legislation organised by EuroVoc top-level domains in thematically related sets of documents.

2. Related work

The EUR-Lex2 database of legal documents of the European Union served as a document collection for several classification methods. (Mencia and Frnkranz, 2007) studied multi-label classification problems, the largest being the categorisation of the EUR-Lex legal documents into the EuroVoc concept hierarchy with almost 4,000 classes. Three algorithms were evaluated: (i) the binary relevance approach, which independently trains one classifier per label; (ii) the multi-class multi-label perceptron algorithm, which respects dependencies between the base classifiers; and (iii) the multi-label pairwise perceptron algorithm, which trains one classifier for each pair of labels, the latest showing a good predictive performance.

Some of our experiments are performed with JEX Indexer – a free, multi-class, multi-label classification tool (Steinberger et al., 2012) provided with pre-trained models for 27 languages, including Bulgarian. The document to be indexed is represented as a vector of the same features (inflected word forms, n-grams, etc.) with their frequency in the document. The training documents (22,692 for Bulgarian covering 2,147 EuroVoc categories) are represented as a log-likelihood-weighted list of features, using the training document set as the reference corpus. The most appropriate categories for the new document are found by ranking the category vector representations according to their cosine similarity with the vector representation of the document to be indexed. JEX uses large numbers of stop words (332 for Bulgarian) that are ignored in the classification process. In order to optimise the profile generation for each class, a number of different parameter settings were optimised by selecting the best-performing setting within a range of values. The following are some of the most important parameters used: How many training documents there must be at least for a class to be trained; How long these training documents must be at least; How often words need to be found in the corpus in order to be used as associates; How statistically relevant a word must be in a training document in order to be considered; How to weigh words depending on the number of descriptors assigned to each training document (Steinberger et al., 2012). The reported precision for Bulgarian is 0,4619, recall – 0,5120 and F1 – 0,4940.

Filtz et al. (2019) uses different approaches to compare the performance of text classification algorithms on existing datasets and corpora of legal documents. For the EUR-Lex legal datasets, the authors show that exploiting the hierarchy of the EuroVoc thesaurus helps to improve classification performance by reducing the number of potential classes while retaining the informative value of the classification itself. Their results suggest that the advantage of using neural networks for the legal document classification problem is lower compared to text classification in other domains.

There are many examples for classification using variants of recurrent or convolutional neural networks (Howard, 2018; Jacovi, 2018). Some recent efforts are towards the so-called Extreme multi-label text classification (XMTC) – the most relevant class labels from an extremely large label collection are assigned to each document (Liu et al., 2017: 115). Kim (2014) reported on a series of experiments with convolutional neural networks (CNN) trained on top of pre-trained word vectors for sentence-level classification tasks and showed that a simple CNN with little hyperparameter tuning and static vectors achieved good results on multiple benchmarks. Liu et al. (2017) applied also deep learning to XMTC,

with a family of Convolutional Neural Network models, which are tailored for multi-label classification and reported results on several benchmark datasets, including EUR-Lex.

Chalkidis et al. (2019) released a new dataset of 57k legislative documents from EUR-Lex, annotated with concepts from EuroVoc. The dataset is substantially larger than previous EUR-Lex datasets and suitable for XMTC. Experiments with several neural classifiers were performed, and it is claimed that BIGRU with self-attention outperforms the current multi-label state-of-the-art methods, which employ label-wise attention.

To sum up, although the neural networks are widely used in classification tasks, there are results showing that the neural networks might not be very appropriate for particular domains, including legislative documents. It is very difficult to produce or reuse large training datasets in the legal domain, and such do not exist (to the best of our knowledge) for legislative documents (in Bulgarian).

3. Problem and Proposed Approach

Our efforts are directed towards the categorisation of Bulgarian legislative documents in top-level EuroVoc domains. Most of the classification approaches use a limited number of classification labels. The EuroVoc thesaurus contains 7,139 descriptors (labels) and is appropriate for the classification of documents in a multi-label classification. In contrast, our task is the classification of legislative documents into one of the top-level domains of EuroVoc: Politics, International relations, European Union, Law, Economics, Finance, Social questions, Education and Communications, Science, Business and Competition, Employment and Working conditions, Transport, Environment, Agriculture, Forestry and Fisheries, Agri-foodstuffs, Production, Technology and Research, Energy, Industry, Geography, International organisations. The limitations of the EuroVoc thesaurus are: it has been designed to meet the needs of systems of general documentation on the activities of the European Union; it cannot cover the various national situations at a sufficiently detailed level[3]. We reduced the classes to 19, excluding Geography and International organisations, as they are not representative for the national legislation.

We decided to make several experiments: a) with JEX indexer converting its multi-label categorisation to one-label categorisation; b) with neural networks using an unbalanced training set for Bulgarian annotated with IATE terms and EuroVoc descriptors; c) with a method measuring the domain-specific IATE terms and EuroVoc descriptors in the documents. The last approach is combined with term extraction, grouping of the primary document and its secondments and categorization by the titles of the documents.

4. Pre-processing of Documents

4.1. Part-of-Speech Tagging and Lemmatisation

For pre-processing Bulgarian legislative documents, we use the pipeline that integrates a sentence splitter, a tokeniser, a part-of-speech tagger, a lemmatiser, a named entity recogniser, a noun phrase parser, an IATE term annotator and a EuroVoc descriptor annotator. All tools are self-contained and part of them are designed to work in a chain, i.e. the output of the previous component is the input for the next component, starting from the sentence splitter and following the strict order for the tokeniser, the POS tagger and the lemmatiser (Bulgarian Language Processing Chain – BGLPC). In particular, we use enhanced versions of the sentence splitter, the tokeniser, the part-of-speech tagger and the lemmatiser for tagging and lemmatisation (Koeva and Genov, 2011). The output is in the CoNLL- U Plus format[4].

4.2. Term Recognition

The term recognition7 is performed via automatic text analysis methods in order to identify words and multiword expressions fulfilling the criteria for terms. The focus texts and the reference texts (texts from literature and news that are supposed not to contain terms) are tagged for part-of-speech and lemmatised. This ensures that each multiword term in the focus texts can be matched against the following linguistic filters (N, AN, AAN, NRN, ANRN, NRAN, ANRAN, NN, ANN, NAN, where A is adjective, N –

[3] https://eur-lex.europa.eu

[4] https://universaldependencies.org/format.html

noun, R – preposition) and that frequencies can be calculated correctly when terms are used in different word forms. For each sequence of part-of-speech tags in the focus texts matching one of the linguistic filters and for each adjective from the reference corpus the following information was indexed: the number of texts in which they occur and the number of all occurrences. Multiword term candidates that contain indexed reference adjectives are eliminated.

To compare the number of occurrences of term candidates in the focus texts with the number of their occurrences in the reference corpus TF-IDF and Log Likelihood algorithms are implemented[5]. The threshold for TF-IDF is set to 0.02. We use the union of the results from Tf-IDF and Log Likelihood. To increase the results the algorithm Dice is applied, which identifies terms similar to those already recognised (with a threshold set to 0.85). Processing the legislative corpus, we extracted 813,118 term candidates.

4.3. Annotation with IATE Terms and EuroVoc Descriptors

The Bulgarian MARCELL corpus has been annotated with terminology from two terminology repositories: IATE – 'Interactive Terminology for Europe'[6], the EU's terminology database used in the EU institutions and agencies, and EuroVoc[7], a multilingual, multidisciplinary thesaurus covering the activities of the EU.

For IATE term and EuroVoc descriptor annotation, a dedicated instrument called TextAnnotator was developed (Koeva et al., 2020). The TextAnnotator[8] calls dictionaries of terms and finds occurrences of these terms in the documents. Both the documents and the dictionaries are structured in the CoNLL-U Plus format (token, part-of-speech tag, lemma, extended grammatical tags) and each token is associated with a term descriptor. The annotation tool matches sequences of lemmas and part-of-speech tags of dictionary entries and lemmas and part-of-speech tags of document tokens. The matching procedure is based on a hash table indexing. For each dictionary entry, a hash key is generated concatenating lemmas and part-of-speech tags within it. All hash keys for a given dictionary are grouped into length classes based on the number of tokens they contain. The algorithm gives a priority to the longest length classes, which ensures the selection of longest matches. When a match is found, the corresponding tokens in the document are annotated with a term (Identification numbers of IATE terms or EuroVoc descriptors) and the processing continues from the end index of the match. The identification numbers (IDs) of the IATE terms point also to the relevant EuroVoc domains and subdomains. There are 45,592 IATE terms for Bulgarian. The annotation takes into account that several terms can be related with one and the same IATE ID (synonymy) and one term can be related with different IDs (polysemy). There are also IATE terms in Bulgarian, which describe concepts specific for other languages, i.e. община (obshtina) IATE ID: 3553038 'regions of Poland'. Such terms were excluded from the annotation (4,641 terms altogether). As a result, 13,799,334 IATE terms and 3,386,437 EuroVoc descriptors were annotated. IATE terms and EuroVoc descriptors are numbered within a given sentence (starting from 1) and the number is repeated for each token belonging to the term. The IATE identification number for the term is listed followed by the numbers of the corresponding EuroVoc descriptor(s).

4.4. Grouping of Documents

The documents are related to a primary legislative act, if such exists (i.e., a law and an instruction to this law). The documents' titles are divided into two parts: a general part that describes the type of the document (i.e. Закон за висшето образование (Zakon za visshetoto obrazovanie – 'The higher education act') and a differential part that describes the topic of the document (i.e. Закон за висшето образование (Zakon za visshetoto obrazovanie 'The higher education act'). The title of a primary document contains exactly two parts, while the titles of satellite documents contain at least two general parts and one differential part. The documents build a group if their differential parts and the nearest general parts match. The titles are lemmatised; then only lemmas are further used for matching. There

[5] The Term Recognition is developed by DImitar Georgiev.

[6] https://iate.europa.eu/home

[7] https://op.europa.eu/en/web/eu-vocabularies/th-dataset/-/resource/dataset/eurovoc

[8] The TextAnnotator is developed by Nikola Obreshkov.

might be differences in using punctuation marks, brackets, capital letters, abbreviations, dates, etc. in the title of the primary legislative act and its secondments. Several procedures are performed to predict possible differences and to match correctly thematically related documents.

4.5. Test Dataset

A manually annotated test corpus was developed including a total of 667 documents. 275 documents[9] were manually annotated with multiple labels among which the most appropriate label was also selected. For the annotation of the remaining 392 documents, 2000 documents were automatically annotated in advance with multiple labels which assisted the manual annotation with the most appropriate label. The features of the EuroVoc thesaurus discussed above were reflected in the fact that in some cases it was difficult even for a human expert to classify a given legislative document to the EuroVoc top-domains (many domains that are subject of legislation are not present, i.e. Health, Culture, etc.).

Code	EuroVoc Domain Name	Number of Documents
04	Politics	129
24	Finance	61
66	Energy	57
08	International relations	56
12	Law	52
44	Employment and working conditions	51
28	Social questions	37
52	Environment	32
56	Agriculture, forestry and fisheries	32
20	Economics	30
32	Education and communications	28
60	Agri-foodstuffs	20
48	Transport	18
68	Industry	15
10	European Union	11
40	Business and competition	10
64	Production, technology and research	10
16	Economics	6
36	Science	3

Table 1: The distribution of documents in the test corpus

5. Experiments

5.1. Categorisation with JEX Indexer

For a given document, JEX Indexer assigns several labels among more than 6,700 EuroVoc descriptors with corresponding likelihood weights. The default settings of the system were used: at least 4 training

[9] The 275 documents were manually annotated by Ts. Dimitrova and V. Stefanova.

documents for a class; at least 100 words per training document; at least 4 occurrences of a word in the corpus which are used as associates; minimum log-likelihood value of 5 to consider a word statistically relevant in a training document. Example 1 shows a document categorised with six labels (category code is the EuroVoc descriptors' ID) and the assigned log-likelihood weight.

```
<document id="bg-81407.txt">
    <category code="4585" weight="0.18519803030161675 "></category>
    <category code="1684" weight="0.18393845477668894 "></category>
    <category code="1234" weight="0.16739820210223233 "></category>
    <category code="365" weight="0.13740067398049766 "></category>
    <category code="1021" weight="0.11888676884171023 "></category>
    <category code="2900" weight="0.11647743754115164 "></category>
</document>
```

Figure 1: JEX Indexer categorisation output for document bg-81407.txt

The one-label categorisation of Bulgarian legislative texts was performed in two steps:

Each document was annotated with weighted EuroVoc descriptors using the JEX Indexer tool.

The annotated descriptors were grouped into one top-domain by the hierarchical relations to broaden terms up to the top-level as well as by the associative relations to related terms.

If more than one descriptor points to a top-domain, the weights of descriptors are summed up. For example, the document bg-81407.txt is classified to the following top-domains:

24 0.725 FINANCE

20 0.184 TRADE

The weight 0.725 for top-domain Finance is calculated by summing the weights of 5 descriptors, assigned by the JEX indexer: 4585 данък върху добавената стойност (danak varhu dobavenata stoynost) 'value added tax', 1234 фискална хармонизация (fiskalna harmonizaciya) 'tax harmonisation', 365 данъчно облекчение (danachno oblekchenie) 'tax relief', 1021 данъчна система (danachna sistema) 'tax system' and 2900 данъчна основа (danachna osnova) 'basis of tax assessment'. The experiment was repeated by setting a threshold for summed log-likelihood values to 0.1, 0.2, and 0.3 respectively.

5.2. Neural Categorisation

First, we tested the JEX language model trained with the European legal texts, but the results were not good. Therefore, we opted to train a new model based on the Bulgarian MARCELL corpus. We use the version of the corpus annotated with IATE terms and EuroVoc descriptions. The annotated terms were linked to the EuroVoc top-domains and the documents were sorted by the number of the top-domain associations. For every EuroVoc top-domain up to 500 training documents were selected; however, for some top-domains the number of associated documents was very small: Energy – 82, Production, Technology and Research – 10, International organisations – 4 and so on.

The generated training corpus was used to train a neural model with TensorFlow[10] and Keras[11]. The built-in Keras tokeniser is used for tokenisation. First, a vocabulary is constructed containing unique tokens in the documents without taking into account their frequency. Then each document is transformed into a sequence of integers based on the presence of its words in the vocabulary. The selected design of the neural network returns the credibility for each label candidate. A threshold is set up to specify the level of credibility.

[10] https://www.tensorflow.org/

[11] https://keras.io/

5.3. Basic Categorisation

The basic categorisation relies on pre-processing: sentence splitting, tokenisation, part-of-speech tagging, lemmatisation, annotation with IATE terms and EuroVoc descriptors. The IATE subject fields link IATE terms with EuroVoc descriptors; the fields of knowledge in which the IATE concepts are used (one IATE term can be linked with several EuroVoc descriptors). The annotated EuroVoc descriptors were grouped into top-domains by the associative relations and the hierarchical relations linking the EuroVoc descriptors. For each document the obtained top-domains are summed up and the numbers of associations to top-domains are sorted and the top-domain with highest number of associations is selected as the document category. Figure 1 shows the basic categorisation pipeline.

5.4. Basic Categorisation Combined with Term Extraction

The legislative domain is a domain with unique or specialised terminology. As we do not analyse the context and disambiguate the word senses, we decided to experiment with term extraction. The basic categorisation is performed in the same manner with the only difference that IATE and EuroVoc dictionaries used for the IATE term and EuroVoc descriptor annotation are filtered to contain only the obtained term candidates.

5.5. Basic Categorisation Combined with Documents' Grouping

We also use the groups formed by a primary legislative act and its satellite documents. The assumption is that all thematically related documents should belong to one category. Based on this assumption two different experiments were performed on top of the Basic classification:

Normal Grouping - All documents within the group are considered as a single document. This experiment is implemented on top of the basic categorisation. For every such document, the EuroVoc term annotations are combined and the EuroVoc top-domain associations are recalculated.

Hierarchical Grouping - Only the primary document is used to represent the whole group. Its EuroVoc term annotations are used to select a class that is then assigned to its related satellite documents.

5.6. Basic Categorisation Combined with Titles' Classification

The method uses the EuroVoc descriptors occurring within the documents' titles and their association with EuroVoc top-domains. The results are scored based on counting the descriptors' lemma matches in different combinations. The most probable top-level domain is the one that has the highest score, which is calculated with the following formula:

score = TotalPhraseMatches * 10 + TotalDescriptorMatches + TotalDescriptorSequenceMatches/100 where:

TotalPhraseMatches is defined as the total number of branches pointing to a top-domain where a lemmatised descriptor is matched with a part of the title (regardless of the word order if the descriptor is a multiword term);

TotalDescriptorsMatches is defined as the total number of single lemmatised descriptors matched in the title for each top-domain;

TotalDescriptorSequenceMatches is defined as the total number of branches pointing to a top-domain where a lemmatised descriptor is matched with a part of the title keeping the word order if the descriptor is a multiword term.

The top category assigned by the titles' categorisation is compared with the category candidates obtained by the basic categorisation. Two scenarios are possible:

The top category of the titles' categorisation is not among the category candidates of the basic categorisation. In this case, the basic categorisation is not affected.

The top category of the titles' categorisation overlaps with one of the category candidates of the basic categorisation. In this case, the respective top-domain annotations count is multiplied by a predefined weight. This may result in reordering the category candidates and promoting a different top category for the current document.

6. Evaluation of Results

For every document the manually annotated class (category) from test corpus was compared with the suggested class by the selected classification method. Initially, per-class precision and recall were evaluated, where

per-class precision = correctly predicted documents with this class / all predicted documents with this class

per-class recall = correctly predicted documents with this class / all documents with this class

Then a macro-averaged precision and a macro-averaged recall were calculated for the whole classifier. This was done with arithmetic mean of the per-class precision and recall. Finally, the harmonic mean of the macro-averaged precision and the macro-averaged recall was used for the F1-score estimation. F1 = 2 * precision * recall / (precision + recall).

The results of the described methods for text categorisation of the Bulgarian legislative documents are presented in Table 2.

Categorization method	Accuracy	Precision	Recall	F1
JEX Indexer	46.36	44.58	45.43	45
Neural model	16.71	15.53	17.38	16.04
Categorisation by Titles	37.39	51.34	40.98	45.58
Basic method	54.42	52.64	70.18	61.16
Basic method + Term Extraction	36.28	40.53	45.42	42.84
Basic method + Normal Grouping	54.03	51.84	69.91	59.54
Basic method + Hierarchical Grouping	54.64	52.08	70.13	60.24
Basic method + Categorisation by Titles (1.3)	57.83	55.04	72.44	62.55
Basic method + Categorisation by Titles (1.5)	58.3	55.23	72.41	62.66
Basic method + Categorisation by Titles (2)	55.66	53.91	57.97	55.87
Basic method + Hierarchical Grouping + Categorisation by Titles (1.5)	58.51	55.37	72.52	62.8

Table 2: The evaluation results

The results achieved with the JEX indexer and the Neural model were not optimal and their improvement may require manual annotation of a large training corpus. For the current task these methods were omitted and the focus were pointed towards the basic method which doesn't need a training corpus. Given the challenges of this classification task, the baseline results achieved by the Basic method were reasonable and seven additional experiments were performed in pursuit of further improvement. The Term Extraction filtering resulted in a dropout of the measured scores. Similarly, the Normal Grouping did not improve the results indicating that the primary document of the group holds the essential (most relevant) information. This hypothesis was also confirmed by the Hierarchical Grouping method where only the primary document was used. It produced better results than the Normal Grouping method and a slight increase of precision and a slight decrease of recall compared to the Basic method. Three experiments were performed with the combination of the Basic method and the Titles classification. For every experiment a different weight was used to control the importance of the Title class when applied to the Basic method classification. The best results were achieved with a weight of 1.5. The last experiment was to combine the Basic method with the Hierarchical Grouping and with the Titles classification weighted with 1.5. This combination led to the best results from all experiments. It

can be seen that the addition of Hierarchical Grouping improved the result of the Basic method and Titles classification both in precision and recall.

7. Conclusions

The categorisation of legislative documents is a part of the NLP pipeline for Bulgarian, which continuously feeds the Bulgarian MARCELL corpus with newly issued legislative documents and makes changes to the data format, organises data in structures, accumulates data with linguistic information, analyses data and provides explicit links between different data segments. The absence of a relevant training corpus with legislative data in Bulgarian (and the stumbling block for creation of a training dataset relating with the relatively small number of documents in the legislative domain) presupposes the limited performance of neural methods of any supervised machine learning approaches. The target result - a large-scale monolingual corpus of Bulgarian national legislation categorised by EuroVoc top-level domains - is achieved by applying a Basic method which relies on the annotation of IATE terms and EuroVoc descriptors within a document. Some restrictions have been applied to reduce the ambiguity effect (manual removal of inappropriate annotations with more than 4 IATE terms and use of the IATE subject fields that link IATE terms with EuroVoc descriptors). The assumption that the legislative documents that are linked to a primary legislative act must belong to the same category, and that the titles of legislative documents contain information about their category led to the performance of seven experiments. The results show that the specific properties of legislative documents (legislative terminology, relations between legislative acts and the structure of the titles used in legislative documents) can be successfully exploited for the document classification task in the legislative domain.

Acknowledgements

The presented work is an outcome of the CEF Telecom project Multilingual Resources for CEF.AT in the Legal Domain1 (MARCELL).

The Term Recognition is developed by Dimitar Georgiev.

References

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I. (2019). *Extreme Multi-Label Legal Text Classification: A case study in EU Legislation.* CoRR abs/1905.10892.

Filtz, E., Kirrane, S., Polleres, A., Wohlgenannt, G. (2019). *Exploiting EuroVoc's Hierarchical Structure for Classifying Legal Documents.* Lecture Notes in Computer Science On the Move to Meaningful Internet Systems: OTM 2019 Conferences, pages 164-181.

Howard, J. and Ruder, S. (2018). *Fine-tuned language models for text classification.* CoRR abs/ 1801.06146, `http://arxiv.org/abs/1801.06146`.

Jacovi, A., Shalom, O.S., Goldberg, Y. (2018). *Understanding convolutional neural networks for text classification.* CoRR abs/1809.08037.

Kim, Y. (2014). *Convolutional Neural Networks for Sentence Classification.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746– 1751.

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

Koeva, S., Obreshkov, N., Yalamov, M. (2020). *Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents.* Proceedings of The 12th Language Resources and Evaluation Conference. European Language Resources Association, 2020, 6988–6994.

Liu, J., Chang, W., Wu, Y. and Yang, Y. (2017). *Deep Learning for Extreme Multi-label Text Classification.* Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17, pages 115–124.

Mencia, E. L. M. and Frnkranz, J. (2007). *Efficient Multilabel Classification Algorithms for Large-Scale Problems in the Legal Domain.* Proceedings of the LWA 2007, pages 126–132.

Steinberger R., Ebrahim, M. and Turchi, M. (2012). *JRC EuroVoc Indexer JEX - A freely available multi-label categorisation tool.* Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'2012), Istanbul, 21-27 May 2012.

Controlling Chat Bot Multi-Document Navigation with the Extended Discourse Trees

Dmitry Ilvovsky
Kazan Federal University,
Kazan, Russia
National Research University
Higher School of Economics,
Moscow, Russia
dilvovsky@hse.ru

Alexander Kirillovich
Kazan Federal University,
Kazan, Russia
alik.kirillovich@gmail.com

Boris Galitsky
Oracle Inc.,
Redwood Shores, CA, USA
bgalitsky@hotmail.com

Abstract

In this paper we learn how to manage a dialogue relying on discourse of its utterances. We define extended discourse trees, introduce means to manipulate with them, and outline scenarios of multi-document navigation to extend the abilities of the interactive information retrieval-based chat bot. We also provide evaluation results of the comparison between conventional search and chat bot enriched with the multi-document navigation.

Keywords: Discourse tree, Dialogue management, Rhetoric structure, Linguistic Linked Open Data

1. Introduction

In this paper we extend the abilities of the interactive chat bot initially developed by (Galitsky, Ilvovsky, 2017) and further improved in (Galitsky, 2019; Galitsky et al., 2019; Galitsky and Ilvovsky 2019). In practice, this chat bot is oriented to work with English language but our approach is language independent. The approach we introduce in this paper is inspired by an idea of a guided search. One source of it is a search methodology designed to show a user an array of different visual possibilities where a searching user may proceed. This is done instead of just navigating to an end point or a terminal answer. This search feature is not looking at images but rather the way those images have been described by users. As particular descriptors show up with sufficient frequency, the system turns them into the categories and sub-categories that accompany search results. This approach is also referred to as faceted search allowing users to narrow down search results by applying multiple filters (Galitsky et al., 2009; Galitsky and McKenna, 2017).

To provide a systematic navigation means to take a user through content exploration, we intend to build upon discourse trees (DTs) for texts and extend the discourse analysis based on RST (Mann and Thompson, 1988) to the level of a corpus of documents. We believe that knowledge exploration should be driven by navigating a discourse tree built for the whole corpus of relevant content. We refer to such a tree as extended discourse tree (EDT). It is a combination of discourse trees of individual paragraphs first across paragraphs in a document and then across documents.

A search engine does not provide a means to navigate through content: it is retained for a search user. Instead, search engine builds an inverse index so that for each query keywords it stores information which paragraph of which document these keywords occur in. Therefore, once a query including multiple documents is submitted, the search engine knows which paragraphs in which documents it should take a search user to.

Most chat bots are designed to imitate human intellectual activity maintaining a dialogue. They try to build a plausible sequence of words to serve as an automated response to user query. Instead, we

focus on a chat bot that helps a user to navigate to the exact, professionally written answer as fast as possible.

In addition to narrowing down, zooming into a certain piece of content as search engines do, a chat bot is expected to provide navigational means for content exploration. Therefore, we extend the notion of search inverse index to the one not only allowing to zoom in based on keywords but also on drilling in / drilling out / drilling back in, based on how documents are interconnected.

2. Dialogue Management Approach

2.1. Controlling Chat Bot Navigating with the Extended Discourse Tree

To control the chat bot navigation in a general case, beyond clarification scenarios, we introduce the notion of an extended discourse tree. A conventional discourse tree expresses the author flow of thoughts at the level of paragraph or multiple paragraphs. Conventional discourse tree becomes fairly inaccurate when applied to larger text fragments, or documents. Hence, we extend the notion of a linguistic discourse tree towards an extended discourse tree, a representation for the set of inter-connected documents covering a topic. For a given paragraph, a DT is automatically built by the discourse parser (Joty et al., 2014). We then combine DTs for the paragraphs of documents to the EDT, which is a basis of an interactive content exploration facilitated by the chat bot. We apply structured learning of extended DTs to differentiate between good, cognitively plausible scenarios and counter-intuitive, non-cohesive ones. To provide cohesive answers, we use a measure of rhetorical agreement between a question and an answer by tree kernel learning of their discourse trees (Galitsky and Ilvovsky, 2017).

On the web, information is usually represented in web pages and documents, with certain section structure. Answering questions, forming topics of candidate answers and attempting to provide an answer based on user selected topic are the operations which can be represented with the help of a structure that includes the DTs of texts involved. When a certain portion of text is suggested to a user as an answer, this user might want to drill in something more specific, ascend to a more general level of knowledge or make a side move to a topic at the same level. These user intents of navigating from one portion of text to another can be represented as coordinate or subordinate discourse relations between these portions.

We merge the links between logical parts of paragraphs and the links between documents (Fig. 1). If at the current step the user is interested in drilling in, we navigate her through an Elaboration relation from nucleus to satellite within a paragraph or Elaboration hyperlink to a more specific document. Conversely, if a user decides that the suggested topic is not exactly what he is looking for and wants to return a higher-level view, the system navigates Elaboration relation in the inverse order from satellite to nucleus at either paragraph or intra-document level. The other navigation option is relying on Contrast or Condition relations exploring controversial topics (these rhetorical relations need to be recognized for inter-document case).

Navigation starts with the route node of a section that matches the user query most closely. Then the chat bot attempts to build a set of possible topics, possible understanding of user intent. To do that, it extracts phrases from elementary discourse units that are satellites of the route node of the DT. If the user accepts a given topic, the navigation continues along the chosen edge; otherwise, when no topic covers the user interest, the chat bot backtracks the discourse tree and proceeds to the other section (possibly of other documents) which matched the original user query second best. Inter-document and inter-section edges for relations such as Elaboration play similar role in knowledge exploration navigation to the internal edges of a conventional DT.

Figure 1: Illustration for the idea of extended DT: intra-paragraph rhetorical relations are combined with inter-document links also labelled as rhetorical relations

2.2. Constructing EDT

To construct EDT, the focus is on building rhetorical links between text fragments (called *elementary discourse units*, or EDU) in different paragraphs or documents. The main difficulty here is to identify a relationship between mentions. The other difficulty is to label an inter-document rhetorical relation. To address it, we form a fictitious text fragment from the respective text fragments of the original paragraph and perform coreferential analysis and discourse parsing.

The input of the EDT algorithm is a set of documents, and an output is an EDT that is encoded as a regular DT with the labels of document identification for each node. The processing flow is as follows:

1. Building a set of all DTs for each paragraph in each document *DTA*;
2. Iterate through all pairs of DT_i and $DT_j \in DTA$;
3. Identify noun phrases and named entities in DT_i and DT_j;
4. Compute overlap and identify common entities E_{ij} between DT_i and DT_j;
5. Establish relationships between occurrences of entities in $E_{i\,j}$ such as *equals, sub-entity, part-of*;
6. Confirm these relationships by forming text fragment merging EDU(E_i) and EDU(E_j) and applying coreference resolution;
7. Form an inter-paragraph rhetorical links R(E_{ij}) for each entity pair occurrence in E_{ij};
8. Classify rhetorical relation for each rhetorical link by forming a text fragment merging EDU(E_i) and EDU(E_j), building its DT and using recognized relation label for this rhetorical link.

To construct conventional DTs, we used existing discourse parser (Joty et al., 2014).

2.3. Example of Navigation

We now present an example of a content exploration scenario based on an extended DT covering three documents (Fig. 2):

Faceted Search
Facets correspond to properties of the information elements. They are often derived by analysis of the text of an item using entity extraction techniques or from pre-existing fields in a database such as author, descriptor, language, and format. Thus, existing web-pages, product descriptions or online collections of articles can be augmented with navigational facets. Within the academic community, faceted search has attracted interest primarily among library and information science researchers, but there is a limited interest of computer science researchers specializing in information retrieval

Entity Extraction
Entity extraction, also known as entity name extraction or named entity recognition, is an information retrieval technique that refers to the process of identifying and classifying key elements from text into pre-defined categories.

Information Retrieval
…

Example 1: Three documents

Exploration scenario is as follows (Fig. 2). Let us imagine that a user is asking a question '*What is faceted search?*'. To understand how it works, this user needs to become fluent with other associated concepts. The chat bot provides further content exploration or search options based on satellite EDUs in the DT of the document '*Faceted search*' (on the top-left). It built multiple DTs (one for each paragraph, two are shown) and formed the following items for content exploration:

- entity extraction;
- information retrieval;
- pre-existing fields in a database;
- augmented with navigational facets.

The user can either follow the link to land on a single piece of information or run a new search to get to multiple search results to choose from. If a user choses 'entity extraction', it is led to the respective document (on the top-right of Fig. 2). The chat bot proceeds to the next iteration, discovering the phrases from satellites of the DT node corresponding to 'entity extraction':

- entity recognition;
- information retrieval.

If a user now selects the second option he would navigate to the 'information retrieval' document.

Whereas a discourse tree of a sentence, paragraph or a document is a well-explored area, algorithms for building a discourse-level representation of a collection of documents in various formats and styles from different sources has not been explored. Irrespectively of the document granularity level, the same relationships such as Elaboration, Contrast and Attribution may hold between the certain portions of text across documents.

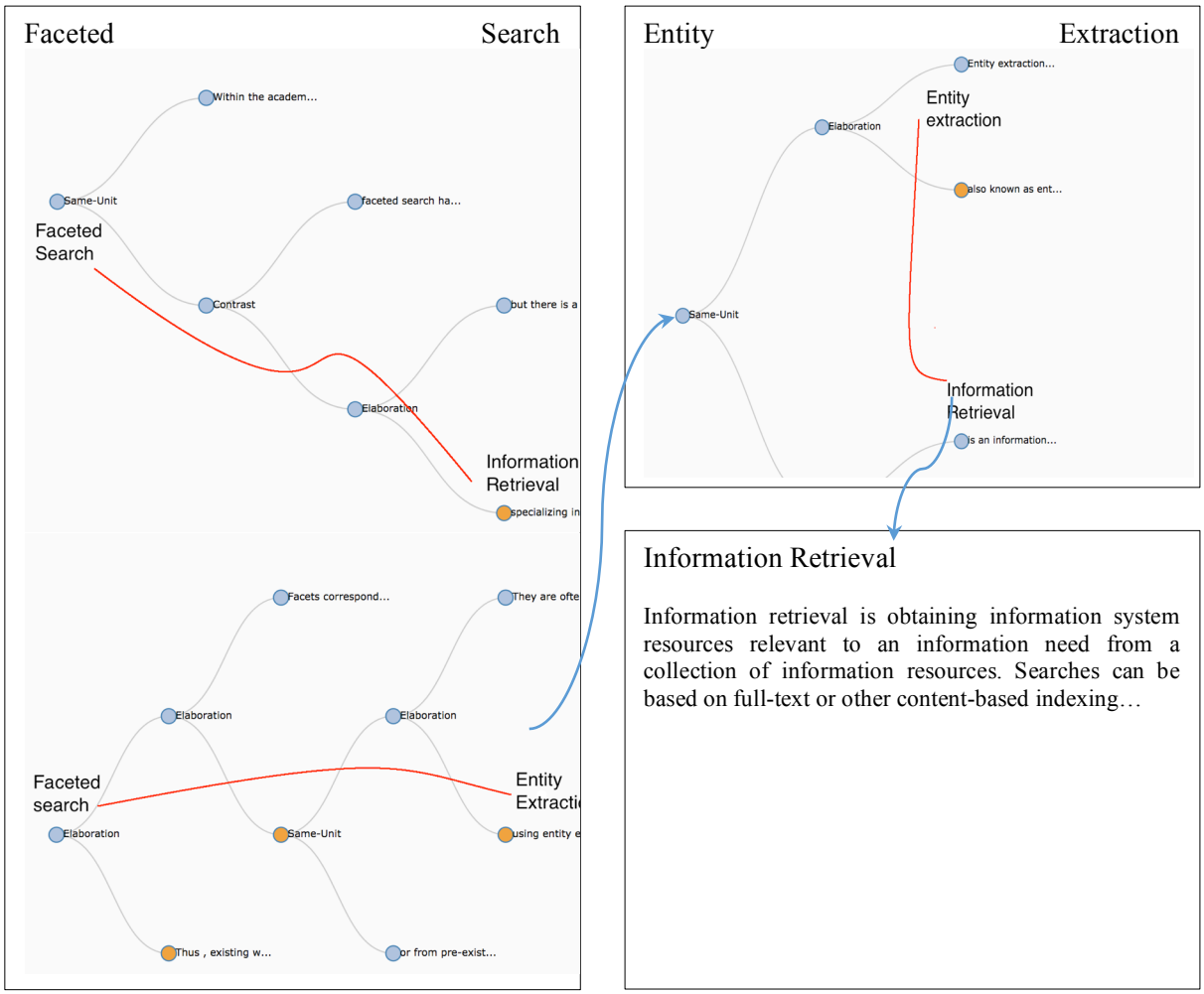


Figure 2: Extended discourse tree for a set of documents used to navigate to a satisfactory answer

3. Evaluation

We compared the efficiency of information access using the proposed chat bot in comparison with a major web search engines such as Google, for the queries where both systems have relevant answers. For search engines, misses are search results preceding the one relevant for a given user. For a chat bot, misses are answers which cause a user to choose other options suggested by the chat bot, or request other topics.

The topics of question included personal finance. Twelve users (author's colleagues) asked the chat bot 15-20 questions reflecting their financial situations, and stopped when they were either satisfied with an answer or dissatisfied and gave up. The same questions were sent to Google, and evaluators had to click on each search results snippet to get the document or a webpage and decide on whether they can be satisfied with it.

The structure of comparison of search efficiency for the chat bot vs the search engine is shown in Fig. 3. The left side of arrows shows that all search results (on the left) are used to form a list of topics for clarification. The arrow on the bottom shows that the bottom answer ended up being selected by the chat bot based on two rounds of user feedback and clarifications. Instead of looking into all search results to find the relevant one (on the left), a user answers a clarification request composed by the chat bot and drills into his topic of interest (on the right). The arrows show how multiple search results on distinct topics are converged to a single clarification request enumerating automatically extracted topics.

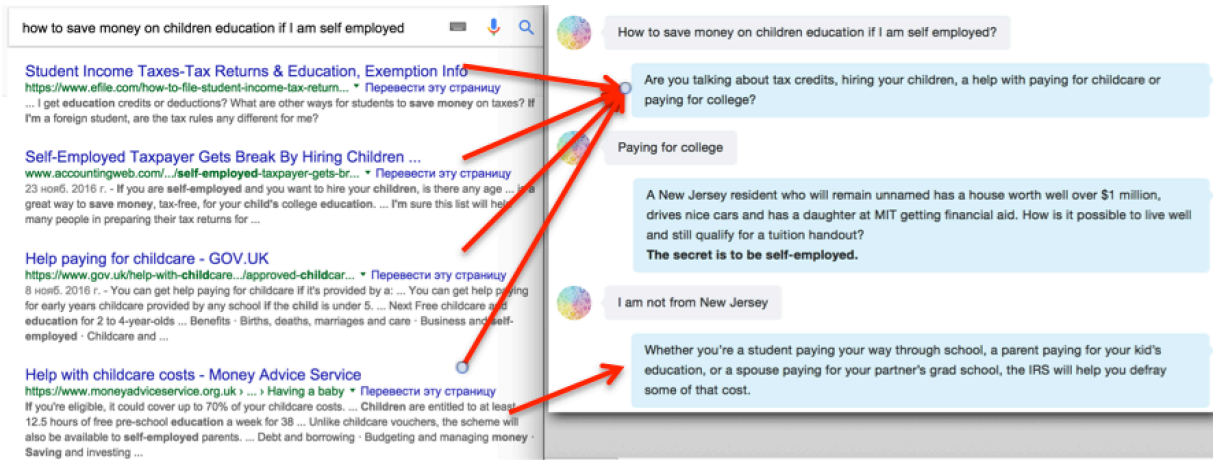


Figure 3: Comparing navigation in a search engine and the chat bot

One can observe (Table 1) that the chat bot time of knowledge exploration session is longer than for the search engine. Although it might seem to be less beneficial for users, businesses prefer users to stay longer on their websites, since the chance of user acquisition grows. Spending 7% more time on reading chat bot answers is expected to allow a user to better familiarize them with a domain, especially when these answers follow the selections of this user. The number of steps of an exploration session for chat bot is a quarter of what is required by a search engine. Traditional ways to measure search engine performance such as MAP and NDCG are also applicable for a comparison between conventional search engines and chat bots with respect to efficiency of information access (Sakai, 2007). We conclude that using a chat bot with extended discourse tree-driven navigation is an efficient and fruitful way of information access, in comparison with conventional search engines and chat bots focused on imitation of a human intellectual activity.

Parameter / search engine	Conventional web search	Chat bot
Average time to satisfactory search result, sec	45.3	58.1
Average time of unsatisfactory search session (ended in giving up and starting a new search,) sec	65.2	60.5
Average number of iterations to satisfactory search result	5.2	4.4
Average number of iterations to unsatisfactory search result	7.2	5.6

Table 1: Comparison for the chat bot and Google search in the domain of personal finance

4. Related Work

Radev (2000) introduced a cross-document structure theory (CST), a paradigm for multi-document analysis. CST takes into account the rhetorical structure of clusters of related textual documents. He specified taxonomy of relations between documents, cross-document links. CST is intended as a foundation to summarize a collection of documents initiated by a user as well as to navigate it by an abstract information-access machine.

To proceed from RST to CST, one cannot employ the deliberateness of writing style, rely on discourse markers within individual documents. However, it is possible to leverage a logical structure

across documents which are systematic, predictable and useful. CST attempts to attach a certain reasoning flow to an imaginary "collective" author of a set of documents.

One of the first studies of rhetorical relations between documents is presented in (Trigg and Weiser, 1987) for scientific papers, such as citation, refutation, revision, equivalence, and comparison. These rhetorical relations are grouped into Normal (inter-document relations) and Commentary (deliberate cross-document relations). However, it is hard to see this model's applicability beyond the scientific domain.

One way to represent the multi-document navigation structure is a multi-document cube. It is a three-dimensional structure that represents related documents with dimensions of *time* (ordered), *source* (unordered) and *position within the document* (ordered).

Discourse disentanglement (such as classification of links between portions of texts or documents) and dialogue/speech/communicative act tagging have been extensively studied (Wang et al., 2011). Discourse disentanglement is the task of splitting a conversation (Elsner and Charniak, 2008) or documents (Wolf and Gibson, 2005) into a sequence of distinct portions of text (sub-discourses). The disentangled discourse is modelled via a tree structure (Grosz and Sidner 1986; Seo et al., 2009), an acyclic graph structure (Rose et al., 1995; Elsner and Charniak, 2008), or a cyclic chain graph structure (Wolf and Gibson, 2005). Speech acts are used to describe the function or role of an utterance in a discourse, similarly to our CDT representation, and have been employed for the analysis of communication means including conversational speech instant messaging, online forums (Kim et al., 2010; Galitsky et al., 2017) and chats (Galitsky and Ilvovsky, 2017). Automated answer scoring benefits from semantic and discourse analyses as well (Wanas et al., 2008). For a more complete review of models for discourse disentanglement and speech act tagging, we refer the reader to (Kim et al., 2010).

Wang et al. (2011) presented the task of parsing user forum threads to determine the labelled dependencies between posts. Three methods, including a dependency parsing approach, are proposed to jointly classify the links (relationships) between posts and the dialogue act (type) of each link. The authors predicted not only the links between posts, but also showed the type of each link, in the form of the discourse structure of the thread.

5. Conclusions and Future Work

We present the first version of a dialogue management system for a chat bot with iterative content exploration that leads a user through a personalized knowledge acquisition session. The chat bot is focused on automated customer support or product recommendation agent that assists a user in learning product features, product usability, suitability, troubleshooting and other related tasks.

The developed dialogue management system is based on the extended discourse trees model. The main contribution of this paper is that it demonstrates applicability of discourse trees in dialog management.

Our current work is undertaken on the following directions:

1) Keeping the topic. In the current version of the system, the chat-bot follows the user's questions, straying off the initial topic. This approach is useful for free conversation systems, but not for task-oriented chat-bots. Currently one of the authors is working on the new approach to dialog management, that tries to avoid digression and keep a user on the main topic of the dialog. We are going to present this new approach at the Dialogue 2020.

2) Linked Open Data integration. In question answering the current version of chat-bot relies only to the data extracted from text documents. Now we are working on complementing these data by the data from Linked Open Data cloud, including domain-independent DBpedia (Lehmann et al., 2015) and our domain-specific mathematical ontology OntoMath^{Edu} (Kirillovich et al., 2020). As an interface between natural language user query and LOD datasets we would rely on the resource from the Linguistic Linked Open Data cloud (Cimiano et al., 2020), such as LLOD representation of WordNet (McCrae et al., 2014), BabelNet (Ehrmann et al., 2014), RuThes (Kirillovich et al., 2017) and FrameNet (Rospocher et al., 2019). We expect that exploitation of LOD cloud can improve user's satisfaction against the baseline obtained in this work.

3) Supporting Russian dialogs. Although the developed approach is language-independent, its actual implementation relies on the discourse parser for English (Joty et.al.,2014) and so now can

work only with English texts. We are going to add support for Russian by retraining the parser on the Russian discourse corpus Ru-RSTreebank (Pisarevskaya et al., 2017). In order to achieve interoperability with the parser format, the corpus will be represented in terms of the OLiA Discourse Extensions ontology (Chiarcos, 2014) and integrated to the Linguistic Linked Open Data cloud.

Acknowledgements

The new results presented in this paper were received by D. Ilvovsky and A. Kirillovich during their work in KFU and were funded by Russian Science Foundation according to the research project no. 19-71-10056. These new results are based on the already published discourse models that were being developed by D. Ilvovsky and B. Galitsky in NRU HSE during the previous years.

References

Cimiano, P., Chiarcos, C., McCrae, J.P., and Gracia, J. (2020). Linguistic Linked Open Data Cloud. In Cimiano, P., et al. *Linguistic Linked Data: Representation, Generation and Applications*, pages 29–41. Springer.

Chiarcos, C. (2014). Towards interoperable discourse annotation: Discourse features in the Ontologies of Linguistic Annotation. In Calzolari, N., et al., Eds., *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 4569–4577. ELRA.

Ehrmann, M., Cecconi, F., Vannella, D., McCrae, J., Cimiano, P., and Navigli, R. (2014). Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0. In Calzolari, N., et al., Eds., *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 401–408. ELRA.

Elsner, M. and Charniak, E. (2008). You Talking to Me? A Corpus and Algorithm for Conversation Disentanglement. In Moore, J.D., et al., Eds., *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-08: HLT)*, pages 834–842. ACL.

Galitsky, B. and Ilvovsky, D. (2017). Chat bot with a Discourse Structure-Driven Dialogue Management. In Martins, A. and Peñas, A., Eds., *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017 Demo)*, pages 87–90. ACL.

Galitsky, B. and McKenna, E.W. (2017). *Sentiment Extraction from Consumer Reviews for Providing Product Recommendations*. US Patent 9646078B2.

Galitsky, B. (2019). Discourse Level Dialogue Management. In Galitsky, B. *Developing Enterprise Chatbots*, pages 365-426. Springer.

Galitsky, B., González, M.P., and Chesñevar, C.I. (2009). A novel approach for classifying customer complaints through graphs similarities in argumentative dialogue. *Decision Support Systems*, 46(3):717–729.

Galitsky, B. and Ilvovsky, D. (2019). On a Chatbot Conducting Virtual Dialogues. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM 2019)*, pages 2925–2928. ACM.

Galitsky, B, Ilvovsky, D., and Goncharova, E. (2019). On a Chatbot Conducting Dialogue-in-Dialogue. In Nakamura, S., et al., Eds., *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2019)*, pages 118–121. ACL.

Grosz, B.J. and Sidner, C.L. (1986). Attention, intention and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Joty, S.R. and Moschitti, A. (2014). Discriminative Reranking of Discourse Parses Using Tree Kernels. In Moschitti, A., et al., Eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 2049–2060. ACL.

Kim, S.N., Wang, L., and Baldwin, T. (2010). Tagging and Linking Web Forum Posts. In Lapata, M. and Sarkar, A., Eds., *Proceedings of the 14th Conference on Computational Natural Language Learning (CoNLL 2010)*, pages 192–202. ACL.

Kirillovich, A., Nevzorova, O., Falileeva, M., Lipachev, E., and Shakirova, L. (2020). OntoMath^{Edu}: a New Linguistically Grounded Educational Mathematical Ontology. In Benzmüller, C. and Miller, B., Eds., *Proceedings of the 13th International Conference on Intelligent Computer Mathematics (CICM 2020)*. Lecture Notes in Artificial Intelligence, vol. 12236. Springer (forthcoming).

Kirillovich, A., Nevzorova, O., Gimadiev. E., and Loukachevitch, N. (2017). RuThes Cloud: Towards a Multilevel Linguistic Linked Open Data Resource for Russian. In Różewski, P. and Lange, C., Eds., *Proceedings of the 8th International Conference on Knowledge Engineering and Semantic Web (KESW 2017)*. Communications in Computer and Information Science, vol. 786, pages. 38–52. Springer, Cham.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2015). DBpedia: A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 6(2):167–195.

Mann, W. and Thompson, S. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

McCrae, J. P., Fellbaum, C., and Cimiano, P. (2014). Publishing and Linking WordNet using lemon and RDF. In Chiarcos, C., et al., Eds., *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL-2014)*, pages. 13–16. ELRA.

Pisarevskaya, D., Ananyeva, M., Kobozeva, M., Nasedkin, A., Nikiforova, S., Pavlova, I., and Shelepov, A. (2017). Towards building a discourse-annotated corpus of Russian. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"*, volume 1, pages 201–212.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources step one: cross-document structure. In Dybkjær, L., et al., Eds., *Proceedings of the 1st SIGdial workshop on Discourse and dialogue (SIGDIAL '00)*, pages 74–83. ACL.

Rose, C.P, Di Eugenio, B., Levin, L.S., and Van Ess-Dykema, C. (1995). Discourse processing of dialogues with multiple threads. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL '95)*, pages 31–38. ACL.

Rospocher, M., Corcoglioniti, F., and Palmero Aprosio, A. (2019). PreMOn: LODifing linguistic predicate models. *Language Resources and Evaluation*, 53:499–524.

Sakai, T. (2007). Alternatives to Bpref. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*, pages 71–78. ACM.

Seo, J., Croft, W.B. and Smith, D.A. (2009). Online community search using thread structure. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09)*, pages 1907–1910. ACM.

Trigg, R.H. and Weiser, M. (1987). TEXTNET: A network-based approach to text handling. *ACM Transactions on Office Information Systems*, 4(1):1–23.

Wanas, N., El-Saban, M., Ashour, H., and Ammar, W. (2008). Automatic scoring of online discussion posts. In *Proceeding of the 2nd ACM workshop on Information credibility on the web (WICOW '08)*, pages 19–26. ACM.

Wang, L., Lui, M., Kim, S.N., Nivre, J., and Baldwin, T. (2011). Predicting thread discourse structure over technical web forums. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pages 13–25. ACM.

Wolf, F. and Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287.

Cross-lingual Transfer Learning for Semantic Role Labeling in Russian

Ilseyar Alimova
Kazan Federal University
Kazan, Russia
alimovaIlseyar@gmail.com

Elena Tutubalina
Kazan Federal University
Kazan, Russia
elvtutubalina@kpfu.ru

Alexander Kirillovich
Kazan Federal University
Kazan, Russia
alik.kirillovich@gmail.com

Abstract

This work is devoted to semantic role labeling (SRL) task in Russian. We investigate the role of transfer learning strategies between English FrameNet and Russian FrameBank corpora. We perform experiments with embeddings obtained from various types of multilingual language models, including BERT, XLM-R, MUSE, and LASER. For evaluation, we use a Russian FrameBank dataset. As source data for transfer learning, we experimented with the full version of FrameNet and the reduced dataset with a smaller number of semantic roles identical to FrameBank. Evaluation results demonstrate that BERT embeddings show the best transfer capabilities. The model with pretraining on the reduced English SRL data and fine-tuning on the Russian SRL data show macro-averaged F1-measure of 79.8%, which is above our baseline of 78.4%.

Keywords: Semantic Role Labeling, Transfer learning, Word embeddings, Deep Learning, FrameNet, FrameBank

1. Introduction

Semantic Role Labeling (SRL) is one of the most critical tasks in natural language processing (Palmer et al., 2010; Solovyev and Ivanov, 2016). The SRL aims to identify the situation a given sentence describes, find sentence constituents expressing the participants of this situation, and identify the roles the participants play.

Recent advances in multilingual neural network models offer new opportunities to improve SRL (Arkhipov et al., 2019; Okamura et al., 2018; Subburathinam et al., 2019). In this work, we take the task a step further from existing monolingual research (Shelmanov and Devyatkin, 2017; Larionov et al., 2019) by exploring multilingual transfer between semantic roles labeled datasets in different languages. Our goal is not to outperform state of the art models, but to ask whether we can transfer knowledge from a high-resource language, such as English, to a low-resource one, e.g., Russian, for SRL. In this work, we seek to answer the following research questions: **RQ1:** Will transfer learning (TL) help improve the results? **RQ2:** How the quality change if the roles in the training and target corpora will be the same? **RQ3:** Which multilingual pre-trained models are most effective for an SRL task?

We conducted experiments on two datasets: a database of Russian lexical constructions FrameBank and an English large-scale semantic database FrameNet. We consider four modern multilingual language models: BERT (Devlin et al., 2019), XLM-R (Conneau and Lample, 2019), MUSE (Lample et al., 2017), LASER (Artetxe and Schwenk, 2019). To our knowledge, this is the first work exploring the interlingual transfer ability for SRL in Russian.

2. Related Work

Various approaches have been proposed for the SRL task in English. Gildea and Jurafsky proposed the statistical classifiers with various lexical and syntactic features combined with knowledge of the predicate verb, noun, or adjective and the prior probabilities of multiple combinations of semantic roles

(Gildea and Jurafsky, 2002). The classifier was tested on the FrameNet corpus. The developed system performed 82% accuracy in identifying the semantic role of pre-segmented constituents and 63% of F-measure on simultaneously segmenting constituents and identifying their semantic role task. Pradhan et al. proposed the SRL system based on the Support Vector Machine classifier (Pradhan et al., 2005). The authors applied a new set of features, including dependency parses features extracted with a combination of Minipar syntactic parse, a chunked syntactic representation, and Charniak parses. The model with a single Charniak parser performed 83.7% of F-measure. A combination of syntactic parsers improved the results on 1,5% of F-measure. Collobert et al. proposed a simple multi-layer neural network that takes as an input the words decoded into a feature vector, by a lookup table operation (Collobert et al., 2011). However, their best system fell short of previous feature-based systems. The modern works apply complicated neural network architectures. He et al. introduced deep highway BiLSTM architecture with constrained decoding (He et al., 2017). The network achieved 83.2% of F-measure on the CoNLL 2005 test set and 83.4 of F-measure on CoNLL 2012 datasets.

The development of the Fremebank corpus led to the growth of studies devoted to SRL for the Russian language. Kuznetsov (2013) proposed a baseline system for SRL in Russian . The system consists of the following parts: text preprocessing module (morphological analysis, lemmatization, syntactic analysis), data enrichment module (mapping text segments annotation to syntax tree nodes), training module (feature extraction, classification, optimization). The system with verbs form, predicate lemmas, and syntactic features obtained 76.1% of F-measure. Adding a combination of semantic and syntactic features increased the results to 76.4% of F-measure. Shelmanov and Devyatkin (2017) applied two neural networks for SRL on the FrameBank corpus. The first neural network model has the simple architecture that acquires all features of an argument: sparse and dense, as a single vector and propagates them through three dense layers. The second complex neural network has the same types of layers. However, the first layer is split into several chunks: a chunk for categorical features, a chunk for an argument embedding, and a chunk for a predicate embedding. The categorical features include various morphological, the relative position of an argument in a sentence, predicate lemma, the preposition of an argument, and the name of a syntax link from argument to its parent features. The authors investigated the ability to learn a model for labeling arguments of "known" and "unknown" predicates that are present and not present in a training set, respectively. The complex neural network achieved 82.3% of micro F-score on "known" predicates and 66.7% on "unknown" predicates and outperformed the simple network on 6.2% and 34.8%, respectively. Larionov et al. (2019) evaluated various pretrained language models, including, word2vec, fasttext, ELMo, BERT, RuBERT. For "known" predicates, the ELMo-based model performed the highest micro F-measure (83.42%), and the RuBERT model outperformed other models in terms of macro F-measure (80.12%). For 'unknown" predicates, ELMo performed the highest metrics both for macro and micro F-measures (37.64% and 55.50%, respectively). A recent study applied a frame-based approach for predicting sentiment attibutes towards named entities in political news (Rusnachenko et al., 2019; Loukachevitch and Rusnachenko, 2020).

To sum up, machine learning approaches with contextual embeddings have a high potential for the SRL task. More recently, multilingual embeddings have been used to achieve state-of-the-art performance on many NLP tasks such as named entity recognition and classification (Devlin et al., 2019; Conneau and Lample, 2019; Artetxe and Schwenk, 2019; Miftahutdinov et al., 2020). The goal of this study is to investigate cross-lingual transfer methods for SRL that exploit resources from existing high-resource language, i.e. English, and fine-tuning on Russian data.

3. Datasets

In this section, we describe two datasets for SRL in Russian and in English. Transfer learning aims to solve the problem on a "target" dataset using knowledge learned from a "source" dataset. We use the English FrameNet dataset (Baker et al., 1998) as source data and the Russian FrameBank dataset (Lyashevskaya and Kashkin, 2015; Lyashevskaya, 2012) as target data. We study two setups for the source side: (i) FULL data and (ii) REDUCED data setup that we describe in Section 3.3.

3.1. FrameBank

FrameBank[1] (Lyashevskaya and Kashkin, 2015; Lyashevskaya, 2012) is a database that consists of a dictionary of Russian lexical constructions and an annotated corpus of their realizations in contemporary written texts. In the dictionary, each verb or other predicate word is followed by a list of constructions in which it serves as a target word. Construction is a morphosyntactic template, where some elements are fixed lexical units, and some are variable slots. A typical construction describes the argument structure of a verb. It consists of the one fixed element, representing the verb and one or more variable slots, representing arguments of this verb. Less frequent constructions are ones describing the argument structure of other parts of speech (POS) and complex idiomatic phrases.

Description of construction elements includes:

- the syntactic rank (Subject, Object, Predicate, Peripheral, Clause);
- the morphosyntactic features (POS, case, and preposition marking);
- the semantic role (Agent, Patient, Instrument, Theme, etc.);
- the lexical-semantic class (person, animal, building, abstract entity, etc.).

The annotated corpus consists of construction realization examples in the Russian National Corpus. Each example is linked to the construction it instantiates, and the parts of the example sentence are linked to the construction slots. These parts are annotated by their actual syntactic and morphological features, which can be different from the features, prescribed by the corresponding construction.

The publicly available version of FrameBank contains 16123 constructions for 1589 target words, realized by 52737 annotation sets.

3.2. FrameNet

FrameNet[2] (Baker et al., 1998) is a large-scale semantic resource, organized as a network of frames. A frame is a description of an abstract situation and its participants, called frame elements. For example, the frame elements of the *Commerce_buy* frame are *Buyer*, *Goods*, *Seller*, etc. Frames are interlinked by several relation types, including inheritance, perspective on, subframes, etc. For example, the *Commerce_buy* frame and its frame elements inherit from the *Getting* frame, and its *Recipient*, *Theme*, *Source* frame elements, respectively. A frame is associated with lexical units, i.e., disambiguated words, evoking this frame. For example, the *Commerce_buy* frame is evoked by "buy" and "purchase" lexical units.

In FrameNet, the network of frames is complemented by the corpus of annotated sentences. In each sentence, one word (typically, a verb) is a lexical unit, evoking a frame, and other sentence constituents express elements of this frame. For example, in the "John bought a car from Mary" sentence, "bought" is a lexical unit, evoking the *Commerce_buy* frame, while "John", "a car" and "Mary" express *Buyer*, *Goods*, *Seller* frame elements, respectively.

The currents version of FrameNet contains 1224 frames, evoked by 13676 lexical units and 202970 annotation sets.

3.3. Linking of FrameNet roles to FrameBank

Concerning the SRL task, there are several significant differences between FrameNet and FrameBank. First, in FrameNet, the frames are defined as generalized language-independent situations, while any FrameBank construction is defined for a particular target word. Second, FrameNet frame elements are defined locally for each particular frame (for example, the *Commercial_transaction* frame defines the roles of the buyer, seller, good, etc., and the *Theft* frame defines the roles of perpetrator, victim, good, etc.). In contrast, FrameBank roles are defined globally for all constructions (for example, the construction for the word *kupit* 'to buy' and for the word *ukrast* 'to steal' use the roles from the same globally defined pool: agent, patient, theme, etc.).

[1] https://github.com/olesar/FrameBank

[2] https://framenet.icsi.berkeley.edu/

To mitigate these differences in the FrameNet, we study two setups for the source side. First, for the FULL setup, we use the entire FrameNet data as a starting point to train a neural model for SRL. We used lexical units as predicates, words annotated in corpus with frame element as arguments, and frame element's name as roles. Second, for the REDUCED setup, we left examples with roles present in the FrameBank corpus. To identify matching roles, we took the translation of roles provided in the FrameBank corpus. The final REDUCED FrameNet corpus includes a total of 86951 examples and 19 roles. For each annotated corpus, we created all possible pairs of predicate and argument and obtained 505 940 samples.

4. Experiments and Evaluation

In this section, we describe the model architecture, pretrained language models, and results of experiments.

4.1. Model

We implemented the neural network proposed in (Larionov et al., 2019). The network contains three input layers for the embedding of an argument, the embedding of a predicate and feature embeddings, and sparse categorical features. The input data fed separately to dense and batch normalization layers. Concatenated outputs of the first layer are fed to dense, batch normalization and dropout layers. The last output dense layer with a softmax activation function makes a classification.

The model takes as an input following features:

- Various morphological characteristics of both an argument and a predicate (case, valency, verb form, etc.);
- Relative position of an argument in a sentence concerning a predicate.
- Preposition of an argument extracted from a syntax tree (including a complex preposition as a single string);
- Name of a syntactic link that connects an argument token to its parent in the syntax tree;
- Argument and predicate lemmas.

We used a maximum of 50 and 15 epochs to train the model on FrameNet and FrameBank, respectively. For both corpora, we utilized the batch size of 32 and Adam optimizer. We applied the implementation of the model from this repository[3].

4.2. Pretrained Language Models

We consider four modern multilingual language models: BERT (Devlin et al., 2019), XLM-R (Conneau and Lample, 2019), MUSE (Lample et al., 2017), LASER (Artetxe and Schwenk, 2019). For arguments and predicates consisting of several words, we used averaged vectors. Further, we provide a detailed description of each model.

BERT (Bidirectional Encoder Representations from Transformers) is a recent neural network model for NLP presented by Google (Devlin et al., 2019). BERT is based on bidirectional attention-based Transformer architecture (Vaswani et al., 2017). In particular, we applied BERT$_{base}$, Multilingual Cased (Multi-BERT), which is pretrained on 104 languages and has 12 heads, 12 layers, 768 hidden units per layer, and a total of 110M parameters. For each predicate and argument, we use the BERT output layer to obtain embeddings without using context in sentences. Besides, we obtained contextualized vectors, when the whole sentence was fed to the input of the network (**BERT-context**).

XLM-R improves the multilingual BERT model by incorporating a cross-lingual task of translation language modeling, which performs masked language modeling on a concatenation of parallel bilingual sentence pairs (Ruder et al., 2019). The model is also based on Transformer architecture (Vaswani et al.,

[3]https://github.com/IINemo/isanlp_srl_framebank

2017). We applied the XLM-R Masked Language Model, which is pretrained on 2.5 TB of CommonCrawl data, in 100 languages, with 8 heads, 6 layers, 1024 hidden units per layer.

MUSE (Multilingual Unsupervised and Supervised Embeddings) is a sentence encoding model simultaneously trained on multiple tasks and multiple languages able to create a single embedding space to 30 languages (Lample et al., 2017). The vectors obtained with fastText library (Bojanowski et al., 2017) pretrained on texts from Wikipedia. The length of the obtained vectors is 300.

LASER (Language-Agnostic SEntence Representations) is a library to calculate and use multilingual sentence embeddings (Artetxe and Schwenk, 2019). LASER is based on encoder-decoder architecture proposed in (Schwenk, 2018). The model was trained on Wikipedia texts and Billions of High-Quality Parallel Sentences on the WEB in 93 languages. The length of the obtained vectors is 1024.

4.3. Corpora preprocessing

For FrameBank corpus, we made the same text processing as in (Larionov et al., 2019). We filtered the dataset keeping only predicates with at least 10 examples and dropped infrequent roles, for which the dataset contains less than 180 samples. The final corpus version contains 52,751 examples for 44 unique semantic roles.

To obtain features the following linguistic processing steps were performed:

- tokenization and sentence splitting with NLTK library (Schneider and Wooters, 2017);
- lemmatization, POS-tagging, and morphological analysis with MyStem library (Segalovich, 2003);
- syntax parsing via UDPipe parser (Straka and Straková, 2017) with model trained on SynTagRus (Nivre et al., 2008).

These steps are implemented using a publicly available IsaNLP library[4].

4.4. Results

We compare all models in terms of macro-averaged precision (P), recall (R), and F1-measure (F). Training and testing sets of FrameBank are adopted from (Larionov et al., 2019) for a fair comparison. The results of multilingual models as well as state-of-the-art **RuBERT** model from (Larionov et al., 2019) are presented in Table 1. **RuBERT** is the Russian Cased BERT pretrained on the Russian part of Wikipedia and news data (Kuratov and Arkhipov, 2019); it has 12 heads, 12 layers, 768 hidden units per layer, and a total of 180M parameters; Multi-BERT was used for initialization, while the vocabulary of Russian subtokens was built on the training dataset.

There are several conclusions to be drawn based on the results in Table 1. First, the models with BERT-context and XLM-R embeddings show the best results among non-pretrained models in terms of F-measure (78.4% and 78.3%, respectively). The model with BERT-context embeddings achieves the highest precision (82.8%), while the model with XLM-R demonstrates the highest recall (76.5%). The model with BERT-based embeddings for individual words shows lower scores than the BERT-context model, where sentences were used to obtain embeddings.

Second, the pretraining on FULL FrameNet improves results for all models except model with XLM-R embeddings. The model achieves the best improvement with BERT embeddings (+2.1%). The model with BERT-context embeddings obtains the best results in terms of recall (83.2%) and F-measure (79.0%) among models pre-trained on full FrameNet corpus. Pretraining on full FrameNet led to an increase in recall metrics for all models on 2.8-8.6%, while the precision metric reduced on 4.6-7.4%.

Third, for the REDUCED setup, the lower number of training examples from FrameNet improves the results for the model with BERT-context embeddings only (+0.8% of F-measure). The precision of the model with BERT-context embedding improves on 6%, while recall reduces on 4.8% compared to a model trained on full FrameNet corpus.

[4]https://github.com/IINemo/isanlp

Model	P	R	F
BERT, Multilingual (Larionov et al., 2019)	-	-	.757
RuBERT, Russian (Larionov et al., 2019)	-	-	.801
Without pretraining on FrameNet			
BERT	.820	.739	.766
BERT-context	**.828**	.746	.784
XLM-R	.815	.765	.783
MUSE	.818	.733	.772
LASER	.811	.720	.762
Pretrained on the Full FrameNet			
BERT	.768	.808	.787
BERT-context	.754	**.832**	.790
XLM-R	.759	.793	.773
MUSE	.772	.782	.777
LASER	.756	.774	.764
Pretrained on Reduced FrameNet			
BERT	.766	.805	.784
BERT-context	.814	.784	**.798**
XLM-R	.739	.793	.762
MUSE	.736	.784	.758
LASER	.747	.786	.764

Table 1: The model performance results.

To sum up, our results demonstrate that models, pretrained on the FULL version of FrameNet and fine-tuned on FrameBank, obtain higher recall and F-measure scores; from the other side, pretraining on English data for SRL decreases precision. The reducing number of FrameNet examples improves results for the model with BERT-context embeddings only.

5. Conclusion

We contribute to the transfer learning research by providing a first study on the effectiveness of exploiting English SRL data to boost Russian SRL performance. We study two setups for the source FrameBank dataset. Our experiments with several multilingual embeddings on the FrameBank dataset show that pretraining on the English FrameNet yield improvement for BERT-, LASER-, and MUSE-based models. Among four models, the model with BERT-based contextualized embeddings obtains the best macro-averaged F1-measure of 79.8%. We have demonstrated that it is beneficial to have the same set of roles in both corpora to order to boost the semantic role labeling performance.

We are currently working on the integration of FrameBank into the Linguistic Linked Open Data (LLOD) cloud (Cimiano et al., 2020; McCrae et al., 2016). According to our project, FrameBank will be interlinked with: **1)** the LLOD representation of FrameNet (Rospocher et al., 2019); **2)** other linguistic resources from the LLOD cloud, such as WordNet (McCrae et al., 2014), BabelNet (Ehrmann et al., 2014) and RuThes Cloud (Kirillovich et al., 2017; Galieva et al., 2017); and **3)** extralingual Linked Open Data resources, including DBpedia (Lehmann et al., 2015).

After that, we are going to retrain our model based on the newly obtained links. We hypothesize that these links can improve the accuracy of SRL against the baseline obtained in the presented paper.

Acknowledgements

The work was funded by Russian Science Foundation according to the research project no. 19-71-10056.

References

Arkhipov, M., Trofimova, M., Kuratov, Y., and Sorokin, A. (2019). Tuning multilingual transformers for named entity recognition on slavic languages. In Erjavec, T., Marcińczuk, M., Nakov, P., Piskorski, J., Pivovarova, L., Šnajder, J., Steinberger, J., and Yangarber, R., Eds., *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2019)*, pages 89–93. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3712.

Artetxe, M. and Schwenk, H. (2019). Margin-based parallel corpus mining with multilingual sentence embeddings. In Nakov, P. and Palmer, A., Eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 3197–3203. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1309.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL '98), Volume I*, pages 86–90. Université de Montréal. https://doi.org/10.3115/980845.980860.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. https://doi.org/10.1162/tacl_a_00051.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). Linguistic linked open data cloud. In *Linguistic Linked Data: Representation, Generation and Applications*, pages 29–41. Springer. https://doi.org/10.1007/978-3-030-30225-2_3.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537. https://dl.acm.org/doi/10.5555/1953048.2078186.

Conneau, A. and Lample, G. (2019). Cross-lingual language model pretraining. In Wallach, H. M., othersLarochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E. B., and Garnett, R., Eds., *Proceedings of the 33rd Conference on Advances in Neural Information Processing Systems (NIPS 2019)*, pages 7059–7069. Curran Associates, Inc. https://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., Eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019), Volume 1 (Long and Short Papers)*, pages 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Ehrmann, M., Cecconi, F., Vannella, D., Mccrae, J. P., Cimiano, P., and Navigli, R. (2014). Representing multilingual data as linked data: the case of babelnet 2.0. In Calzolari, N. et al., Eds., *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, pages 401–408. European Language Resources Association. https://www.aclweb.org/anthology/L14-1628/.

Galieva, A., Kirillovich, A., Khakimov, B., Loukachevitch, N., Nevzorova, O., and Suleymanov, D. (2017). Toward domain-specific russian-tatar thesaurus construction. In *Proceedings of the International Conference IMS-2017*, page 120–124. Association for Computing Machinery. https://doi.org/10.1145/3143699.3143716.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288. https://doi.org/10.1162/089120102760275983.

He, L., Lee, K., Lewis, M., and Zettlemoyer, L. (2017). Deep semantic role labeling: What works and what's next. In Barzilay, R. and Kan, M.-Y., Eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), Volume 1: Long Papers*, pages 473–483. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1044.

Kirillovich, A., Nevzorova, O., Gimadiev, E., and Loukachevitch, N. (2017). Ruthes cloud: Towards a multilevel linguistic linked open data resource for russian. In Różewski, P. and Lange, C., Eds., *Proceedings of the 8th International Conference on Knowledge Engineering and Semantic Web (KESW 2017)*, Communications in Computer and Information Science, vol. 786, pages 38–52. Springer. https://doi.org/10.1007/978-3-319-69548-8_4.

Kuratov, Y. and Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for russian language. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"*, pages 333–339. http://www.dialog-21.ru/media/4606/kuratovyplusarkhipovm-025.pdf.

Kuznetsov, I. (2013). Semantic role labeling system for russian language. In Joho, H. and Ignatov, D. I., Eds., *ECIR 2013–Doctoral Consortium*, pages 15–18.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2017). Unsupervised machine translation using monolingual corpora only. *arXiv:1711.00043*. http://arxiv.org/abs/1711.00043.

Larionov, D., Shelmanov, A., Chistova, E., and Smirnov, I. (2019). Semantic role labeling with pretrained language models for known and unknown predicates. In Angelova, G., Mitkov, R., Nikolova, I., and Temnikova, I., Eds., *Proceedings of Recent Advances of Natural Language Processing (RANLP 2019)*, pages 620–630. Incoma Ltd. https://doi.org/10.26615/978-954-452-056-4_073.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2015). Dbpedia – a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195. https://doi.org/10.3233/SW-140134.

Loukachevitch, N. and Rusnachenko, N. (2020). Sentiment frames for attitude extraction in russian. *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"*, pages 526–537.

Lyashevskaya, O. and Kashkin, E. (2015). Framebank: A database of russian lexical constructions. In Khachay, M. Y., Konstantinova, N., Panchenko, A., Ignatov, D., and Labunets, V. G., Eds., *Revised Selected Papers of the 4th International Conference on Analysis of Images, Social Networks and Texts (AIST 2015)*, Communications in Computer and Information Science, vol. 542, pages 350–360. Springer. https://doi.org/10.1007/978-3-319-26123-2_34.

Lyashevskaya, O. (2012). Dictionary of valencies meets corpus annotation: A case of russian framebank. In Fjeld, R. V. and Torjusen, J. M., Eds., *Proceedings of the 15th EURALEX International Congress (Euralex 2012)*, pages 1023–1030. University of Oslo. https://euralex.org/publications/dictionary-of-valencies-meets-corpus-annotation-a-case-of-russian-framebank/.

McCrae, J. P., Fellbaum, C., and Cimiano, P. (2014). Publishing and linking wordnet using lemon and rdf. In Chiarcos, C., McCrae, J. P., Osenova, P., and Vertan, C., Eds., *Proceedings of the 3rd Workshop on Linked Data in Linguistics (LDL-2014)*, pages 13–16. European Language Resources Association.

McCrae, J. P., Chiarcos, C., Bond, F., Cimiano, P., Declerck, T., de Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., Osenova, P., Pareja-Lora, A., and Pool, J. (2016). The open linguistics working group: Developing the linguistic linked open data cloud. In Calzolari, N. et al., Eds., *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 2435–2441. European Language Resources Association. https://www.aclweb.org/anthology/L16-1386.

Miftahutdinov, Z., Alimova, I., and Tutubalina, E. (2020). On biomedical named entity recognition: Experiments in interlingual transfer for clinical and social media texts. In Jose, J. M. et al., Eds., *Proceedings of the 42nd European Conference on Information Retrieval (ECIR 2020)*, Lecture Notes in Computer Science, vol. 12036, pages 281–288. Springer. https://doi.org/10.1007/978-3-030-45442-5_35.

Nivre, J., Boguslavsky, I., and Iomdin, L. (2008). Parsing the syntagrus treebank of russian. In Scott, D. and Uszkoreit, H., Eds., *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 641–648. https://dl.acm.org/doi/10.5555/1599081.1599162.

Okamura, T., Takeuchi, K., Ishihara, Y., Taguchi, M., Inada, Y., Iizuka, M., Abo, T., and Ueda, H. (2018). Improving japanese semantic-role-labeling performance with transfer learning as case for limited resources of tagged corpora on aggregated language. In Politzer-Ahles, S., Hsu, Y.-Y., Huang, C.-R., and Yao, Y., Eds., *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 2018)*, pages 503–512. Association for Computational Linguistics. https://www.aclweb.org/anthology/Y18-1058.

Palmer, M., Gildea, D., and Xue, N. (2010). *Semantic Role Labeling*. Morgan & Claypool.

Pradhan, S., Ward, W., Hacioglu, K., Martin, J. H., and Jurafsky, D. (2005). Semantic role labeling using different syntactic views. In Knight, K., Ng, H. T., and Oflazer, K., Eds., *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 581–588. Association for Computational Linguistics. https://doi.org/10.3115/1219840.1219912.

Rospocher, M., Corcoglioniti, F., and Palmero Aprosio, A. (2019). Premon: Lodifing linguistic predicate models. *Language Resources and Evaluation*, 53:499–524. https://doi.org/10.1007/s10579-018-9437-8.

Ruder, S., Søgaard, A., and Vulić, I. (2019). Unsupervised cross-lingual representation learning. In Nakov, P. and Palmer, A., Eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019): Tutorial Abstracts, July 28, 2019, Florence, Italy*, pages 31–38. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-4007.

Rusnachenko, N., Loukachevitch, N., and Tutubalina, E. (2019). Distant supervision for sentiment attitude extraction. In Mitkov, R. and Angelova, G., Eds., *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1022–1030. INCOMA Ltd. https://doi.org/10.26615/978-954-452-056-4_118.

Schneider, N. and Wooters, C. (2017). The nltk framenet api: Designing for discoverability with a rich linguistic resource. In Specia, L., Post, M., and Paul, M., Eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017): System Demonstrations*, pages 1–6. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-2001.

Schwenk, H. (2018). Filtering and mining parallel data in a joint multilingual space. In Gurevych, I. and Miyao, Y., Eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), Volume 2: Short Papers*, pages 228–234. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2037.

Segalovich, I. (2003). A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In Arabnia, H. R. and Kozerenko, E. B., Eds., *Proceedings of the 2003 International Conference on Machine Learning, Models, Technologies and Applications (MLMTA'03), June 23–26, 2003, Las Vegas, Nevada, USA*, pages 273–280. CSREA Press.

Shelmanov, A. and Devyatkin, D. (2017). Semantic role labeling with neural networks for texts in russian. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"*, volume 2, pages 245–256. http://www.dialog-21.ru/media/3945/shelmanovaodevyatkinda.pdf.

Solovyev, V. and Ivanov, V. (2016). Knowledge-driven event extraction in russian: corpus-based linguistic resources. *Computational intelligence and neuroscience*, 2016. https://doi.org/10.1155/2016/4183760.

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In Hajič, J. and Zeman, D., Eds., *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99. Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-3009.

Subburathinam, A., Lu, D., Ji, H., May, J., Chang, S.-F., Sil, A., and Voss, C. (2019). Cross-lingual structure transfer for relation and event extraction. In Inui, K., Jiang, J., Ng, V., and Wan, X., Eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pages 313–325. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1030.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., Eds., *Proceedings of the 31st Conference on Advances in Neural Information Processing Systems (NIPS 2017)*, pages 5998–6008. Curran Associates, Inc. http://papers.nips.cc/paper/7181-attention-is-all-you-need.

Description Logic Based Formal Representation of Adjectives

Maria Grits
University of Koblenz-Landau
Institute for Computational Visualistics
maria.gritz@yandex.ru

Abstract

The paper introduces a system of rules for description logic based formal representation of adjectives used in both attributive and predicative functions and involved in a variety of syntactic relations. The system was developed to convey the semantics of adjectives by virtue of concept and role constructors implemented in description logic formalisms known as SHOIN(D) and SROIQ(D). The system is intended to be integrated into a large-scale system of formalization rules devised for the development of description logic based definitions of domain terms. The proposed system of rules was tested and evaluated on two sets of syntactic units that contain attributive and predicative adjectives represented in English and Russian languages.

Keywords: formal representation of semantics, attributive adjective, predicative adjective, formal definition, description logics

1. Introduction

The application of description logics (DL) for formal representation of semantics conveyed by units of various lexical categories is motivated by the extensive implementation of ontologies for natural language processing within the framework of Semantic Web development initiative (Horrocks, 2008; Ding, 2010; Yu, 2014). In order to provide for Question Answering over Linked Data, each query has to obtain a formal representation of its semantics that could be mapped to a network of classes, properties, data values, and individuals that constitute an ontology as a knowledge base (Fazzinga and Lukasiewicz, 2010; Mehta et al. 2015). For this reason, ontology-based question answering systems, instantiated with ORAKEL (Cimiano et al., 2008), Pythia (Unger and Cimiano, 2011), and AMUSE (Hakimov et al., 2018), require a solid set of rules to formalize meanings of lexical units of a natural language query.

With the view to facilitating the process of formalization, software developers compile a comprehensive ontology and augment it with an extensible lexicon. Lexicon units have their meanings defined through mappings to units of the ontology, the mappings are provided by virtue of lexicon models, instantiated with LexInfo (Cimiano et al., 2011), LexOnto (Cimiano et al., 2007), and OntoLex-Lemon (Cimiano et al., 2016; McCrae et al., 2017). If lexicon models were enhanced with formal definitions of lexicon units' semantics, a lexicon unit's semantics could be specified by virtue of interrelated ontology units rather than through a single ontology unit (Gritz, 2018a). This accurate specification of lexicon units' meanings could enhance formal representations of the semantics of queries and, therefore, contribute to the development of ontology-based semantic search technology. Furthermore, units introduced within formal definitions are supposed to bridge gaps within class and property taxonomies of an underlying domain ontology.

Even though several sets of formalization rules were designed to convert dictionary-based definitions of lexicon units' semantics into description logic based formal definitions (DL-definitions) (Völker et al., 2007; Azevedo et al., 2014), a comprehensive system of formalization rules is still required to obtain DL-definitions on a regular basis in an automatic or a semi-automatic fashion. The

current research aims to contribute to the development of a system of formalization rules by devising a set of rules used for DL-based formal representation of semantics conveyed by attributive and predicative adjectives, the formal representations are intended to be applicable in the process of DL-definitions formation. For the purpose of formal representation, we implement the concept and role constructors that are applied in SHOIN(D) and SROIQ(D) description logics, which are compatible with Web Ontology Language (OWL) standards (Horrocks and Patel-Schneider, 2004; Horrocks et al., 2006).

The rest of the paper is divided into four sections. Section 2 provides a critical review of the current practice of DL-based adjective formal representation. Section 3 introduces the rules for the DL-based formal representation of adjectives and exemplifies their application. Section 4 summarizes the results of the rules testing, instantiates the successful implementation of the rules, and analyses the failures. Section 5 outlines the conclusions and objectives for future work.

2. Related research work

Traditionally an adjective is supposed to act as a modifier of a noun, which is syntactically related to the adjective, and to be used in a sentence either in the predicative or in the attributive function (Kennedy, 2012). Within the framework of formal semantics, adjectives undergo an entailment-based classification stemming from the assumption that an entity denoted by an adjective-noun compound might be independently referred to by one or both units of the compound. The entailment-based adjective typology, discussed by Kamp and Partee (1995), Bouillon and Viegas (1999), and McNally (2016), distinguishes three classes: intersective, subsective, and non-subsective adjectives. As the current formalization practice suggests, an adjective of any class should acquire a DL-based representation, with semantic and derivational properties of the adjective being considered, class and property taxonomies of an underlying ontology being utilized.

2.1. Intersective adjectives in description logic notation

An entity denoted by an adjective-noun compound containing an intersective adjective might be independently referred to by the adjective and by a modified noun: $\forall x(AN(x) \rightarrow A(x))$; $\forall x(AN(x) \rightarrow N(x))$. Within the framework of the current DL-based formalization practice, an intersective adjective is formalized through an existential restriction imposed on a certain property. Restriction specifying classes are nominated by the lexemes that are derivationally or semantically related to the intersective adjective.

For instance, Amoia and Gardent (2006) exploit an existential restriction to describe an adjective as a lexeme undertaking a theta role of a derivationally related verb: $Afloat \equiv \exists Theme^{-1}.Float$. McCrae et al. (2014) and Walter et al. (2017) define an intersective adjective by virtue of an existential restriction imposed via a class nominated by a derivationally associated noun: $Belgian \equiv \exists nationality.Belgium$. Ding et al. (2019) augmented this approach, using singleton sets to impose existential restrictions: $American \equiv \exists nationality.\{United\_States\}$, and implementing negation: $Alive \equiv \neg\exists deathDate.\top$. Gangemi et al. (2016) formalize an adjective-noun compound through an intersection of a modified noun represented class and a class described through an existential restriction imposed on the *hasQuality* property by virtue of an adjective nominated class: $CanadianSurgeon \equiv Surgeon \sqcap \exists hasQuality.Canadian$.

2.2. Subsective adjectives in description logic notation

Whenever an individual is denoted by an adjective-noun compound containing a subsective adjective, the individual can be independently referred to by a modified noun: $\forall x(AN(x) \rightarrow N(x))$. However, the adjective cannot unveil the class membership of the individual: $\forall x(AN(x) \nrightarrow A(x))$.

When it comes to subsective adjective formalization, one has to deal with the representation of concept inclusion and gradability. Amoia and Gardent (2006) define an adjective denoted class as a subclass of a class obtained by imposing an existential restriction on the object property *has_property*. The restriction is imposed through a class denoted by a derivationally related noun, with the latter class undergoing intersection with a class represented by an existential restriction applied to the datatype property *has_measure*: $Tall \sqsubset \exists has\_property.(Tallness \sqcap \exists has\_measure.Top)$. Pareti and Klein (2011) enhanced this approach, by introducing conceptually related nouns to define an existential restriction imposed on the *hasProperty* object property and clarifying the threshold values applied to the

hasMeasure datatype property: $Expensive \equiv \exists hasProperty.(Cost \sqcap \exists hasMeasure.(\geq, X))$. Gangemi et al. (2016) impose an adjective specified existential restriction on the property *hasIntensionalQuality* in order to outline an adjective-noun compound represented class as a subclass of a class labeled with a modified noun: $SkillfulSurgeon \equiv \exists hasIntensionalQuality.Skillful \sqsubseteq Surgeon$.

2.3. Non-subsective adjectives in description logic notation

Non-subsective adjectives are referred to as intensional modifiers since they are applied to modify the intension of a syntactically bound noun (McNally, 2016). Non-subsective adjectives are divided into two groups: ordinary non-subsective adjectives and privative adjectives (Morzycki, 2016). Ordinary non-subsective adjectives, instantiated by the adjectives: *alleged*, *probable*, and *potential*, cannot be used to define the class membership of an individual represented by an adjective-noun compound: $\forall x(AN(x) \nrightarrow A(x))$. Moreover, it is unfeasible to identify an individual with a class denoted by a modified noun: $\forall x(AN(x) \nrightarrow N(x))$. In order to provide a description logic based formal representation of an adjective-noun compound, Gangemi et al. (2016) use classes denoted by an ordinary non-subsective adjective and a modified noun to impose existential restrictions on *hasModality* and *associatedWith* properties accordingly: $AllegedThief \equiv \exists hasModality.Alleged \sqcap \exists associatedWith.Thief$.

Privative adjectives appear to negate the core semantic properties of a modified noun: $\forall x(AN(x) \rightarrow \neg N(x))$, and, simultaneously, extend a modified noun's intension so that the noun could denote a broader class (Partee, 2010). Following this conception, Gangemi et al. (2016) apply negation to convey the semantics of both privative adjectives and intersective adjectives with privative readings: $FakeSphinx \equiv \neg Sphinx \sqsubseteq Sphinx\_(broad)$, $StoneLion \equiv \exists hasQuality.Stone \backslash Lion \sqsubseteq Lion\_(broad)$. Alternatively, a privative adjective and a modified noun might be used to specify existential restrictions imposed on *hasIntensionalProperty* and *associatedWith* properties: $FakeSphinx \equiv \exists hasIntensionalQuality.Fake \sqcap \exists associatedWith.Sphinx$ (Gangemi et al., 2016).

The proposed approaches were developed in order to enhance question answering systems over knowledge bases (Ding et al., 2019), knowledge extraction tools (Gangemi et al., 2016), ontology revision applications (Pareti and Klein, 2011), and ontology lexicons (McCrae et al., 2014; Walter et al., 2017). The techniques for formal representation of adjectives were intended to harness certain properties and classes of existing ontologies. On the contrary, the proposed set of rules is intended to be applicable for the development of ontologies and associated lexicons from scratch.

3. A system of rules for description logic based formalization of attributive and predicative adjectives

Following the functional approach to the semantics of adjectives (Partee, 2010; Morzycki, 2016), we currently implement compositional type-theoretic semantics in order to obtain formal representations of adjectives and introduce the resulting formulas in a string of symbols composing a DL-based formal definition. Therefore, semantic values of syntactic nodes of a parsed natural language definition are represented functionally: within each branch, sister nodes are correlated as a function and its argument, with a corresponding parent node representing the value of the applied function (see Figure 1). The parent node is implemented further either as a function or as an argument of a function within the process of functional application that carries on until the root node of the definition is reached (Chierchia and McConnell-Ginet, 2000; Winter, 2016).

Under the assumption that a DL-definition retains its truth-value in the whole scope of possible worlds (Gritz, 2018b), we represent the process of functional application, using the semantic types produced by a combination of the elementary types *e* and *t*: *e* stands for an entity on a domain and typically characterizes a proper noun, *t* stands for a truth-value and typically characterizes a syntactically well-formed definition as a declarative sentence (Chierchia and McConnell-Ginet, 2000).

In order to obtain a DL-definition, three type combinations are utilized: <e, t>, <<e, t>, t>, and <<e, t><e, t>>. The type <e, t> assigns the role of the characteristic function to a syntactic node that denotes a concept *C*, which is related to an individual by virtue of a concept assertion: *C:a*. The type <<e, t>, t> delivers a semantic value of a phrase introducing a defined term and performing the subject role within

the main clause of a definition. This type reserves an argument position to be filled in by a concept denoted by the verb phrase occupying the sister node. Finally, the type <<e, t><e, t>> is used to mark all other functions that should return the denotation of a parent node, applying a concept denoted by a sister node as an argument. This type is implemented to characterize the semantics of the copular verb *to be*, bearing the identity function, and an article used as a determiner. A function of the type <<e, t><e, t>> expressed by definite and indefinite articles is not explicated within formalization examples in the current paper in order to make formal descriptions more concise.

The implementation of these semantic types is instantiated through the process of formalization of a natural language definition of the term *Agentive: An agentive is an agent* (the formal definitions are represented in first-order logic and in description logic notations).

$$NP(x) = \lambda x.Agent(x)/< e,t >$$
$$V(x) = \lambda NP.NP(x)/\ll e,t >< e,t \gg$$
$$VP(x) = \lambda x.Agent(x)/< e,t >$$
$$NP_{subj}(x) = \lambda VP.\forall x(Agentive(x) \leftrightarrow VP(x))/ << e,t > t >$$
$$\forall x(Agentive(x) \leftrightarrow Agent(x))/t$$

$$NP = Agent/< e,t >$$
$$V = \lambda NP.NP/\ll e,t >< e,t \gg$$
$$VP = Agent/< e,t >$$
$$NP_{subj} = \lambda VP.Agentive \equiv VP/<< e,t > t >$$
$$Agentive \equiv Agent/t$$

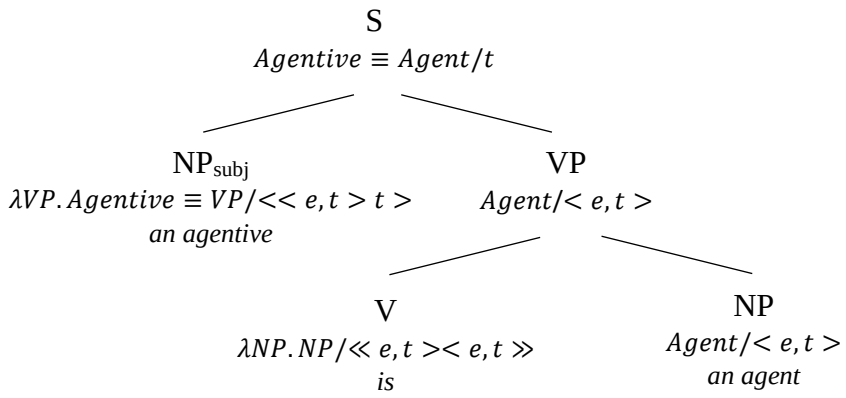


Figure 1: An analysis of the syntactic structure and semantics conveyed by a natural language definition of the term *Agentive*

In the current example, the lexical meaning of a term is defined by virtue of a synonym. We have proposed a system of rules to obtain DL-based formal representations of attributive and predicative adjectives that might be used to describe the lexical meaning of a term within the right part of a definition (see Table 1). An adjective acquires either the semantic type <<e, t><e, t>>, bearing the role of a function with an argument conveyed by virtue of a sister node, or the type <e, t>, performing the role of a function's argument (see Table 2 for examples).

The DL-based concept descriptions obtained by virtue of the rules acquire interpretations within the framework of model-theoretic semantics (Baader et al., 2007). Model-theoretic semantics is applied as a referential theory of meaning that studies meaning as a relation of symbols to objects. In other words, the meaning of a resulting concept description is determined by attaching an interpretation function *I* from possible worlds to subsets on a domain: $C^I: W \rightarrow D$ (Fitting, 2015). If the interpretation function returns a non-empty subset, the concept description is considered to be satisfiable. The resulting DL-based concept descriptions are checked for satisfiability on a domain in order to evaluate the rules proposed for DL-based formal representation of adjectives (see Table 1).

Rules for formal representation of adjectives	Interpretations for satisfiability check of the resulting concept descriptions
$Adj^{attr}_{\ll e,t><e,t\gg} = \lambda NP_{<e,t>}.A \sqcap NP$ (1)	$\|\|NP'\|\| = \{x \in \Delta^I \| x \in A^I \wedge x \in \|\|NP\|\|\}$
$Adj^{attr}_{\ll e,t><e,t\gg} = \lambda NP_{<e,t>}.\exists A^{-}.\top \sqsubseteq NP$ (2)	$\|\|NP'\|\| = \{x \in \Delta^I \| \exists x.(y,x) \in A^I \rightarrow x \in \|\|NP\|\|\}$
$Adj^{attr}_{\ll e,t><e,t\gg} = \lambda NP_{<e,t>}.\exists A.NP$ (3)	$\|\|NP'\|\| = \{x \in \Delta^I \| \exists y.(x,y) \in A^I \wedge y \in \|\|NP\|\|\}$

$Adj^{pred}_{<e,t>} = A$ (4)	$\|VP\| = \{x \in \Delta^I \| x \in A^I\}$/ $\|VP\| = \{x \in \Delta^I \| \exists y.(x,y) \in V^I \wedge y \in A^I\}$
$Adj^{pred}_{<<e,t><e,t>>} = \lambda NP^{extsubj}_{<e,t>}.NP^{extsubj} \sqsubseteq A$ (5)	$\|AdjP\| = \{x \in \Delta^I \| x \in \|NP^{extsubj}\| \rightarrow x \in A^I\}$
$Adj^{pred}_{<<e,t><e,t>>} = \lambda NP^{obl/adj}_{<e,t>}.\exists A\_P.NP^{obl/adj}$ (6)	$\|AdjP\|$ $= \{x \in \Delta^I \| \exists y.(x,y) \in A\_P^I \wedge y \in \|NP^{obl/adj}\|\}$
$Adj^{pred}_{<<e,t><e,t>>} = \lambda VP^{comp/adj}_{<e,t>}.A \sqcap VP^{comp/adj}$ (7)	$\|AdjP\| = \{x \in \Delta^I \| x \in A^I \wedge x \in \|VP^{comp/adj}\|\}$
$Adj^{pred}_{<<e,t><e,t>>} = \lambda VP^{comp/adj}_{<e,t>}.A \sqsubseteq VP^{comp/adj}$ (8)	$\|AdjP\| = \{x \in \Delta^I \| x \in A^I \longrightarrow x \in \|VP^{comp/adj}\|\}$
$Adj^{pred}_{<<e,t><e,t>>} = \lambda VP^{comp/adj}_{<e,t>}.A \sqsubseteq \neg VP^{comp/adj}$ (9)	$\|AdjP\| = \{x \in \Delta^I \| x \in A^I \longrightarrow x \notin \|VP^{comp/adj}\|\}$

Table 1: The rules for description logic based formal representation of adjectives.

Evaluation of the rules

Within the process of the development of the rules, the adjective type heterogeneity hypothesis (Morzycki, 2016) was assumed in order to differentiate property-denoting adjectives (type <e, t>) from adjectives functioning as predicate modifiers (type <<e, t>, <e, t>>). However, a semantic type of an adjective is supposed to be determined by its syntactic relations rather than by its entailment-based category (see Section 2).

In the current research, we maintain the discrimination of intersective, subsective, and non-subsective adjectives used in the attributive function. Since a modified noun phrase and an attributive intersective adjective might be entailed to denote the class membership of an individual represented by an adjective-noun compound, we use the intersection constructor to represent the compound's semantics formally: $BlueFence \equiv Blue \sqcap Fence$. Hence, the resulting compound is supposed to denote an intersection of two sets on a domain: $\|Blue\ Fence\| = \{x \in \Delta^I \| x \in Blue^I \wedge x \in Fence^I\}$ (see Rule 1 and Example 1).

In order to retain the analytic truth of a DL-definition despite the use of subsective and non-subsective adjectives in the attributive function, we impose specific existential restrictions on adjective nominated roles so as to provide a formal account for a scope of domain entities that hold for: $\lambda x.AN(x)$. We assume that a subsective adjective represents a binary relation with a set of observers as its domain and a set of observed objects as its range. We use the inverse role constructor, impose an existential restriction on the resulting role, and introduce an inclusion axiom within a DL-definition of an adjective-noun compound: $ProfessionalArtist \equiv \exists professional^{-}.\top \sqsubseteq Artist$. We define a subset denoted by the adjective-noun compound within a set of entities denoted by a modified noun phrase: $\|Professional\ Artist\| = \{x \in \Delta^I \| \exists x.(y,x) \in professional^I \rightarrow x \in Artist^I\}$ (see Rule 2 and Example 2).

In order to represent the meaning of a noun phrase including a non-subsective adjective used as a modifier, we impose an existential restriction on an adjective nominated role by virtue of a concept represented by a modified noun phrase (see Rule 3 and Example 3). Rule 3 is supposed to be applicable to ordinary non-subsective adjectives: $AllegedCriminal \equiv \exists alleged.Criminal$, and privative adjectives in the attributive function: $FakePearl \equiv \exists fake.Pearl$, since the resulting formulae do not imply unintended entailments and successfully convey the adjectival modification of a noun phrase's intension.

Whereas in the previous examples attributive adjectives were viewed as units incorporated into noun phrases, predicative adjectives are represented as units of verb phrases. Rule 4 provides a formal account of a predicative adjective that is related formally to a modified noun phrase in one of the two ways: as an atomic concept standing in an intersection with a noun phrase nominated concept (see Example 4.1); as a concept imposing an existential restriction on a predicate nominated role to deliver a concept standing in an intersection with a noun phrase nominated concept (see Example 4.2). Example 4.1 illustrates the case when a predicative adjective is bound by the copula verb *to be* (see Table 2). For Rules 4, 5, 7, 8, 9 to give a proper formal account for predicative uses of subsective and privative adjectives (e.g. *Tanja is skillful, The document is fake*), the corresponding general concepts are presumed to denote domains of adjective nominated roles: $\exists skillful^{-}.\top \sqsubseteq Skillful$, $\exists fake.\top \sqsubseteq Fake$.

Rule 5 was designed to formalize a predicative adjective related within an open predicative complement (or an open predicative adjunct) to an external subject expressed with a noun phrase that is used as an object of the main clause predicate. The predicative adjective is formalized as a concept

related through inclusion to a concept represented by the external subject used to specify an existential restriction imposed on a role nominated by the main clause predicate: $\|paints\ a\ fence\ blue\| = \{x \in \Delta^I | \exists y.\ (x,y) \in paints^I \wedge y \in Fence^I \rightarrow y \in Blue^I\}$ (see Example 5). Rule 6 represents a predicative adjective related to a noun phrase used as an oblique argument or an adjunct. The adjective denotes a role with an existential restriction being imposed on it by virtue of a concept represented by the noun phrase: $\|busy\ at\ work\| = \{x \in \Delta^I | \exists y.\ (x,y) \in busy\_at^I \wedge y \in Work^I\}$ (Example 6). The noun phrase is supposed to be introduced within a prepositional phrase; the head of the prepositional phrase undergoes concatenation with the adjective in the process of formalization (see Rule 6).

Rules 7, 8, and 9 were designed for formal representation of a predicative adjective attaching a clausal complement with an omitted subject, an open predicative complement, or an open predicative adjunct. The choice between the rules depends either on the grammatical form of a verb used as a predicate within the complement/adjunct or on the semantics of the predicative adjective. A concept expressed by the predicative adjective is supposed to undergo intersection with a concept represented by a related verb phrase: $\|busy\ packing\ a\ bag\| = \{x \in \Delta^I | x \in Busy^I \wedge x \in \|pack\ a\ bag\|\}$, whenever a predicate of the complement/adjunct is delivered by virtue of a participle (see Rule 7). In contrast, a concept expressed by the predicative adjective is supposed to denote a concept subsumed by a verb phrase introduced concept or by the negation of that concept: $\|busy\ to\ pack\ a\ bag\| = \{x \in \Delta^I | x \in Busy^I \rightarrow x \notin \|pack\ a\ bag\|\}$ (see Rules 8 and 9), whenever a predicate of the complement/adjunct is delivered by means of an infinitive. The negation of a concept expressed by the verb phrase is used in case the predicative adjective bears privative semantics: i.e. the adjective indicates the fact that the agent characterized by the adjective (see Example 9) does not perform the action denoted by the related verb phrase.

A worker paints a **blue** fence (1)	A man invites a **professional** artist (2)
$NP = Fence/< e,t >$ $\boldsymbol{Adj = \lambda NP.\ Blue \sqcap NP/\ll e,t >< e,t \gg}$ $NP_{obj} = Blue \sqcap Fence\ /< e,t >$ $V = \lambda NP_{obj}.\exists paints.\ NP_{obj}/\ll e,t >< e,t \gg$ $VP = \exists paints.\ (Blue \sqcap Fence)/< e,t >$ $NP_{subj} = \lambda VP.Worker \sqcap VP/<< e,t > t >$ $Worker \sqcap \exists paints.\ (Blue \sqcap Fence)/t$	$NP = Artist/< e,t >$ $\boldsymbol{Adj = \lambda NP.\exists professional^{-}.\top \sqsubseteq NP/\ll e,t > < e,t \gg}$ $NP_{obj} = \exists professional^{-}.\top \sqsubseteq Artist\ /< e,t >$ $V = \lambda NP_{obj}.\exists invites.\ NP_{obj}/\ll e,t >< e,t \gg$ $VP = \exists invites.\ (\exists professional^{-}.\top \sqsubseteq Artist)\ / < e,t >$ $NP_{subj} = \lambda VP.Man \sqcap VP/<< e,t > t >$ $Man \sqcap \exists invites.\ (\exists professional^{-}.\top \sqsubseteq Artist)/t$
A policeman arrests an **alleged** criminal (3)	A child is **happy** (4.1)
$NP = Criminal/< e,t >$ $\boldsymbol{Adj = \lambda NP.\exists alleged.\ NP\ /\ll e,t >< e,t \gg}$ $NP_{obj} = \exists alleged.\ Criminal\ /< e,t >$ $V = \lambda NP_{obj}.\exists arrests.\ NP_{obj}/\ll e,t >< e,t \gg$ $VP = \exists arrests.\ \exists alleged.\ Criminal/< e,t >$ $NP_{subj} = \lambda VP.Policeman \sqcap VP/<< e,t > t >$ $Policeman \sqcap \exists arrests.\ \exists alleged.\ Criminal/t$	$\boldsymbol{Adj = Happy/< e,t >}$ $V = \lambda Adj.\ Adj/\ll e,t >< e,t \gg$ $VP = Happy/< e,t >$ $NP = \lambda VP.Child \sqcap VP/<< e,t > t >$ $Child \sqcap Happy/t$
A child feels **happy** (4.2)	A worker paints a fence **blue** (5)
$\boldsymbol{Adj = Happy/< e,t >}$ $V = \lambda Adj.\exists feels.\ Adj/\ll e,t >< e,t \gg$ $VP = \exists feels.\ Happy/< e,t >$ $NP = \lambda VP.\ Child \sqcap VP/<< e,t > t >$ $Child \sqcap \exists feels.\ Happy/t$	$NP_{obj} = Fence/< e,t >$ $\boldsymbol{Adj = \lambda NP_{obj}.\ NP_{obj} \sqsubseteq Blue/\ll e,t >< e,t \gg}$ $AdjP = Fence \sqsubseteq Blue/< e,t >$ $V = \lambda AdjP.\exists paints.\ AdjP/\ll e,t >< e,t \gg$ $VP = \exists paints.\ (Fence \sqsubseteq Blue)/< e,t >$ $NP_{subj} = \lambda VP.Worker \sqcap VP/<< e,t > t >$ $Worker \sqcap \exists paints.\ (Fence \sqsubseteq Blue)/t$
A woman is **busy** at work (6)	A woman is **busy** packing a bag (7)
$NP_{obl/adj} = Work/< e,t >$ $\boldsymbol{Adj = \lambda NP_{obl/adj}.\exists busy\_at.\ NP_{obl/adj}/ \ll e,t >< e,t \gg}$ $AdjP = \exists busy\_at.\ Work/< e,t >$ $V = \lambda AdjP.\ AdjP/\ll e,t >< e,t \gg$	$NP_{obj} = Bag/< e,t >$ $V = \lambda NP_{obj}.\exists pack.\ NP_{obj}/\ll e,t >< e,t \gg$ $VP_{comp/adj} = \exists pack.\ Bag/< e,t >$ $\boldsymbol{Adj = \lambda VP_{comp/adj}.\ Busy \sqcap VP_{comp/adj}/ \ll e,t >< e,t \gg}$

$VP = \exists busy\_at.Work/< e,t >$ $NP_{subj} = \lambda VP.Woman \sqcap VP/<< e,t > t >$ $Woman \sqcap \exists busy\_at.Work/t$	$AdjP = Busy \sqcap \exists pack.Bag/< e,t >$ $V = \lambda AdjP.AdjP/\ll e,t >< e,t \gg$ $VP = Busy \sqcap \exists pack.Bag/< e,t >$ $NP_{subj} = \lambda VP.Woman \sqcap VP/<< e,t > t >$ $Woman \sqcap Busy \sqcap \exists pack.Bag/t$
A woman feels **eager** to pack a bag (8)	A woman feels **reluctant** to pack a bag (9)
$NP_{obj} = Bag/< e,t >$ $V = \lambda NP_{obj}.\exists pack.NP_{obj}/\ll e,t >< e,t \gg$ $VP_{comp/adj} = \exists pack.Bag/< e,t >$ $\boldsymbol{Adj = \lambda VP_{comp/adj}.Eager \sqsubseteq VP_{comp/adj}/}$ $\boldsymbol{\ll e,t >< e,t \gg}$ $AdjP = Eager \sqsubseteq \exists pack.Bag/< e,t >$ $V = \lambda AdjP.\exists feels.AdjP/\ll e,t >< e,t \gg$ $VP = \exists feels.(Eager \sqsubseteq \exists\ pack.Bag)/< e,t >$ $NP_{subj} = \lambda VP.Woman \sqcap VP/<< e,t > t >$ $Woman \sqcap \exists feels.(Eager \sqsubseteq \exists\ pack.Bag)/t$	$NP_{obj} = Bag/< e,t >$ $V = \lambda NP_{obj}.\exists pack.NP_{obj}/\ll e,t >< e,t \gg$ $VP_{comp/adj} = \exists pack.Bag/< e,t >$ $\boldsymbol{Adj = \lambda VP_{comp/adj}.Reluctant \sqsubseteq \neg VP_{comp/adj}/}$ $\boldsymbol{\ll e,t >< e,t \gg}$ $AdjP = Reluctant \sqsubseteq \neg\exists\ pack.Bag/< e,t >$ $V = \lambda AdjP.\exists feels.AdjP/\ll e,t >< e,t \gg$ $VP = \exists feels.(Reluctant \sqsubseteq \neg\exists\ pack.Bag)/< e,t >$ $NP_{subj} = \lambda VP.Woman \sqcap VP/<< e,t > t >$ $Woman \sqcap \exists feels.(Reluctant \sqsubseteq \neg\exists\ pack.Bag)/t$

Table 2: Examples of application of the rules for DL-based formal representation of adjectives

4. Implementation of the system of rules for description logic based formalization of attributive and predicative adjectives

In order to test the proposed system of rules, 400 syntactic units were extracted and formalized: 200 English syntactic units were retrieved from a Dictionary of Linguistics and Phonetics (Crystal, 2008) and the British National Corpus[1], 200 Russian syntactic units were derived from a Dictionary of Linguistic Terms (Akhmanova, 2012) and the Russian National Corpus[2]. Attributive and predicative adjectives were equally represented in both languages. The precision of formalization rules for attributive adjectives equals 96,5%, the rules yield satisfiable formal expressions for predicative adjectives with the precision of 93,5%.

Rules 1 and 2 are successfully applied to provide formal representations of intersective and subsective adjectives in both Russian and English languages (we implement the automatic transliteration in accordance with the ISO-9 standard[3]): $Stilističeskaâ \sqcap Kategoriâ$ (*stilističeskaâ kategoriâ*, "a stylistic category"), $\exists major^{-}.\top \sqsubseteq Component$ (major components). Nevertheless, whenever a subsective or an intersective adjective yields a binary relation between entities of a set denoted by a modified noun phrase, Rules 1 and 2 fail to obtain satisfiable concept descriptions. For instance, Rule 1 returns an intersection of two concepts: $Divergent \sqcap Form$, whereas the phrase *divergent forms* should be formalized as: $Form \sqcap \exists divergent.Form$. Rule 2 represents a subsective adjective by virtue of a role that binds a set of observers with a set of observed entities. For this reason, the concept description: $\exists alternative^{-}.\top \sqsubseteq Grammar$, is an invalid representation of the compound *alternative grammars*, and the expression: $\exists vzaimosvâzannoe^{-}.\top \sqsubseteq Izmenenie$, does not yield the semantics of the following compound: взаимосвязанные изменения (*vzaimosvâzannye izmeneniâ*, "interrelated changes"). On the contrary, Rule 3 returns only valid expressions for both languages: $\exists fallacious.Argument$ (fallacious argument), $\exists vozmožnyj.Predšestvennik$ (*vozmožnyj predšestvennik*, "a possible predecessor").

Rule 4 has proved to be efficacious, being applied flawlessly to formalize Russian and English sentences. Rule 4 is utilized when predicative adjectives are related to modified nouns via the copular verb *to be*, used in finite and infinite forms: $Grammar \sqcap Adequate$ (Grammars are adequate), $Značenie \sqcap Proizvodnoe$ (*značenie, âvlâûŝeesâ proizvodnym*, "a meaning being derived"). Rule 4 is also applicable in cases a predicative adjective is related to the main predicate by virtue of the copular verb *to be*: Grammar $\sqcap \exists said\_to\_be.Adequate$ (Grammars are said to be adequate) or a preposition, which is subjected to concatenation: $Predmet \sqcap \exists myslitsâ\_kak.Izvestnyj$ (*Predmet myslitsâ kak izvestnyj*, "An entity is regarded as familiar"). The negation constructor is inserted to represent a

[1] https://www.english-corpora.org/bnc/
[2] http://www.ruscorpora.ru/new/
[3] https://www.translitteration.com/transliteration/en/russian/iso-9/

predicative adjective whenever a negative particle is used: $\check{C}lenPredlo\check{z}enii\hat{a} \sqcap \neg Glavnyj$ (*členy predloženiâ, ne âvlâûŝiesâ glavnymi*, "members of a sentence not being main"), or a negative prefix is utilized: $Word \sqcap \neg Variable$ (The words are invariable).

Rule 5 is flawlessly implemented for both Russian and English syntax: $Technology \sqcap \exists make.((Rural \sqcap Life) \sqsubseteq Feasible)$ (The technologies make rural life feasible), $Govor\hat{a}\hat{s}ij \sqcap \exists s\check{c}itaet.(Mysl' \sqsubseteq \check{C}u\check{z}a\hat{a})$ (*Govorâŝij sčitaet mysl' čužoj*, "The speaker considers the idea to be borrowed"). Rule 5 alike Rule 4 is appropriate to use for predicative adjectives bound by means of the copular verb *to be*: $Theory \sqcap \exists assumes.(Assertion \sqsubseteq True)$ (The theory assumes the assertion to be true), or by virtue of a preposition, both units are omitted in the process of formalization: $Vid \sqcap \exists Predstavl\hat{a}et.(Dejstvie \sqsubseteq Posto\hat{a}nnoe)$ (*vid, predstavlâûŝij dejstvie kak postoânnoe*, "the aspect representing an action as constant"). Rule 6 returns valid formal representations of predicative adjectives that attach noun phrases as oblique arguments or adjuncts in English and in Russian languages: $Form \sqcap \exists present\_in.Language$ (The form is present in the language), $Dviženie \sqcap \exists neobhodimoe\_dl\hat{a}.\exists Proiznesenie.Zvuk$ (*dviženiâ neobhodimye dlâ proizneseniâ zvukov*, "the movements necessary for the articulation").

Rule 7 is applicable to yield a valid representation of a predicative adjective that attaches a clausal complement with an omitted subject, an open predicative complement, or an open predicative adjunct, a predicate included in the attached verb phrase being expressed through a participle: $Woman \sqcap \exists fell.Silent \sqcap \exists stare\_into.Night$ (A woman fell silent staring into the night), $Pisatel' \sqcap Prav \sqcap Točen \sqcap \exists opisat'.\exists harakter.Geroin\hat{a}$ (*Pisatel' prav i točen, opisyvaâ harakter geroini*, "The writer is right and exact describing the character of the heroine"). However, whenever a predicative adjective and a predicate incorporated into an attached verb phrase imply different agents as their arguments, Rule 7 produces an unsatisfiable expression: $Digital \sqcap Signal \sqcap Possible \sqcap \exists use.(Available \sqcap Technology)$ (Digital signals are possible using available technology). Rule 7 also fails in case a predicative adjective binds a verb phrase incorporated in a prepositional phrase: $\exists source\_of.Energy \sqcap Capable\_of \sqcap \exists being\_used\_in.\exists production.SpeechSound$ (A source of energy is capable of being used in speech sound production), since the predicative adjective is intended to render a binary relation on a domain.

Rules 8 and 9 replace Rule 7 in case a predicate of an attached complement/adjunct is expressed by virtue of an infinitive, and these rules have also proved to be efficient for both languages: $Separation \sqcap Slow \sqsubseteq \neg\exists produce.(Audible \sqcap Friction)$ (The separation is slow to produce audible friction), $Slovo \sqcap Vyraženie \sqcap Sposobnoe \sqsubseteq \exists vystupat'\_v.(Sintaksi\check{c}eska\hat{a} \sqcap Funkci\hat{a})$ (*slova i vyraženiâ, sposobnye vystupat' v sintaksičeskoj funkcii*, "words and expressions able to perform a syntactic function"). Nevertheless, both rules produce invalid expressions whenever a predicative adjective and a predicate of an attached complement/adjunct imply different agents as their arguments: $Affricate \sqcap Easy \sqsubseteq \exists define.\top$ (Affricates are easy to define), $Sentence \sqcap Problematic \sqsubseteq \neg\exists analyse.\top$ (Sentences are problematic to analyze).

In case a predicative adjective and a predicate of an attached complement/adjunct are related to the same external subject, but the subject (whether overt or omitted) refers to a set of events: Мужчина находит занятным подделывать банковские чеки (*Mužčina nahodit zanâtnym poddelyvat' bankovskie čeki, "The man finds it enjoyable to fake bank cheques")*, the proposed system of rules yields a concept description that fails to receive an adequate interpretation on a domain: $Mužčina \sqcap \exists nahodit.(Zanâtnyj \sqsubseteq \exists poddelyvat'.(Bankovskij \sqcap \check{C}ek))$. An adjective phrase formal representation, instantiated by: $Zanâtnyj \sqsubseteq \exists poddelyvat'.(Bankovskij \sqcap \check{C}ek)$, acquires the type <e, t>, yet there is no such entity on a domain that could be characterized as enjoyable and produce fake bank cheques at the same time. As far as Rule 9 is concerned, a double negation exemplified by an invalid formal expression: $Triangle \sqcap \neg Able \sqsubseteq \neg(\neg 3have.Side)$ (A triangle is unable not to have three sides), which implies an affirmative false statement: $Triangle \sqcap \neg Able \sqsubseteq 3have.Side$, allows us to deduce that Rule 9 is also inapplicable in case a predicate of an attached complement/adjunct binds a negative particle.

5. Conclusion

As a result of the conducted research, a comprehensive set of rules for description logic based formal representation of attributive and predicative adjectives was devised in order to contribute to Question Answering over Linked Data and to improve the technologies for ontology lexicon modeling. The system was developed as an integral part of the formalization technology intended to provide DL-based definitions of domain terms. The system was designed to be implemented in a semi-automatic fashion: syntactic features and semantic characteristics of adjectives and related syntactic units should be specified manually so that high rates of precision could be achieved.

In the current research, the emphasis was put on the development of rules for both predicative and attributive adjectives. A scope of efficient rules for DL-based formal representation of predicative adjectives involved in a variety of syntactic relations was proposed. The implementation of existential restrictions on roles for the representation of attributive adjectives' semantics resulted in a novel technique for subsective and non-subsective adjectives formalization, with a description logic being used as a first-order formalism. The set of rules allows flexibility in formalization, representing adjectives' semantics through the implementation of both concept and role constructors.

The proposed set of formalization rules is supposed to deliver DL-based concept descriptions that should be incorporated in a rigid structure of a DL-definition that is essentially a chain of concept intersections. The DL-definitions yield fairly complicated concept descriptions designed to provide accurate delimitations of subsets denoted by defined terms on a domain. The proposed set of rules delivers a limited selection of DL-based concept descriptions. The concept descriptions are intended to provide satisfiable descriptions of subsets denoted by defined terms, rather than to provide accurate formal representations of various syntactic structures applied in natural language definitions. Therefore, the devised system of rules has to be augmented with the formalization solutions introduced for the syntactic structures that failed to acquire valid formal representations by virtue of the proposed system.

References

Akhmanova, O. S. (2012). *A Dictionary of Linguistic Terms*. Sixth edition. Moscow: URSS.

Amoia, M. and Gardent, C. (2006). Adjective Based Inference. In *Proceedings of the Workshop on Knowledge and Reasoning for Language Processing (KRAQ'06)*. Trento, Italy. Published by Association for Computational Linguistics, USA, 20–27.

Azevedo, R., Freitas, F., Rocha, R., Menezes, J., Rodrigues, C., and Gomes, M. (2014). Representing Knowledge in DL ALC from Text. *Procedia Computer Science*, (35):176–185.

Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., and Patel-Schneider, P. F., Eds. (2007). *The Description Logic Handbook: Theory, Implementation, and Applications*. Second edition. Cambridge, UK: Cambridge University Press.

Bouillon, P. and Viegas, E. (1999). The Description of Adjectives for Natural Language Processing: Theoretical and Applied Perspectives. In *Proceedings of the TALN'99 Workshop on Description des Adjectifs pour les Traitements Informatiques. Traitement Automatique des Langues Naturelles*. Cargèse, France, July 12 – 17, 1999. 20–30.

Chierchia, G. and McConnell-Ginet, S. (2000). *Meaning and Grammar: An Introduction to Semantics*. Second edition. Cambridge, MA: MIT Press.

Cimiano, P., McCrae, J. P., and Buitelaar, P., Eds. (2016). *Lexicon Model for Ontologies: Community Report.* Final Community Group Report. https://www.w3.org/2016/05/ontolex.

Cimiano, P., Buitelaar, P., McCrae, J., and Sintek, M. (2011). LexInfo: A Declarative Model for the Lexicon-Ontology Interface. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, 9(1):29–51.

Cimiano, P., Haase, P., Heizmann, J., Mantel, M., and Studer, R. (2008). Towards Portable Natural Language Interfaces to Knowledge Bases — the Case of the ORAKEL System. *Data & Knowledge Engineering*, 65(2):325–354.

Cimiano, P., Haase, P., Herold, M., Mantel, M., and Buitelaar, P. (2007). LexOnto: A Model for Ontology Lexicons for Ontology-based NLP. In Buitelaar, P., Choi, K. S., Gangemi, A., and Huang, C. R., Eds., *Proceedings of the OntoLex07 Workshop held in conjunction with the 6th International Semantic Web Conference (ISWC'07)*. Busan, South Korea.

Crystal, D. (2008). *A Dictionary of Linguistics and Phonetics.* Sixth edition. Oxford: Blackwell.

Ding, J., Hu, W., Xu, Q., and Qu, Y. (2019). Mapping Factoid Adjective Constraints to Existential Restrictions over Knowledge Bases. In Ghidini, C. et al., Eds., *Proceedings of the 18th International Semantic Web Conference (The Semantic Web – ISWC 2019), Part 1*. Auckland, New Zealand, October 26–30, 2019. 164–181.

Ding, Y. (2010). Semantic Web: Who is Who in the Field – A Bibliometric Analysis. *Journal of Information Science*, 36(3):335–356.

Fazzinga, B. and Lukasiewicz, T. (2010). Semantic Search on the Web. *Semantic Web*, 1(1, 2):89–96.

Fitting, M. (2015). *Intensional Logic*. In Zalta, E. N., Ed., the Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/sum2015/entries/logic-intensional.

Gangemi, A., Nuzzolese, A. G., Presutti, V., and Recupero, D. R. (2016). Adjective Semantics in Open Knowledge Extraction. In Ferrario, R. and Kuhn, W., Eds., *Formal Ontology in Information Systems (Frontiers in Artificial Intelligence and Applications)*, Vol. 283, 167–180.

Gritz, M. (2018). Lexical Meaning Formal Representations Enhancing Lexicons and Associated Ontologies. In Basile, P., Basile, V., Croce, D., Dell'Orletta, F., and Guerini, M., Eds., *Proceedings of the 2nd Workshop on Natural Language for Artificial Intelligence (NL4AI 2018) co-located with 17th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2018)*. Trento, Italy, November 22 – 23, 2018. CEUR Workshop Proceedings, Vol. 2244, 102–115. http://ceur-ws.org/Vol-2244/.

Gritz, M. (2018). Towards Lexical Meaning Formal Representation by virtue of the NL-DL Definition Transformation Method. In *Proceedings of the 3rd International Conference on Computational Linguistics in Bulgaria*. Sofia, Bulgaria, May 28 – 29, 2018. Institute for Bulgarian Language, Bulgarian Academy of Sciences, Sofia, Bulgaria, 23–33.

Hakimov, S., Jebbara, S., and Cimiano, P. (2018). *AMUSE: Multilingual Semantic Parsing for Question Answering over Linked Data*. https://arxiv.org/pdf/1802.09296.pdf.

Horrocks, I. (2008). Ontologies and the Semantic Web. *Communications of the ACM*, 51(12):58–67.

Horrocks, I. and Patel-Schneider, P. F. (2004). Reducing OWL Entailment to Description Logic Satisfiability. *Journal of Web Semantics*, 1(4):345–357.

Horrocks, I., Kutz, O., and Sattler, U. (2006). The Even More Irresistible *SROIQ*. In Doherty, P., Mylopoulos, J., and Welty, C. A., Eds., *Proceedings of the 10th International Conference on Principles of Knowledge Representation and Reasoning (KR 2006)*. Lake District, UK, June 2 – 5, 2006. AAAI Press, 57–67.

Kamp, H. and Partee, B. (1995). Prototype Theory and Compositionality. *Cognition*, (57):129–191.

Kennedy, C. (2012). Adjectives. In Russell, G. and Fara, D. G., Eds., *Routledge Companion to Philosophy of Language*. New York: Routledge, 328–341.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In *Proceedings of eLex 2017 Conference*. Leiden, the Netherlands, September 19 – 21, 2017. 19–21.

McCrae, J. P., Unger, C., Quattri, F., and Cimiano, P. (2014). Modelling the Semantics of Adjectives in the Ontology-Lexicon Interface. In Zock, M., Rapp, R., and Huang, C. R., Eds., *Proceedings of 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*. Dublin, Ireland. 198–209.

McNally, L. (2016). Modification. In Aloni, M. and Dekker, P., Eds., *Cambridge Handbook of Semantics*. Cambridge, UK: Cambridge University Press, 442–467.

Mehta, A., Zatakia, S., and Deulkar, K. (2015). Comparative Study of Web Search Methods Using Ontology. *International Journal of Computer Science and Information Technologies*, 6(6):5497–5499.

Morzycki, M. (2016). *Modification (Key Topics in Semantics and Pragmatics)*. Cambridge, UK: Cambridge University Press.

Pareti, P. and Klein, E. (2011). Learning Vague Concepts for the Semantic Web. In Novacek, V., Huang, Z., and Groza, T., Eds., *Proceedings of the Joint Workshop on Knowledge Evolution and Ontology Dynamics in conjunction with the 10th International Semantic Web Conference*. Bonn, Germany, October 24, 2011. CEUR Workshop Proceedings, Vol. 784. http://ceur-ws.org/Vol-784/.

Partee, B. (2010). Privative Adjectives: Subsective Plus Coercion. In Bäuerle, R., Reyle, U., and Zimmermann, T. E., Eds., *Presuppositions and Discourse: Essays Offered to Hans Kamp*. Amsterdam: Elsevier, 273–285.

Unger, C. and Cimiano, P. (2011). Pythia: Compositional Meaning Construction for Ontology-based Question Answering on the Semantic Web. In Muñoz, R., Montoyo, A., and Métais, E., Eds., *Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems (NLDB)*. Alicante, Spain, June 28 – 30, 2011. Heidelberg: Springer, 153–160.

Völker, J., Hitzler, P., and Cimiano, P. (2007). Acquisition of OWL DL Axioms from Lexical Resources. In Franconi, E., Kifer, M., and May, W., Eds., *Proceedings of the 4th European Semantic Web Conference (ESWC'07)*. Innsbruck, Austria, June 3 – 7, 2007. Springer, 670–685.

Walter, S., Unger, C., and Cimiano, P. (2017). Automatic Acquisition of Adjective Lexicalizations of Restriction Classes: a Machine Learning Approach. *Journal on Data Semantics*, 6(3):113–123.

Winter, Y. (2016). *Elements of Formal Semantics. An Introduction to the Mathematical Theory of Meaning in Natural Language (Edinburgh Advanced Textbooks in Linguistics)*. Edinburgh: Edinburgh University Press.

Yu, L. (2014). *A Developer's Guide to the Semantic Web.* Second edition. Heidelberg: Springer.

Linguistic vs. encyclopedic knowledge. Classification of MWEs on the base of domain information

Zara Kancheva
IICT-BAS
zara@bultreebank.org

Ivaylo Radev
IICT-BAS
radev@bultreebank.org

Abstract

This paper reports on the first steps in the creation of linked data through the mapping of BTB-WordNet and the Bulgarian Wikipedia. The task of expanding the BTB-WordNet with encyclopedic knowledge is done by mapping its synsets to Wikipedia pages with many MWEs found in the articles and subjected to further analysis. We look for a way to filter the Wikipedia MWEs in the effort of selecting the ones most beneficial to the enrichment of BTB-WN.

Keywords: MWEs; Wordnet; Wikipedia.

1. Introduction

The state of the field shows that language resources used alone do not perform well in each and every NLP task. In recent years researchers started to align various lexical resources in projects such as BabelNet (Navigli and Ponzetto, 2012), SemLink (Palmer, 2009), Predicate Matrix (de Lacalle et al., 2014) and Uby (Gurevych et al., 2012). Building relations between linguistic and semantic resources and using this kind of new data to generate knowledge graphs has its benefits for languages with less lexical resources.

With the development of huge electronic corpora and advancements in corpus linguistics MultiWord Expressions (MWEs) receive more and more attention from researchers. A paper (Sag et al., 2002) estimates that the number of MWEs in the lexicon of a person is more than 40%. MWEs are omnipresent in all text data and can not be skipped in tasks such as word sense disambiguation, named entity linking and coreference resolution.

Some tendencies in contemporary linguistics have changed drastically and the observation of (Kiefer, 1988) that "theoretical linguists and lexicography seem each to go their own ways, they do not seem to show much interest in other's preoccupations" are no longer factual. There are several projects that aim to integrate linguistic and encyclopedic knowledge, most commonly by the merge of a dictionary or WordNet with Wikipedia or Wiktionary.

One of the main challenges occurring from the integration of the two types of knowledge - of the linguistic system and of the world (Kecskes, 2013) - is related with the question how much and what types of encyclopedic information is useful to add to our language resource? A lot of work is already done on the mapping of dictionaries and WordNets with Wikis, but it is interesting and challenging to focus on the MWE distribution in the resulting dataset.

In the process of manual mapping with CLaRK system (Simov et al., 2004) between Bulgarian WordNet (BTB-WN) and Wikipedia we plan to introduce to BTB-WN all MWEs related to the mapped Wikipedia articles. Usually these MWEs are with a head word corresponding to the title of the Wikipedia article — for example, 'Wine' vs. 'Red wine', 'White wine', 'Sparkling wine', etc. From a linguistic perspective this determines relation head-dependent. From semantic point of view the relations are more diverse. In this mainly we determine sub-concepts, but by different features.

The structure of the paper is as follows: the next section outlines the related work. Section 3 explores different domains of encyclopedic knowledge in Wikipedia. Section 4 concludes the paper.

2. Related Work

Among the most outstanding works on the alignment of linguistic and encyclopedic knowledge with WordNets are: BabelNet - combining multilingual WordNet and Wikipedia; Uby - combining WordNet, GermaNet, Wiktionary, Wikipedia, FrameNet and VerbNet for English and German; the mapping of the Princeton WordNet with the English Wikipedia (McCrae, 2018); and the mapping of the plWordNet onto the Princeton WordNet (Rudnicka et al., 2017).

For the purposes of this research we work with the BTB-WN (Osenova and Simov, 2018), which was build in several steps. It started as an translation of Core WordNet and was expanded with concepts from Bulgarian Treebank (BulTreeBank (Osenova et al., 2012)), frequency list and currently Bulgarian Wikipedia. At the moment BTB-WN contains about 25 000 synsets - the last 15 percent of them came from the expansion with around 13 000 articles from Bulgarian Wikipedia in attempt to map it to the BTB-WN (Simov et al., 2019).

Currently the Wikipedia in Bulgarian has 259 927 content pages, which makes it about 23 times smaller than the English version, but it is still a very useful resource to extract world knowledge from. It contains data for concepts (similarly to WordNet) and instances of concepts - notable Named Entities (NEs) for persons, locations and events (often excluded in WordNet). Building knowledge graphs upon the relations between concepts and their instances and using these graphs to train, test and improve NLP systems is deemed to be very impactful in positive manner. Being a communal free to use and edit resource, Wikipedia is constantly expanding with new articles and reflects the creation of new inventions and products or the emergence of new celebrities and events.

Recent paper (Laskova et al., 2019) presents an overview of MWEs in BTB-WN, where the MWEs are presented as several types of phrases by their head-word: multiword Nouns (Noun+Noun (N); Adj+N; Numeral+N); Verbs (Verb+N; Verb+Adv; Verb+PP); Adjectives (Adv+Adj; Adj+PP) and Adverbials (Prep+N; Prep+Adj; Adv+Adv) in accordance with the classification developed within WG 4 of PARSEME COST Action[1]; and treated afterward with a catena-based modeling. Working with the same resource we will use the same classification method in our work.

A similar approach in dealing with MWEs is presented in (Koeva et al., 2016). The paper reports on classification of MWEs based on morphosyntactic, structural and semantic criteria and using semi-automatic methods to compile a MWE dictionary for Bulgarian. The work discusses a repository of 86373 'nominal' and 'verbal' MWEs, based on the head word.

MWEs could be defined as "lexical units larger than a word that can bear both idiomatic and compositional meanings" (Masini, 2005). (Sprenger, 2003) uses different term for the same linguistic phenomena - fixed expressions - and describes them as "combinations of two or more words that are typically used to express a specific concept. (...) these combinations are stored in the mental lexicon of native speakers and as a whole refer to a (linguistic) concept".

There is no single generally accepted typology of the MWE, different researchers classify them at several levels - morphology, lexicology, syntax and semantics. One of most detailed classifications is that of (Sag et al., 2002). It does not take into consideration the head-word type of the MWEs like the approach of (Laskova et al., 2019), (Koeva et al., 2016), it divides MWEs in two general types - lexicalised and institutionalized phrases. The first group is for phrases that have at least partially idiosyncratic syntax or semantics, or contain 'words' which do not occur in isolation and it has three subtypes:

- Fixed expressions - fully lexicalized expressions, that do not undergo morphosyntactic variation and internal modification (for example *in short*, *ad hoc*).
- Semi-fixed expressions - these expressions undergo some degree of lexical variation and are further divided in three types:
  - Non-decomposable Idioms — the only type of lexical variation observable in this group is inflection (*kick the bucket*) and reflexive form (*wet oneself*).

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

  - Compound Nominals — these phrases inflect for number (*car parks, parts of speech*).
  - Proper Names — the phrases in this group are syntactically highly idiosyncratic (*San Francisco 49ers, Oakland Raiders*), so they require different approach for analysis, depending on their instances.

- Syntactically-flexible expressions - this subtype exhibits a much wider range of syntactic variability than the semi-fixed expressions and are divided by the types of variations possible:
  - Verb-particle Constructions - these constructions consist of a verb and one or more particles (*write up, look up*).
  - Decomposable Idioms - phrases of this subtype (for example *let the cat out of the bag, sweep under the rug*) are very challenging for analysis, because they are syntactically variable to varying degrees.
  - Light Verbs - these constructions contain a noun used in a normal sense and a verb with bleached, rather than idiomatic meaning (*make a mistake, give a demo*).

The second type of MWEs in this classification is institutionalized phrases and it contains conventionalized phrases that are semantically and syntactically compositional, but statistically idiosyncratic (*traffic light, fresh air*).

Another approach on the differentiation of MWEs, that is not intended as a classification, but could give an interesting perspective on the subject is given in (Hüning and Schlücker, 2015), where twelve groups are outlined:

- Proverbs (*a bird in the hand is worth two in the bush*, quotations (*shaken, not stirred*) and commonplaces (*one never knows*).
- Metaphorical Expressions (*as sure as eggs is eggs*).
- Verbal Idioms (*to kick the bucket*).
- Particle/Phrasal Verbs (*to make up*).
- Light Verb Constructions/Composite Predicates (*to have a look*).
- Syntactic/Quasi Noun Incorporation (German *Auto waschen* 'to wash car').
- Stereotyped Comparisons/Similes (*as nice as pie*).
- Binomial Expressions (*shoulder to shoulder*).
- Complex Nominals (*man about town*).
- Collocations (*strong tea*).
- Fossilized/Frozen Forms (*all of a sudden*).
- Routine Formulas (*Good morning*).

3. Domain Specific MWEs

For the aims of this research we have semi-automatically extracted 13173 MWEs from 14512 Wikipedia pages. These pages contained over 30 000 links to other pages in Wikipedia from which manually we selected "true" MWEs, excluding person names, annual events (13th/14th Summer Olympics), titles of movies, books, music albums and songs. The initial set of 14512 Wikipedia pages were selected on the basis of mappings between Bulgarian Wikipedia and BTB-WN (Laskova et al., 2019). So far 1628 MWEs were preliminary added as synsets in the BTB-WordNet without being domain classified; 2187 MWEs have been domain classified in preparation to be added as synsets and 9358 MWEs are ongoing

the process of domain classification. An effort to use Wikipedia categories to match the WMEs to the respective domains was made, but they were too noisy and unstructured. Many domains are represented in Wikipedia, but here we outline the most prominent ones divided in six conditional groups: Physics and astronomy, Chemistry, Geography, Biology and medicine, Social, Other.

Domain	**No MWEs**
Chemistry	30
Physics and Astronomy	89
Biology and Medicine	130
Geography	801
Social	960
Other	177
Total	2187

Table 1: Classified MWEs from Wikipedia

As already mentioned, we will apply the MWE classification of (Laskova et al., 2019). All of the extracted MWEs are nouns and most of them are type of Adj+N; smaller part of them are NN and Numeral+N. MWEs in these domains can be divided in two groups Named Entities (NEs) and terms / terminology / concepts. Our main concern are the terms. We also include NEs of global scope such as the event of WW2 (and large scale operations as D-Day) or Summer Olympics as a sports forum (but not its iterations).

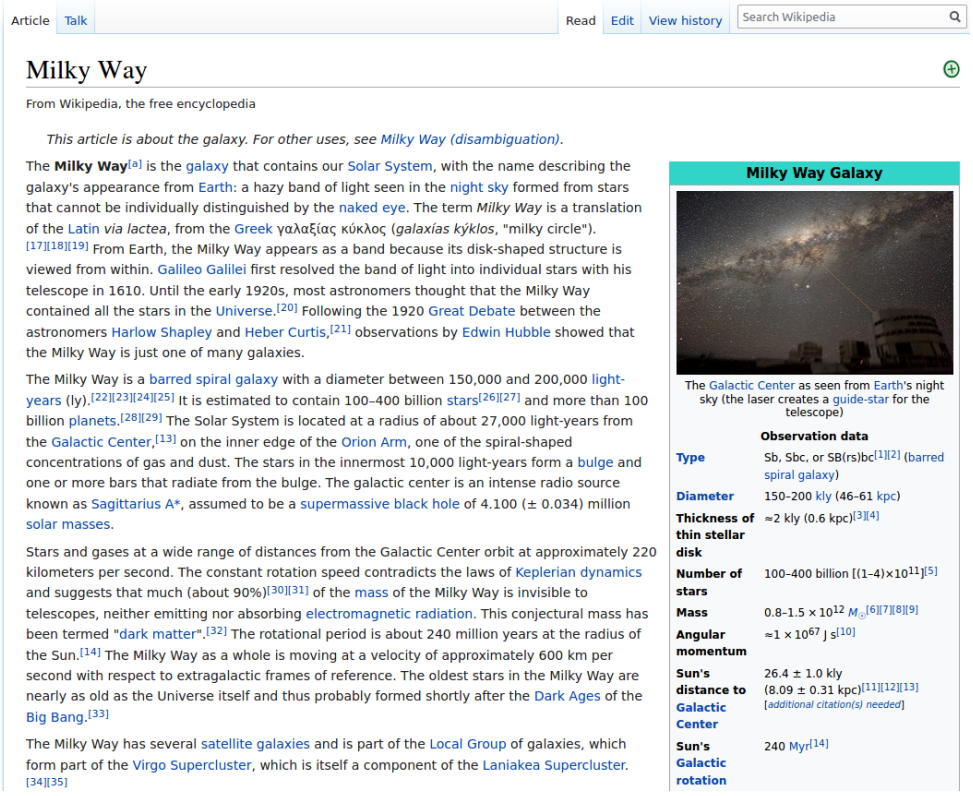
It is important to outline (though, it was somehow predictable) that there are no proverbs, metaphorical expressions and verbal constructions among the extracted MWEs, because of the characteristics of the Wikipedia content, which most frequently concerns entities and events, constructed by nouns and adjectives. Typically Wikipedia contains articles about famous geographical objects and terminology of different fields of science. MWEs that are proper names and terms may not be of the greatest interest for linguists, but they are valuable for our current research. Various NLP tasks need both linguistic and encyclopedic knowledge, thus enriching BTB-WN with as much as possible synsets will be beneficial for our work. Also this type of data can be used in further modelling of MWEs.

After the manual determination of MWEs, we have automatically divided them by their category in Wikipedia, which helps with the domain typology to a certain extent, but is definitely tricky. The categories in Wikipedia are thousands and tend to specify, rather than to generalise the topics of the content, so they would form a very detailed and hard to apply classification of MWEs. Additionally, every article could and very often does belong to more than one category (for example the article for *Ammonium nitrate* appears in three categories - *Ammonium compounds, Nitrates, Explosive chemicals* and does not directly point to the more general category *Chemistry*. The intention of the research at this stage is to focus exactly on common domains and less on their specific subclasses (for now), so we will classify the MWEs on the one hand by the science branch that they belong to, and on the other hand - by their linguistic features.

3.1. Physics and Astronomy Domain

Wikipedia contains many MWEs (for example Fig 1) which are NE to astral objects like asteroids, comets and planets of the type 3 Juno and 81P/Wild, that follow taxonomic patterns and are considered infinite. Such MWEs are interesting but of little importance to the expansion of the BTB-WN. They are typically constructed of noun and number: asteroids and spaceflight programs tend to contain the name of an Ancient greek or roman gods or mythological characters (6 Хеба (*6 Heba*, “6 Hebe”). Scientific laws, theories, principles very often include the name(s) of its inventor and thus they are constructions of noun, preposition and surname - Уравнения на Максуел (*Uravneniya na Maksuel*, “Maxwell’s equations”), Принцип на Паули (*Printsip na Pauli*, “Pauli exclusion principle”).

Other MWEs that are much more valuable are the terms for physical phenomena and units of measure, because they do not contain numbers and proper nouns. Examples for this type of MWEs in the do-



Article Talk Read Edit View history Search Wikipedia

Milky Way

From Wikipedia, the free encyclopedia

This article is about the galaxy. For other uses, see Milky Way (disambiguation).

The **Milky Way**[a] is the galaxy that contains our Solar System, with the name describing the galaxy's appearance from Earth: a hazy band of light seen in the night sky formed from stars that cannot be individually distinguished by the naked eye. The term *Milky Way* is a translation of the Latin *via lactea*, from the Greek γαλαξίας κύκλος (*galaxías kýklos*, "milky circle").[17][18][19] From Earth, the Milky Way appears as a band because its disk-shaped structure is viewed from within. Galileo Galilei first resolved the band of light into individual stars with his telescope in 1610. Until the early 1920s, most astronomers thought that the Milky Way contained all the stars in the Universe.[20] Following the 1920 Great Debate between the astronomers Harlow Shapley and Heber Curtis,[21] observations by Edwin Hubble showed that the Milky Way is just one of many galaxies.

The Milky Way is a barred spiral galaxy with a diameter between 150,000 and 200,000 light-years (ly).[22][23][24][25] It is estimated to contain 100–400 billion stars[26][27] and more than 100 billion planets.[28][29] The Solar System is located at a radius of about 27,000 light-years from the Galactic Center,[13] on the inner edge of the Orion Arm, one of the spiral-shaped concentrations of gas and dust. The stars in the innermost 10,000 light-years form a bulge and one or more bars that radiate from the bulge. The galactic center is an intense radio source known as Sagittarius A*, assumed to be a supermassive black hole of 4.100 (± 0.034) million solar masses.

Stars and gases at a wide range of distances from the Galactic Center orbit at approximately 220 kilometers per second. The constant rotation speed contradicts the laws of Keplerian dynamics and suggests that much (about 90%)[30][31] of the mass of the Milky Way is invisible to telescopes, neither emitting nor absorbing electromagnetic radiation. This conjectural mass has been termed "dark matter".[32] The rotational period is about 240 million years at the radius of the Sun.[14] The Milky Way as a whole is moving at a velocity of approximately 600 km per second with respect to extragalactic frames of reference. The oldest stars in the Milky Way are nearly as old as the Universe itself and thus probably formed shortly after the Dark Ages of the Big Bang.[33]

The Milky Way has several satellite galaxies and is part of the Local Group of galaxies, which form part of the Virgo Supercluster, which is itself a component of the Laniakea Supercluster.[34][35]

Milky Way Galaxy

The Galactic Center as seen from Earth's night sky (the laser creates a guide-star for the telescope)

Observation data

Type	Sb, Sbc, or SB(rs)bc[1][2] (barred spiral galaxy)
Diameter	150–200 kly (46–61 kpc)
Thickness of thin stellar disk	≈2 kly (0.6 kpc)[3][4]
Number of stars	100–400 billion [(1–4)×10^{11}][5]
Mass	0.8–1.5 × 10^{12} $M_{\odot}$[6][7][8][9]
Angular momentum	≈1 × 10^{67} J s[10]
Sun's distance to Galactic Center	26.4 ± 1.0 kly (8.09 ± 0.31 kpc)[11][12][13] *[additional citation(s) needed]*
Sun's Galactic rotation	240 Myr[14]

Figure 1: Wikipedia page with astronomy MWE

main are: слънчев вятър (*slanchev vyatar*, "solar wind"), магнитно поле (*magnitno pole*, "magnetic field"), горен/долен/странен кварк (*goren/dolen/stranen kvark*, "up/down/strange quark"). There were observed two types of measure units - Adj+N (конска сила (*konska sila*, "horsepower") and N+Prep+N (километър в час (*kilometar v chas*, "kilometre per hour").

Because of the constant new findings and inventions of the modern science this domain is one of the most productive in Wikipedia, so it could be considered as a regular source for MWEs extraction.

3.2. Chemistry Domain

This is the domain with the fewest amount of MWEs - only 30 and they are maybe the most homogeneous group. Most of the MWEs here are chemical compounds, which are traditionally built of Adj+N (for example глюкуронова киселина (*glyukuronova kiselina*, "glucuronic acid"), but there are also other types of terms with the same structure such as периодична система (*periodichna sistema*, "periodic table") and ковалентна връзка (*kovalentna vrazka*, "covalent bond"). We also have concepts like селитра (*selitra*, "saltpeter") and its sub-types (hyponyms): натриев нитрат (*natriev nitrat*, "sodium nitrate"); амониев нитрат (*amoniev nitrat*, "ammonium nitrate") ; калиев нитрат (*kaliev nitrat*, "potassium nitrate").

The most complex in lexical structure MWEs in this domain are the terms that contain preposition and proper name like Принцип на Льо Шателие-Браун (*Printsip na Lyo Shatelie-Braun*, "Le Chatelier's principle") and Процес на Фишер-Тропш (*Protses na Fisher-Tropsh*, "Fischer–Tropsch process").

3.3. Geography Domain

This is the second largest domain and contains Wikipedia pages mostly for NEs, that may be put in two groups: locations (LOC) such as mountains, deserts, lowlands, bodies of water, islands, archipelagos, capes, hemispheres, etc. (for example южен полюс (*yuzhen polyus*, "South pole") and geopolitical locations (LOC-GPE) as countries, regions, departments, provinces, cities, kingdoms, counties, colonies

(for example Лос Анджелис (*Los Angelis*, “Los Angeles”); Обединеното кралство (*Obedinenoto kralstvo*, “The United Kingdom”).

There are several instances of peninsula with NEs - Скандинавски полуостров (*Skandinavski poluostrov*, “Scandinavian Peninsula”); Корейски полуостров (*Koreiski poluostrov*, “Korean Peninsula”); Баха Калифорния (*Baha California*, “Baja California”).

Lots of NEs that are settlements in Bulgaria will be added as instances of the synsets for village, town, city.

In this domain we also include climate zones and types (умерен климат (*umeren klimat*, “temperate climate”) and natural phenomena and disasters such as storms, volcanic eruptions, etc. (Ел Ниньо (*El Ninyo*, “El Niño”), Вранчанско земетресение (*Vrachansko zemetresenie*, “Vrancea earthquake”).

The geography domain is also very rich in terms that are not named entities: тектонска плоча (*tektonska plocha*, “tectonic plate”), морско равнище (*morsko ravnishte*, “sea level”).

3.4. Biology and Medicine Domain

The MWEs in these domains most frequently are constructed of two components. Here there are names of species of animals, plants and mushrooms: червена лисица (*chervena lisica*, “Red Fox”); бял бряст (*bial briast*, “european white elm”), бяла мухоморка (*byala muhomorka*, “destroying angel”); of body organs or diseases - костен мозък (*kosten mozak*, “bone marrow”); бели кръвни тела (*beli kravni tela*, “white blood cells”), метаболитен синдром (*metaboliten sindrom*, “metabolic syndrome”); branches or subfields of biology and medicine - молекулярна генетика (*molekulyarna genetika*, “molecular genetics”), ветеринарна медицина (*veterinarna meditsina*, “veterinary medicine”); and other types of domain specific terms - застрашен вид (*zastrashen vid*, “endangered species”), вечнозелено растение (*vechnozeleno rastenie*, “evergreen plant”).

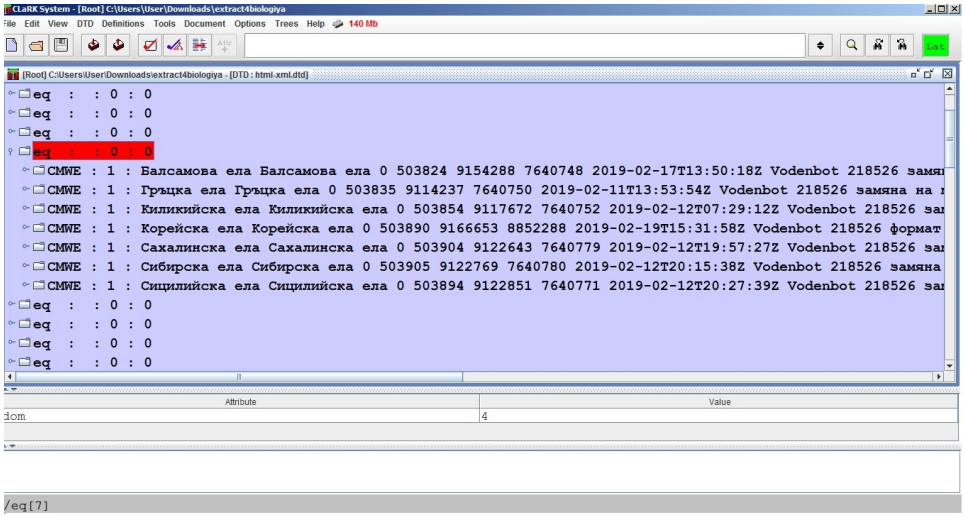


Figure 2: The Wikipedia category “Firs” with articles for the different species in CLaRK system[2]

Some exceptions from the two component structure are: the terms for some disorders that include the name of their discoverer (as it is with the principles in the physics domain) such as Синдром на Турет (*Sindrom na Turet*, “Tourette syndrome”); subtypes of disease like рак на дебелото черво (*rak na debeloto chervo*, “Colorectal cancer”); terms like оцеляване на най-приспособения (*otselyavane na nai-prisposobeniya*, “survival of the fittest”).

3.5. Social Domain

This is the largest and most prominent domain that contains concepts related to society and humans, thus making it the most heterogeneous. Here these types of MWEs can be found: sport events and teams;

[2]http://bultreebank.org/en/clark/

wars, battles and crisis; pacts, contracts and unions; armies and legions; languages and linguistic terms; famous buildings; the parts of the Bible; holidays; art styles; institutions and organizations.

The longest MWEs in this domain are the different types of institutions and organizations (of course not all of them are so complex - Върховен съд (*Varhoven sad*, "Supreme court") such as Българска народна македоно-одринска революционна организация (*Bulgarska narodna makedono-odrinska revolyutsionna organizatsiya*, "Bulgarian people's Macedonian-Adrianople revolutionary organization"), Координационен комитет за контрол на износа (*Koordinatsionen komitet za kontrol na iznosa*, "Coordinating committee for multilateral export controls").

Some Wikipedia pages contain information about hyponyms of a concept like: президентска република (*prezidentska republika*, "presidential republic") and парламентарна република (*parlamentarna republika*, "parliamentary republic") as sub-types (hyponyms) of република (*republika*, "republic").

As observed in (Sag et al., 2002) sports team names usually contain a place or organization name (for example Бостън Селтикс (*Bostan Seltiks*, "Boston Celtics"). The case with sports competitions and different types of festivals is similar - Токио 2020 (*Tokio 2020*, "Tokyo 2020"); Филмов фестивал в Кан (*Filmov festival v Kan*, "Cannes Film Festival"). Events of wars and battles are built of at least two lexical elements and usually denominate the place or time/duration of their occurance - Първа световна война (*Parva svetovna voina*, "World War I"); Битка при Вердюн (*Bitka pri Verdyun*, "Battle of Verdun"). Some of these concepts are annual and similar to the productivity of NEs in the astronomy domain and are skipped.

There are many MWEs for organizations in different fields - I Германски легион (*Parvi germanski legion*, "1st Germanic Legion") and occupations - министър на отбраната (*ministar na otbranata*, "minister of defence"). Languages and language families are always MWEs (бретонски език (*bretonski ezik*, "Breton language"), тюркски езици (*tyurkski ezitsi*, "Turkic languages") Many holidays and currencies appear in this group too - Рождество Христово (*Rozhdestvo Hristovo*, "Feast of the Nativity"), суринамски долар (*surinamski dolar*, "Surinamese dollar").

3.6. "Other" Domain

This group contains heterogeneous MWEs, that can not be placed in the previous categories and are too little to be in separate groups. Most of them could be generalized as artefacts - there are products, inventions, man-made entities. A quite big part of this group consists of nautical and aviation terminology - types of ships, ship elements, military aircraft are very well presented in the Bulgarian Wikipedia. Here the MWEs always have two components - adjective and noun - like батарейна палуба (*batareina palouba*, ("gun deck"), бойна рубка (*boina rubka* , ("conning tower"), минен трал (*minen tral*, ("mine roller"). There are exceptions like the names of fighter aircraft and bombers, that usually contain a proper name and numbers (Messerschmitt Bf 109, Avia B-135, Albatros C.III).

Another big group is formed by types of weapons and ammunition, which are also frequently constructed of Adj+N in Bulgarian (in English they usually are compound nouns) for example гладкоцевно оръжие *gladkotsevno orazhie*, "smoothbore"), but could be more complex in some cases - ръчен противотанков гранатомет (*rachen protivotaknov granatomet*, "rocket-propelled grenade"), междуконтинентална балистична ракета (*mezhdukontinentalna balistichna raketa*, "intercontinental ballistic missile"). The tendencies in the MWEs for vehicles, machines and their components, musical instruments are quite the same like the before mentioned groups - асинхронен двигател (*asinhronen dvigatel*, "asynchronous motor"), бронирана кола (*bronirana kola*, "armoured car"), бас китара (*bas kitara*, "bass guitar"). Rarely constructions with preposition could be observed - автомобил с повишена проходимост (*avtomobil s povishena prohodimost*, "sport utility vehicle"), but we can see more everyday artefacts like: вятърна мелница (*viaturna melnica*, ("wind mill"); спален чувал (*spalen choval*, ("sleeping bag"); автомобилна гума (*avtomobilna guma*, ("automobile tyre")

Different types of food and drinks are included in the Other domain and they do not diverge in type and number of lexical elements from the already mentioned MWEs in this domain. Typical examples are червено вино (*cherveno vino*, "red wine"), пейл ейл (*peil eil*, "pale ale"), бяло саламурено сирене

(*byalo salamureno sirene*, "white brine cheese").

Another group is related to mathematics and IT: аксиоматичен метод (*aksiomatichen metod*, "axiomatic system"); закон за големите числа (*zakon za golemite chisla*, "law of large numbers"); уеб дизайн (*uoeb dizain*, "web design"); син екран на смъртта (*sin ekran na smurtta*, "blue screen of death").

4. Conclusion

Aligning lexical resources like WordNet with encyclopedic knowledge from Wikipedia has proven very beneficial in the NLP field. This is even more true about relatively small sized resource that is BTB-WN. It has already been expanded once by 15% with general concepts from Wikipedia and now we are working on MWEs specialized expansion with another 15%, but this may be an underestimate.

Although doing this kind of work manually is very time consuming our experience shows that in the case of Bulgarian Wikipedia attempting to do this kind of domain classification automatically using only the data (in the form of its categories and hierarchy) from Wikipedia is not beneficial enough.

It is possible to use automatic methods in the future to produce synsets for NEs related to Bulgaria. For example all of the PERs and LOC-GPEs in Bulgarian Wikipedia can be added with the instance-of relation to the respective type of occupation/profession and settlement (village, town and city).

In regard to the domain distribution of the extracted MWEs it could be summarized that the fields of social sciences, sport and art and the geography domain are the most numerous.

Acknowledgements

This work was partially supported by the *Bulgarian Ministry of Education and Science under the National Research Programme "Young scientists and postdoctoral students" approved by DCM # 577 / 17.08.2018* and by *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant number DO01-272/16.12.2019.

References

de Lacalle, M. L., Laparra, E., and Rigau, G. (2014). Predicate matrix: extending semlink through wordnet mappings. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 903–909, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). Uby - a large-scale unified lexical-semantic resource based on lmf. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590, Avignon, France, April. Association for Computational Linguistics.

Hüning, M. and Schlücker, B. (2015). *Multi-word expressions*, volume 1, pages 450–467. De Gruyter Mouton, January.

Kecskes, I. (2013). *Encyclopedic Knowledge, Cultural Models, and Interculturality*, pages 81–104. OUP USA, December.

Kiefer, F. (1988). Linguistic, conceptual and encyclopedic knowledge: Some implications for lexicography. In Magay, T. and Zigány, J., Eds., *Proceedings of the 3rd EURALEX International Congress*, pages 1–10, Budapest, Hungary, September. Akadémiai Kiadó.

Koeva, S., Stoyanova, I., Todorova, M., and Leseva, S. (2016). Semi-automatic compilation of the dictionary of bulgarian multiwordexpressions. *Proceedings of GLOBALEX 2016: Lexicographic Resources for Human Language Technology, Workshop at LREC2016, Portorož, Slovenia*, May.

Laskova, L., Osenova, P., Simov, K., Radev, I., and Kancheva, Z. (2019). Modeling MWEs in BTB-WN. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 70–78, Florence, Italy, August. Association for Computational Linguistics.

Masini, F. (2005). Multi-word expressions between syntax and the lexicon: The case of italian verb-particle constructions. *SKY Journal of Linguistics*, 18:145–173, January.

McCrae, J. P. (2018). Mapping WordNet Instances to Wikipedia. In *Proceedings of Ninth Global WordNet Conference*, pages 62–69. The Global WordNet Association.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and applicationof a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Osenova, P. and Simov, K. (2018). The datadriven Bulgarian WordNet: BTBWN. *Cognitive Studies — Études cognitives*, 18.

Osenova, P., Simov, K., Laskova, L., and Kancheva, S. (2012). A treebank-driven creation of an ontovalence verb lexicon for Bulgarian. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2636–2640, Istanbul, Turkey. LREC 2012.

Palmer, M. (2009). Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*, page 9–15.

Rudnicka, E. K., Piasecki, M. T., Piotrowski, T., Łukasz Grabowski, and Bond, F. (2017). Mapping WordNets from the perspective of inter-lingual equivalence. *Cognitive Studies — Études cognitives*, 17(1373):1–17.

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 Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer-Verlag, February.

Simov, K., Simov, A., Ganev, H., Ivanova, K., and Grigorov, I. (2004). The CLaRK System: XML-based Corpora Development System for Rapid Prototyping. *Proceedings of LREC 2004*, pages 235–238, May.

Simov, K., Osenova, P., Laskova, L., Radev, I., and Kancheva, Z. (2019). Aligning the Bulgarian BTB WordNet with the Bulgarian Wikipedia. In *Proceedings of the 10th Global WordNet Conference*, pages 290–297. Oficyna Wydawnicza Politechniki Wrocławskiej.

Sprenger, S. (2003). *Fixed expressions and the production of idioms*. Ponsen and Looijen BV, Wageningen.

It Takes Two to Tango – Towards a Multilingual MWE Resource

Svetlozara Leseva
Institute for Bulgarian Language
Bulgarian Academy of Sciences
zarka@dcl.bas.bg

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

Ivelina Stoyanova
Institute for Bulgarian Language
Bulgarian Academy of Sciences
iva@dcl.bas.bg

Abstract

Mature wordnets offer the opportunity of digging out interesting linguistic information otherwise not explicitly marked in the network. The focus in this paper is on the ways the results already obtained at two levels, derivation and multiword expressions, may be further employed. The parallel recent development of the two resources under discussion, the Bulgarian and the Romanian wordnets, has enabled interlingual analyses that reveal similarities and differences between the linguistic knowledge encoded in the two wordnets. In this paper we show how the resources developed and the knowledge gained are put together towards devising a linked MWE resource that is informed by layered dictionary representation and corpus annotation and analysis. This work is a proof of concept for the adopted method of compiling a multilingual MWE resource on the basis of information extracted from the Bulgarian, the Romanian and the Princeton wordnet, as well as additional language resources and automatic procedures.

Keywords: wordnets, Bulgarian, Romanian, derivation, verbal multiword expressions, linked resources

1. Introduction

For almost a decade the development of the Bulgarian and the Romanian wordnets (BulNet and RoWN respectively) has involved shared research interests directed towards the enrichment of the two resources with qualitative information. Another relevant concern has been making linguistic information already existing in the wordnets accessible to computer processing: although the human specialist is able to spot different types of linguistic information in the two resources, it must be encoded in such a way that computer programmes can also easily access and use it. One of the avenues pursued along these lines has been digging derivational relations out of existing synsets and marking them explicitly in the two wordnets. Another strand of research involving joint efforts has been the encoding and exploration of multiword expressions (MWEs), and in particular several types of verbal multiword expressions (VMWEs) in wordnet synsets. The importance of MWEs has been widely acknowledged by linguistics and computational linguistics (Sag et al., 2002), as has been the significance of the ability of language processing systems to access resources in which such information is explicitly marked (Savary et al., 2019).

The results of these past and ongoing efforts have led to the idea of creating a lexical resource presenting a full description of MWEs that unifies the information available in wordnet with detailed morphological, syntactic, semantic, word order, pragmatic and derivational information. The greater goal is, using knowledge about Romanian, Bulgarian and English MWEs, to propose a framework for the description of MWEs that is applicable across languages, while also adaptable to language-specific features. Below we report on the ongoing work for the languages under study – Bulgarian and Romanian – with a recourse to the description of English VMWEs.

We start with a brief presentation of the development of the two wordnets under discussion and their enrichment with further relations (section 2). An interlingual analysis of the results of this enrichment is given in section 3. After that, we present the process of annotating verbal multiword expressions in the two wordnets with several multilingually defined types (section 4), while section 5 contains the results of the comparison between the types and the frequency of these verbal expressions in the two wordnets, as well as their interpretation. The work towards creating a multilingual linked VMWE resource that is currently underway is described in section 6.

2. BulNet and RoWN

The beginning and evolution of the Bulgarian wordnet (BulNet) (Koeva, 2010) and of the Romanian wordnet (RoWN) (Tufiş et al., 2013) were previously presented in Barbu Mititelu et al. (2017). Below we present some of the work carried out and made available through the two wordnets, which has inspired and informed our work on MWEs.

BulNet and RoWN were developed following the expand method (Rodriguez et al., 1998) and in compliance with two main principles: hierarchy preservation and conceptual density. Thus, the two wordnets preserve the structure of the original Princeton wordnet (PWN) and are aligned to it and, consequently, to each other and to any other wordnet aligned to PWN, this being a valuable asset for multilingual research and applications.

The interest in adding derivational relations to the two wordnets sprang up independently in the two teams: Koeva (2008) discusses important theoretical aspects of adding derivational relations to a wordnet, their multilingual relevance in the case of aligned wordnets, and presents the way in which the derivational relations from PWN were transferred and filtered in order to be included in BulNet. Barbu Mititelu (2013a) presents the methodology, heuristics and tools used for adding derivational relations to RoWN, as well as their importance for language applications.

For Romanian, Barbu Mititelu (2013b) presents in details the steps taken in the process of adding derivational relations among words of all parts of speech in RoWN (nouns, adjectives, verbs, adverbs). An initial phase of automatic identification of possible derivationally related pairs of any part of speech makes use of an exhaustive list of Romanian affixes. However, the resulting pairs require manual investigation for two reasons: on the one hand, some pairs contain false positives given that the beginning of a word can be misinterpreted as a prefix or the ending as a suffix, when this is a mere coincidence: consider the pair *val* 'wave' – *aval* 'downriver'; the latter can be morphologically misanalyzed as being formed from the former with the prefix *a-*, but this is not the case: *aval* is a French borrowing, where it is formed from Latin elements *ad-* 'at' and *vallis* 'valley'. On the other hand, manual validation is necessary because there must be a semantic connection between the words in a derivational relation; as such, there is a derivational relation between the words *drive* and *driver* when they are considered with the senses 'operate or control a vehicle' and 'the operator of a motor vehicle', respectively. There is also a derivational relation between them when they are considered with the senses 'push, propel, or press with force' and 'someone who drives animals that pull a vehicle', respectively. However, there is no such relation between them when considered with the senses interchanged (see Barbu Mititelu (2012) for a more detailed explanation).

As part of an effort along a similar avenue and in line with the theoretical considerations and the analyses proposed in Koeva (2008), Dimitrova et al. (2014) report on the steps, decisions and the theoretical motivation involved in the process of adding verb-noun derivational relations to BulNet. The adopted procedure was to start from the morphosemantic relations encoded in PWN (Fellbaum et al., 2009) and, using morphology-based heuristics, to identify and validate the derivational pairs in the corresponding BulNet synsets. Similar issues have been observed as the ones described for Romanian, particularly false positives and other errors due to overgeneration or failure of the procedures to capture different phonetic variants. An example of a false positive is represented by the pair *pod-slon-ya* 'give shelter' and *slon* 'elephant'. The results of the automatic procedures were therefore validated manually.

Besides marking the formal (i.e. morphological) relation between the words, their semantics was also described in terms of a set of predefined relations. For the noun-verb pairs these relations were bor-

rowed form PWN: Agent, Body-part, By-means-of, Destination, Event, Instrument, Location, Material, Property, Result, State, Undergoer, Uses, Vehicle. They all apply to the Bulgarian and Romanian pairs. With a view to discovering more derivational relations and attaching semantics to them in the already adopted framework, Koeva et al. (2016) proposed a machine learning method for automatic identification and classification of morphosemantic relations between pairs of potentially derivationally related verbs and nouns. The method employs the previously validated verb-noun derivationally related pairs and a number of linguistic features derived from the training data. The method is applicable to classifying MWEs as well, to the extent that the morphosemantic relations between single words would hold for MWEs headed by these single words.

3. Interlingual Comparison between Noun-Verb Relations in BulNet and RoWN

Annotating the morpho-semantically related noun-verb pairs in the two languages offered important insights into the derivational morphology of Bulgarian and Romanian as reflected in the respective wordnets (Tarpomanova et al., 2014): quantitatively, we found a richer system of suffixes in Bulgarian, as well as richer polysemy displayed by them. However, an important number of similarities could also be identified. Firstly, the same relations tend to be best or better represented in both wordnets: Agent and Event are the best represented from two perspectives: number of suffixes involved and frequency in the networks. The latter could also be regarded as a result of the similar objectives followed when deciding on the the wordnets development (see section 2). Almost all the other relations have a similar distribution in the annotated data for both languages[1].

Secondly, polysemous suffixes occurring in both languages are specialized for a certain set of relations, but have a preferred reading: e.g.: Bg. *-tel* forms nouns from verbs that bear the semantic relations Agent, Material, Instrument, By-means-of, Undergoer, and Uses, but Agent is by far the prevalent one; Ro *-(ă)tură* creates nouns that establish one of the following semantic relations – Event, Result, By-means-of, Instrument, Material, Uses – with the root verb, with Event being the best represented. There are suffixes occurring in both languages and showing high similarity in their semantics[2]: e.g. the suffix *-tor* is productive in both languages and in the vast majority of cases serves to derive nouns expressing the relation Agent, but may also be found with other relations such as: Instrument, Material, By-means-of, Uses.

4. Identifying and Classifying VMWEs in BulNet and RoWN

Wordnets contain both simple words and word combinations. The manual inspection of the latter has led to distinguishing (Barbu Mititelu et al., 2019), on the one hand, between free combinations with a compositional meaning (annotated with the label NONE) and expressions and, on the other hand, among several types of verbal multiword expressions, which were defined in the PARSEME shared task 1.0 (Savary et al., 2017) and then refined for shared task 1.1 (Ramisch et al., 2018). The VMWE labels used for annotating the VMWEs in BulNet and RoWN are: VID (verbal idioms), LVC.full (light verb constructions in which the verb is semantically bleached), LVC.cause (light verb constructions in which the verb has a causative meaning), IRV (inherently reflexive verbs), for both languages, and IAV (inherently adpositional verbs) only for Bulgarian (although such verbs also exist in Romanian, but the preposition remains underspecified in synsets). Table 1 illustrates all the types of labels with examples from the two wordnets.

[1]Further analysis of the data (Barbu Mititelu et al., 2015) involved comparison of the annotated noun-verb pairs in Bulgarian and Romanian with the corresponding English ones and that revealed the same tendency in the productivity of relations in all three languages, with Event, Agent and Result being the best represented. At the same time, the data confirmed the tendency towards conversion displayed by English.

[2]Larger data could further confirm these results as well as refine them.

VMWE type	**Example from BulNet**	**# in BulNet**	**Example from RoWN**	**# in RoWN**
VID	*cheta mezhdu redovete* 'read between the lines'	775	*citi printre rânduri* 'read between the lines'	614
LVC.full	*vzema uchastie* 'take part'	465	*lua parte* 'take part'	102
LVC.cause	*hvărlyam văv văztorg* 'cause to go into ecstasies'	63	*lăsa loc* 'allow for'	42
IRV	*gnevya se* 'become angry'	1,822	*[se] înfuria* 'become angry'	989
IAV	*razbiram ot* 'be good at'	39	not annotated	-

Table 1: Types of VMWEs in BulNet and RoWN and their distribution.

These types of VMWEs were defined within a multilingual context involving almost thirty languages from different families and displaying various characteristics. However, the annotation of the corpora participating in the shared tasks did not involve parallel corpora and neither was any interlingual analysis of VMWEs at the sense level made. Should there be such annotation available in the two wordnets, it would be possible to study interlingual equivalents.

5. An Interlingual Account of VMWEs

Previous analyses on VMWEs as represented in BulNet and ROWN, cf. Barbu Mititelu et al. (2019), have shown a number of parallels and differences between Bulgarian and Romanian VMWEs. In the paper under discussion, we report on 3,656 multitoken literal-to-literal pairs in corresponding synsets. These include VMWEs proper and multitoken free phrases (marked as NONE); their distribution is presented in Table 2 (for the purpose of comparison, suffix-based aspectual pairs in Bulgarian are counted as a single VMWE).

		BulNet			
		VID	LVC	IRV	NONE
RoWN	VID	**192**	16	99	140
	LVC	41	**44**	75	138
	IRV	151	64	**2,023**	148
	NONE	49	5	96	**263**

Table 2: Distribution of VMWE literal-to-literal correspondences between BulNet and RoWN.

5.1. Interlingual Analysis of the Data

With a big overlap of 72.7% reported between the VMWE types in the data under discussion, there are also plenty of examples of the same meaning being lexicalized by different types of VMWEs across the two languages and even in the same language (different-type MWE literals in one synset), a trend that is even more relevant when comparing or describing multiple languages.

As discussed therein, the interlingual correspondence is most consistent in the category of reflexive verbs (IRVs), which is to be expected given the similar semantics attached to reflexive verbs in the two languages (Slavcheva, 2006). In light of a dictionary-based approach to accounting for MWEs, IRVs do not pose considerable difficulties, as they constitute a recognized part of the vocabulary in both languages and their description follows the general guidelines for single words, as the reflexive component does not vary.

Other consistent correspondences are found in the class of verbal idioms (VIDs), which, as the authors admit, might be due to the fact that the choice of VIDs to encode was more or less influenced by internationally established idioms (respectively, calques in the two languages) already implemented

in PWN, such as {*read between the lines*:1}, {*send a message*:1}, among others. Nonetheless, these are expressions that are established in the languages under discussion and observe their morphological and syntactic peculiarities.

Correspondences are less marked in the domain of light-verb constructions (LVCs) for a couple of reasons: first of all, this class of VMWEs is not consistently represented in BulNet and ROWN – LVCs have usually been implemented to make up for lexical gaps; secondly, the teams working on the two wordnets have adopted different strategies, including a considerable difference in the number of light verbs recognized – 118 verbs for Bulgarian and 21 verbs for Romanian. Nonetheless, as the annotated data from the PARSEME corpora show, LVCs are pervasive in the two languages, so one of the objectives in proposing a dictionary-based resource is to properly account for this category of MWEs that is underrepresented in the two wordnets, as well as in many dictionaries.

As the type of VMWE that lexicalizes a particular sense is largely idiosyncratic, we would like the VMWE type to be an integral part of the entry of each individual VMWE: it should be assigned to or validated (if already available) individually for each VMWE literal (if there are more than one in a synset) and should be accessible for processing as the VMWE type enables the prediction of certain morphological, syntactic, word-order and other properties of the respective unit. Therefore, we have encoded the VMWE type as one of the features for description at the semantic level.

Given these observations, our efforts are directed primarily to capturing the linguistic features of LVCs and VIDs.

6. Towards a Multilingual Linked VMWEs Resource

The description of the various linguistic levels below is based on a proposal for the semi-automatic compilation of a MWE dictionary made in Stoyanova et al. (2016), which was further expanded to accommodate: (i) a stand-off format with links to wordnet synsets and literals; (ii) other levels of description, such as the VMWE types adopted in PARSEME, information about the connotation and the derivational potential of VMWEs; (iii) a multilingual testing setting for the description of VMWEs (Stoyanova et al., 2019).

The linked VMWE resource proposed harnesses several previously developed resources: (i) the three wordnets discussed above: RoWN, BulNet and PWN, which inform the general framework and provide a substantial part of the VMWE inventory as well as rich semantic and pragmatic information for the VMWEs included in them; (ii) VMWE annotated corpora for the two languages developed under the PARSEME initiative(Ramisch et al., 2018); (iii) single-word derivational patterns and instances for Romanian (Barbu Mititelu, 2013b) and Bulgarian (Dimitrova et al., 2014; Koeva et al., 2016) and MWE-to-MWE patterns for the two languages (Barbu Mititelu and Leseva, 2018).

6.1. Levels of Description

Below is a summary of the levels of description proposed.

1. **Technical information**. The technical information supports the linking between the dictionary entries and the respective wordnets, particularly through the unique synset ID and an additionally employed VMWE ID which serves both to identify a VMWE as part of a particular synset and to distinguish it from other VMWEs in the same synsets or from identical VMWE literals in other synsets. This allows us to: (a) access all the synset-level linguistic information provided; (b) make references to a particular VMWE uniquely, e.g., in the description of derivatives.

2. **Morphological description** which includes several types of information:

    - The lemma of the VMWE (non-abstract lemma) and a lemmatized form of each component of the VMWE (abstract lemma) (Savary, 2008). The parallel use of both types of lemma is motivated as follows: the non-abstract lemma is the human readable lemma, while the abstract one helps identifying VMWEs in a lemmatized corpus and assigning each such corpus occurrence of a VMWE a linguistically proper lemma that will link it to the wealth of information associated with the respective dictionary entry.

- A regular morphosyntactic representation which consists of the unrestricted set of forms of the expression's head and the unrestricted set of forms of the non-head components. This type of description is relevant for VMWEs with a full paradigm of the verbal head and its dependents and is typical of IRVs and IAVs, as well as of many LVCs. This set can be obtained from the in-house morphologic lexicons each team has.
- Restrictions on the paradigmatic realization of the verbal head with respect to one or more morphosyntactic features, such as person, number, tense, mood, polarity, etc., e.g. RO *nu privi cu ochi buni* (not watch with eyes good, 'regard with disfavour') is always used with the negative marker *nu* 'not'; the same goes for BG *ne iskam akăl nazaem* (not want brains to borrow, 'to not need unsolicited advice').
- Restrictions on the inflected forms of the dependent components of a VMWE. Such a field is defined for each dependent and is used to explicitly encode any restrictions on a dependent's possible forms as part of the VMWE. Considering the above example, the noun *ochi* ('eyes') is restricted to the plural indefinite form, while the BG *akăl* ('brains') is restricted to the singular indefinite form.

3. **Syntactic description**. The syntactic description is based on the UD framework as it aims at achieving universality, while offering the possibility to define language characteristics in the same framework (`https://universaldependencies.org/u/dep/index.html`).

   - Internal syntactic structure of the VMWE, which describes the number of components, the syntactic category of each them and the syntactic relations between the components. The most common structures found across Romanian and Bulgarian VMWEs as reflected in the analysed data are illustrated in Table 3.

Relation	**Description of relation**	**Example RO**	**Example BG**
V + obj	linking V to the entity acted upon or undergoing change	*da declarație* (give declaration, 'declare')	*vzemam reshenie* ('make a decision')
V + obl	linking V to a nominal as a non-core (oblique) argument or adjunct	*inceta din viață* (cease from life, 'die')	*poemam v svoi rătse* ('take into one's own hands')
V + advmod	linking V to a (non-clausal) adverb or adverbial phrase that modifies the predicate	*da afară* (give outside, 'remove from job, fire')	*vzemam predvid* ('take into account')
V + nsubj	the VMWEs is made up of a verb and a subject	*fura somnul* (steal sleep-the, 'fall asleep')	*zvezdata mi izgryava* ('one's star is rising')
V + nsubj + obl	linking V to a subject and a non-core (oblique) argument or adjunct	*îngheța sângele în vine* (freeze blood-the in veins, 'get cold feet')	*krăvta zamrăzva v zhilite mi* (blood-the freezes in my veins, 'get cold feet')
V + obj + obl	linking V, an object and a non-core (oblique) argument or adjunct	*găsi drumul în viață* (find road-the in life, 'find one's way in life')	*tsepya stotinkata na dve* (split the penny in two, 'be stingy')

Table 3: Types of syntactic structures across Romanian and Bulgarian VMWEs.

   - Possible dependents of the MWE elements. Some MWE components may have dependents, others may not. These dependents are either arguments, that is, obligatory dependents of the VMWE components, or adjuncts, i.e. dependents that are not required for the sentence to be grammatical.

     Argument dependents are usually predetermined by the head verb's argument structure. Consider, for instance RO *citi printre rânduri*, BG *cheta mezhdu redovete* and EN *read between*

the lines, which have an identical syntactic structure: the verbal head takes dependents of the type subject (nsubj or csubj in UD terminology) and direct object (obj in UD) in order to form a grammatical sentence and these positions need to be posited as slots in the VMWE description that need to be filled by a suitable phrase in order for the VMWE to form grammatical utterances. The dependent MWE component, in this case the prepositionally introduced noun, cannot be or is rarely modified by another word.

In contrast, the dependents of some VMWEs, light-verb constructions in particular, may readily take adjuncts of their own, e.g. BG *vzemam reshenie* ('make a decision') > *vzemam vazhno reshenie* ('make an important decision'), where the dependent noun, which is an object, may be modified (nmod in UD).

Both types of possible dependents must be encoded in the syntactic description, especially as many of them may intervene between the components of the VMWE; keeping track of them may provide useful information about the distance between the individual elements of a MWE in running text – a peculiarity that affects the automatic recognition of MWEs.

- Any restrictions on the word order of the VMWE components and of the possible dependents are encoded in this field. For example, in the RO VMWE *da ortul popii* (give coin-the to-priest-the, 'die') the obj *ortul* always precedes the indirect object *popii* (iobj in UD); in the BG VMWE *na star krastavichar krastavitsi prodavam* (to an old cucumber-seller cucumbers sell, 'to try to cheat someone with experience') the obl *na star krastavichar* precedes the obj *krastavitsi*, while the verb can be either at the front or at the end. This information is useful in MWE recognition.

4. **Semantic description**. The semantic description unites the idiosyncratic information about the basic properties of MWEs that predetermine their morphosyntactic behaviour, with lexical, usage and pragmatic information available from wordnet and possibly from other resources.

   - The MWE type – defined according to the guidelines adopted in the PARSEME shared task edition 1.1 (Ramisch et al., 2018).
   - The wealth of semantic information – the explanatory definition (gloss), single-word and MWE synonyms, other semantic and derivational relations, usage examples – that is accessible through the linking to wordnet and pertains to the entire synset.
   - Usage and register information. This field provides relevant restrictions on the usage of VMWEs, which may be automatically retrieved from the respective wordnet, if available, or added by a lexicographer. For example, many idioms are specific to the informal use, e.g. RO *bate la ochi* (beat at eyes, 'catch someone's eye'); BG *udryam kyoravoto*:1 (hit the blind, 'hit the jackpot'), and need to be accordingly marked.
   - The positive, negative or neutral connotation of a given VMWE whose value may be obtained either from available resources, such as SentiWordNet (Baccianella et al., 2010), or supplied manually. In the former case the connotation values are assigned from the respective synsets which have been assigned values from SentiWordNet (transferred automatically to BulNet and RoWN). For instance, the corresponding synsets BG: {*puka mi:1, dreme mi:2, davam pet pari:1, dam pet pari:1, davam puknata para:1, dam puknata para:1*}, RO: {*da doi bani:1, da două parale:1*}, EN: {*care a hang:1, give a hoot:1, give a hang:1, give a damn:1*} are assigned a positive value of +0.125 and a negative value of –0.375. Even so, manual validation is needed as the connotation value of individual literals may be language specific.

5. **Derivational information**. The derivational potential of MWEs has been tackled to a certain extent in the PARSEME initiative. The Romanian and the Bulgarian perspective on VMWE-to-MWE derivation, including a description of the semantic, syntactic and other changes that take place in the process of derivation, has been discussed in Barbu Mititelu and Leseva (2018), which we follow to a great degree. We adopt a verb-centric approach, regardless of the actual direction of the derivation, and currently focus on verb-to-noun derivation such as the one exemplified by the pairs: RO *spăla*

creierul – spălarea creierului, BG *promivam mozăka – promivane na mozăka*, EN *brainwash – brainwashing*. The derivational information is presented as a list of possible derivatives for each VMWE lexicon entry. Derivatives are encoded in the form that occurs in the respective wordnet and are accompanied by the ID of the synset to which they belong. If the derivatives are not implemented in the wordnet yet, then the ID remains unspecified.

6.2. Procedures for Semi-automatic Description

Below we present the baseline VMWE resource, which incorporates various levels of linguistic description for each of the languages. It was compiled using a series of automatic procedures and heuristics. The original VMWE inventory consists of all the synsets in BulNet and RoWN that contain at least one VMWE. Consequently, their correspondences in Princeton wordnet were also included regardless of whether they contain VMWEs. The baseline resource consists of 944 synsets which have a VMWE in both Bulgarian and Romanian with 2,744 literals on the Bulgarian side and 1,533 literals on the Romanian side. For 340 out of the 944 synsets there is a VMWE correspondence in English with a total of 662 VMWE literals.

The automatically retrievable information for each field of the description was assigned. Where possible, default values were determined, which need to be checked manually. The default values depend on a number of factors: (i) the form in which the VMWEs components are found in the lemma of the VMWE: if a component participates in a VMWE in its citation form, its full paradigm is its default value (not considering other factors); if the component in the VMWE's lemma is in a different form, it is most likely restricted with respect to the relevant grammatical category: consider, for instance, the VMWE *make advances* – the nominal component *advances* is in the plural in the lemma of the MWE and is unlikely to be found in the singular; (ii) the type of the MWE – for example, LVCs are more permissible than VIDs with respect to the modification of the dependents. Below we present the types of information that are automatically retrievable from the description of the MWEs in the wordnets under discussion.

1. **Automatic tagging and further morphological analysis**. The MWEs in the three languages are automatically POS-tagged using available programming tools. The BG data were annotated using the Bulgarian Language Processing Chain[3] and the RO and the EN MWEs were processed using the UDPipe with a Romanian and an English language model[4] respectively. The tagging was used in the grammatical description of the MWEs, in particular, for identifying (i) the POS tags of the MWEs components; (ii) the MWE's abstract lemma; (iii) the lexico-grammatical and grammatical features, such as verb aspect (in BG), number and definiteness for nominal components, etc.. As illustrated by the example above (*make advances*), the form in which a component is fixed in the non-abstract lemma, such as the one retrievable from wordnet, helps in predicting the possible variations of this component's grammatical properties (or a part of them).

2. **Syntactic analysis**. On the basis of the morphosyntactic tagging we derive the linear order of the components and we identify the basic internal syntactic structure of the MWEs, in particular: (i) the head and the dependents; (ii) the possible modifiers of the components (e.g. an NP dependent may take an adjective modifier); (iii) their basic word order and word order variations (e.g., the position of the reflexive particle in IRVs in BG and RO); (iv) the default values for the possible modifiers of the dependents based on the PARSEME type: 'yes' for LVCs, 'no' for VIDs and IRVs.

3. **Semantic description**. We extracted the available semantic information such as the synset ID, the definition, synonyms, semantic relations, register restrictions, etc. from the relevant synsets in the wordnets.

4. **Derivational information**. The derivational information is retrieved from wordnet as well by collecting all the synsets labelled as derivationally related to the one to which the MWE under discussion belongs regardless of the language for which the derivation applies. Further, we select

[3]http://dcl.bas.bg/dclservices/index.php
[4]http://ufal.mff.cuni.cz/udpipe

the multiword derivatives and analyse the matching components between the original MWE verb (literal in the verb synset) and the potential derivatives.

Table 4 shows the linking of corresponding MWE entries in BG *zatvaryam si ochite* 'close one's eyes', RO *închide ochii* 'close the eyes' and EN *turn a blind eye* with the components of their description. As the respective wordnet synsets do not have derivatives encoded, regularly produced derivatives – such as eventive nouns derived from verbs, e.g. BG *zatvarym si ochite* 'close one's eyes' – *zatvaryane na ochite* 'closing of the eyes' – need to be additionally extracted from corpus data or from available (lexicographic) resources.

Feature	**BG**	**RO**	**EN**
PWN ID	eng-30-00801977-v	eng-30-00801977-v	eng-30-00801977-v
MWE ID	bg_427	ro_265	en_3
Lemma ID	*zatvaryam si ochite*	*închide ochii*	*turn a blind eye*
Abstract lemma ID	*zatvaryam svoy oko*	*închide ochi*	EN *turn a blind eye*
Components	1_zatvaryam_V 2_svoy_PronP 3_oko_N	1_închide_V 2_ochi_N	1_turn_V 2_a_DET 3_blind_A 4_eye_N
Syntactic structure	V + obj	V + obj	V + obj
Verbal head	*zatvaryam*	*închide*	*turn*
Gram. features	1_VLITsr1_IMPERF	1_Vmip3s	1_VB
Dependents	2_svoy_PFPZ 3_oko_NCNpd	2_ochi_Ncmpd	2_a_DET 3_blind_A 4_eye_Ns
Restrictions	3_Npd	2_Npd	4_Ns
Modifiers	No	No	No
Word order	V_PronP order changes	–	fixed
PARSEME type	VID	VID	VID
Synonyms	bg_428: *zatvorya si ochite*	–	–
Register	Informal	Informal	Informal
Sentiment	–0.5 / +0.0	–0.0 / +0.0	–0.5 / +0.0

Table 4: An example of linked corresponding MWE entries in BG, RO and EN. (The POS notation is unified across the languages. POS: V – verb, N – noun, A – adjective, Adv – adverb, P – preposition, Pron – pronoun, DET – determiner, etc. The morphological features are partially unified so as to facilitate the use of the uniform notation of restrictions: PERF/IMPERF – verb aspect, s/p – singular/plural, 0/d – indefinite/definite, etc.).

7. Conclusions

The construction of the linked VMWE resource is work in progress and we are currently focused on the manual validation of the entries and the addition of missing linguistic information. Apart from providing description of Romanian and Bulgarian VMWEs in the adopted format, we are also interested in testing the applicability of the description cross-linguistically for capturing language-specific features towards obtaining a more fine-grained typology of syntactic and semantic types of VMWEs.

While the proposal makes use of widely recognized frameworks, such as aligned wordnets, the UD formalism, PARSEME VMWEs types, derivational morphology and semantics, our effort is aimed at accommodating them in a unified, data-driven framework and at providing a linked data formalism.

Acknowledgements

This work was carried out under the project *Enhancing Multilingual Language Resources with Derivationally Linked Multiword Expressions* between the Institute for Bulgarian Language at the Bulgarian Academy of Sciences and the Research Institute for Artificial Intelligence at the Romanian Academy.

References

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 7th Conference on Language Resources and Evaluation (LREC 2010), May 17-23, Valletta, Malta*, pages 2200–2204.

Barbu Mititelu, V. and Leseva, S. (2018). *Multiword expressions: Insights from a multi-lingual perspective*, chapter Derivation in the domain of multi–word expressions, pages 215–246. Berlin: Language Science Press.

Barbu Mititelu, V., Rizov, B., Tarpomanova, E., Leseva, S., and Dimitrova, T. (2015). Noun-Verb Derivation in the Bulgarian, Romanian and English Wordnets – A Comparative Approach. In *Proceedings of the 11th International Conference "Linguistic Resources and Tools for Processing the Romanian Language"*, pages 53–64.

Barbu Mititelu, V., Leseva, S., and Tufiș, D. (2017). The Bilateral Collaboration for the Post-BalkaNet Extension of the Bulgarian and the Romanian Wordnets. In *Proceedings of the International Jubilee Conference of the Institute for Bulgarian Language*, volume 2, pages 192–200.

Barbu Mititelu, V., Stoyanova, I., Leseva, S., Maria Mitrofan, T. D., and Todorova, M. (2019). Hear about Verbal Multiword Expressions in the Bulgarian and the Romanian Wordnets Straight from the Horse's Mouth. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 2–12.

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

Barbu Mititelu, V. (2013a). Increasing the Effectiveness of the Romanian Wordnet in NLP Applications. *Computer Science Journal of Moldova*, 21(3(63)).

Barbu Mititelu, V. (2013b). *Reţea semantico-derivaţională pentru limba română*. Editura Muzeul Literaturii Române, Bucharest.

Dimitrova, T., Tarpomanova, E., and Rizov, B. (2014). Coping with derivation in the Bulgarian wordnet. In *Proceedings of the 7th Global Wordnet Conference*, pages 109–117.

Fellbaum, C., Osherson, A., and Clark, P. E. (2009). *Responding to In- formation Society Challenges: New Advances in Human Language Technologies*, volume 5603, chapter Putting Semantics into WordNet's "Morphosemantic" Links, pages 350–358. Springer Lecture Notes in Informatics.

Koeva, S., Leseva, S., Stoyanova, I., Dimitrova, T., and Todorova, M. (2016). Automatic prediction of morphosemantic relations. In *Proceedings of the Eighth Global Wordnet Conference*, pages 168–176. University Al. I. Cuza Publishing House.

Koeva, S. (2008). Derivational and Morphosemantic Relations in Bulgarian Wordnet. *Intelligent Information Systems*, XVI.

Koeva, S. (2010). Bulgarian Wordnet – current state, applications and prospects. In *Bulgarian-American Dialogues*.

Ramisch, C., Cordeiro, S., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaitė, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A., and Walsh, A. (2018). Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *The Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240.

Rodriguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertagna, F., and Roventini, A. (1998). The top-down strategy for building eurowordnet: Vocabulary coverage, base concepts and top ontology. *Computers and the Humanities*, 32(2-3):117–152.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. pages 1–15.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., and Doucet, A. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*.

Savary, A., Cordeiro, S., and Ramisch, C. (2019). Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91. Association for Computational Linguistics.

Savary, A. (2008). Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, 1(2)).

Slavcheva, M. (2006). Semantic descriptors: The case of reflexive verbs. In *Proceedings of the 5th Language Resources and Evaluation Conference*, pages 1009–1014.

Stoyanova, I., Koeva, S., Todorova, M., and Leseva, S. (2016). Semi-automatic Compilation of a Very Large Multiword Expression Dictionary for Bulgarian. In *Proceedings of GLOBALEX 2016: Lexicographic Resources for Human Language Technology, Workshop at LREC-2016, Portorož, Slovenia, May 24, 2016*, pages 86–95.

Stoyanova, I., Leseva, S., Barbu Mititelu, V., Todorova, M., and Cristescu, M. (2019). Wrapping our Heads Around VMWEs and their Derivatives. In *Proceedings of the 14th International Conference "Linguistic Resources and Tools for Natural Language Processing"*, pages 153–166.

Tarpomanova, E., Leseva, S., Todorova, M., Dimitrova, T., Rizov, B., Barbu Mititelu, V., and Irimia, E. (2014). Noun-Verb Derivation in the Bulgarian and the Romanian WordNet – A Comparative Approach. In *Proceedings of the First International Conference Computational Linguistics in Bulgaria*, pages 23–31.

Tufiş, D., Barbu Mititelu, V., Ştefănescu, D., and Ion, R. (2013). The Romanian Wordnet in a Nutshell. *Language Resources and Evaluation*, 47(4):1305–1314.

Generating Natural Language Numerals with TeX

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

Milena Veneva
Independent Researcher
milena.p.veneva@gmail.com

Abstract

Sometimes one needs to produce a text in which many numbers have to be written out in words. Writing such a text and ensuring it is error-free can be a burden, especially if the author is not fluent in the language. Such may occur when working on a reference grammar, a research paper or presentation, or a problem on number names for a contest in linguistics. A remedy is to prepare the text with TEX and let some parts be generated automatically. The human effort this takes is to compose a grammar that describes the features of the numeral system. This paper discusses how this is done.

Keywords: number names, number systems, numerals, TEX, typesetting

1. Introduction

Sometimes one needs to produce a text in which many numbers have to be written out in words. Writing such a text and ensuring it is error-free can be a burden, especially if the quantity of numbers is very large, they change a lot during the editing, or the author is not fluent in the language. Such may occur when working on a reference grammar, a research paper or presentation, or a problem on number names for a contest in linguistics (Derzhanski and Veneva, 2018).

This burden can become lighter if the text is prepared with TEX (Knuth, 1986). Some parts can then be generated automatically (Derzhanski, 2013), and number names are a prime candidate. The human effort this takes is to compose a grammar that describes the features of the numeral system, but we leave the bulk of the typing and the proofreading to TEX.

It should be noted that in this day 'producing a document with TEX' tends to mean writing it in LATEX 2$_\varepsilon$ (LATEX 2$_\varepsilon$, 2018) or another such extension, but in fact the arithmetic operations, conditionals, switch-case statements, and other programming commands which facilitate the process of writing a semi-self-generating grammar pertain to pure TEX, albeit freely used within LATEX 2$_\varepsilon$.

2. The Problems

The method has been employed for generating numerals in eight languages in the statements and solutions of linguistic problems on number names that have been assigned at different instalments of the International Linguistic Olympiad (IOL) (www.ioling.org/) or national-scale contests in linguistics in Bulgaria (Derzhanski, 2009: Chapters 12 and 13). Here are the sources, the languages, their ISO 639-3 codes, families and countries where spoken:

1. IOL7 (Evgenia Korovina and Ivan Derzhanski): Sulka (sua: isolate, Papua New Guinea);
2. IOL8 (Ksenia Gilyarova): Drehu (dhv: Austronesian, New Caledonia);
3. IOL10 (Ksenia Gilyarova): Umbu-Ungu (ubu: Trans-New Guinea, Papua New Guinea);
4. IOL13 (Milena Veneva): Arammba (stk: South-Central Papuan, Papua New Guinea);
5. IOL13 (Milena Veneva): Classical Nahuatl (nci: Uto-Aztecan, Aztec Empire);
6. IOL15 (Milena Veneva): Birom (bom: Atlantic-Congo, Nigeria);

7. Winter Mathematics Contest 2000 (Ivan Derzhanski): Georgian (kat: Kartvelian, Georgia);
8. National Contest in Linguistics 2001 (Ivan Derzhanski): Yoruba (yor: Atlantic-Congo, Nigeria).

Table 1 summarises the principal features of the number systems of these languages, as well as Bulgarian as a point of comparison; that is, the answers to the following questions:

1. What is the base of the number system, and are there supplementary bases (such are often 5 and/or 10, and then perhaps 15, when the principal base is 20)?
2. Does the base have alternative (suppletive) names?
3. Are there any other numbers that play a base-like part in the number system?
4. Does the language use subtraction, or better, do the numbers just below the base behave – or are they formed – in an unusual way?
5. Does the language use overcounting (Menninger, 1969; Hanke, 2005)?
6. What, if any, arithmetic operations are marked?
7. Is the order of addends ($+$) and multiplicands ($\times$) ascending ($\nearrow$) or descending ($\searrow$)?
8. Are there any (morpho)phonological changes in the derivation of number names?

language	sua			dhv	ubu	stk	nci	bom	kat	yor	bul
1: base	20	20	20+	20+	24	6	20+	12	20+	20+	10
2: other names	no			yes	no	yes	yes	yes	no	no	no
3: other bases	3	4	no	no	4	no	no	no	no	no	no
4: subtraction (‘$-$’)	no			no	no	no	no	yes	no	yes	no
5: overcounting (‘$\neg$’)	no			no	yes	no	no	no	no	no	no
6: operations	$+, \times 2$			$+$	$\neg$	no	$+$	$+, \times$	$+, \times$	$+, -$	$+$
7(a): word order $+$	$\nearrow$			$\nearrow, \searrow$	$\searrow$	$\searrow$	$\searrow$	$\searrow$	$\searrow$	$\nearrow$	$\nearrow$
7(b): word order $\times$	$\searrow$			$\nearrow$	$\searrow$	$\nearrow$	$\nearrow$	$\searrow$	$\nearrow$	$\searrow$	$\nearrow$
8: (morpho)phonology	no			yes	no	no	yes	yes	yes	no	yes

Table 1: Linguistic phenomena in several number name systems.

3. The Idea

The idea of writing a computer program to convert a number to words is not original. It can be found under the form of a popular programming exercise on applying conditional and switch-case operators and manipulating strings of characters. For instance, problem #5.6 in (Dreyfus and Gangloff, 1975) concerns composing a program in Fortran IV to write out a given one- or two-digit number in French. Likewise, problem #31 in (Todorova et al., 2008) shows one of the ways to convert a two-digit number input from the keyboard to its Bulgarian name in C++.

4. TEX Definitions

Typesetting with TEX (Knuth, 1986) is akin to writing a program in several ways.

One is that frequently used constructions can be formulated as macro definitions—control sequences that can be evoked every time we need them. They can be mathematical formulae, words, sentences or even whole text passages. This reduces the number of keystrokes, typing errors and inconsistencies.

Another is that information of various types can be stored in variables (registers), which can be assigned values and performed operations on (in particular, integer arithmetics).

Finally, there are flow of control constructions of the if-then-else kind (depending on the outcome of a numeric comparison or another boolean condition) and the switch type (depending on the non-negative integer value of a variable).

5. Implementation

5.1. Bulgarian

Bulgarian has a decimal number system; up to 99 multiplication is expressed by juxtaposition and addition by the preposition **na** ‘on, over’ and the conjunction **i** ‘and’. The number names in this range are

formed as follows:

Rule no.		Lines
1	**edno** 1, **dve** 2, **tri** 3, **chetiri** 4, **pet** 5, **shest** 6, **sedem** 7, **osem** 8, **devet** 9	#20
2	**deset** 10	#18
3	α·**na**·**deset** = $10 + \alpha$ $\quad (1 \leq \alpha \leq 9)$ (if $\alpha = 1$, the stem is **edi**; if $\alpha = 2$, the stem is **dva**)	##17–18
4	β·**deset** $\quad (2 \leq \beta \leq 9)$ $\quad$ (if $\beta = 2$, the stem is **dva**)	#13
5	β·**deset i** $\alpha = \beta \times 10 + \alpha$ $\quad (1 \leq \alpha \leq 9, 2 \leq \beta \leq 9)$ (if $\beta = 2$, the stem is **dva**)	##13–14, #20

Table 2: Rules for Bulgarian

Figure 1 shows the TEX macros for generating number names in the range $[1; 99]$. The top macro is `\blg` (e. g., `\blg{42}` produces **chetirideset i dve** $4 \times 10 + 2$), and it turns to the auxiliary `\bln`, which yields numerals in the range $[1; 9]$, in the general case for both the `\tens` and the `\ones`, into which the argument is split in lines 9–10. Since by default counting is done in the neuter gender but the number 2 as a multiplier in 20 and both 1 and 2 as addends in the second decade are in the masculine, the flag `\ifneutrum` is set to indicate that a neuter form is required.

```
1  \newcommand \bln[1]{\ifcase #1
2      \or ed\ifneutrum no\else i\fi \or dv\ifneutrum e\else a\fi
3      \or tri\or chetiri\or pet\or shest\or sedem\or osem\or devet\fi
4  }
5  \newcount \tens \newcount \ones
6  \newif \ifneutrum
7  \newcommand \blg[1]{%
8      \ifnum #1<100
9          \tens=#1\divide \tens by 10
10         \ones=-\tens \multiply \ones by 10\advance \ones by #1
11         \neutrumfalse
12         \ifnum 1<\tens
13             \bln{\tens}deset%
14             \ifnum 0<\ones \space i \fi
15         \fi
16         \ifnum 1=\tens
17             \ifnum 0<\ones \bln{\ones}na\fi
18             deset%
19         \else
20             \ifnum 0<\ones \neutrumtrue \bln{\ones}\fi
21         \fi
22     \fi
23 }
```

Figure 1: Macro definitions for generating Bulgarian numerals up to 99.

5.2. Birom

The use of macros to avoid typos is most opportune when the number names contain many diacritics, which happens to be the case in Birom. The number names up to 120 (the range featured in the problem; in fact the same rules hold for $[121; 131]$ as well, and only fail to do so at 132) obey the rules in Table 3.

The TEX macros for generating these numerals in Birom are shown in Figure 2. The main macro

Rule no.		Lines
1	gwīnìŋ 1, bà 2, tàt 3, nààs 4, tùŋūn 5, tìīmìn 6, tàāmà 7, rwīīt 8	#30
2	ʃāā-$\alpha = 12 - \alpha$ $(1 \leq \alpha \leq 3)$: ʃāātàt 9, ʃāābà 10, ʃāāgwīnìŋ 11	#30
3	kūrū 12	#21
4	bā-kūrū bī-$\bar{\gamma}$ $= \gamma \cdot 12$ $(2 \leq \gamma \leq 8)$, bā-kūrū ʃāā-bī-$\bar{\gamma}$ $= (12 - \gamma) \cdot 12$ $(1 \leq \gamma \leq 2)$ (the tone in the first syllable of γ becomes middle)	##22–23
5	β ná $\left\{ \begin{array}{l} \text{gwɛ̄} \ (\delta = 1) \\ \text{vɛ̀} \quad (2 \leq \delta \leq 11) \end{array} \right\} \delta = \beta + \delta \qquad (\beta = k \cdot 12)$	##21–28

Table 3: Rules for Birom

```
\newcommand \biron [3]{%
    \newcount \numm \numm=#3%
    \ifnum 8<\numm \textesh\={a}\={a}%
        \advance \numm by-12\numm =-\numm
    \fi
    #1\ifcase \numm \or gw\={\i}n\'{\i}\ng%
        \or b#2{a}\or t#2{a}t\or n#2{a}#2{a}s%
        \or t#2{u}\ng\={u}n\or t#2{\i}\={\i}m\'{\i}n%
        \or t#2{a}\={a}m\'{a}\or rw\={\i}\={\i}t\fi
}
\newcount \biggestBM
\newcount \biggestBA
\newcommand \birom [1]{%
    \newcount \numb \numb=#1
    \ifnum 121>\numb
        \ifnum 11<\numb
            \biggestBM=\numb
            \divide \biggestBM by 12
            \biggestBA=\biggestBM
            \multiply \biggestBA by 12
            \ifnum 1=\biggestBM k\={u}r\={u}%
            \else b\={a}k\={u}r\={u}
                \biron{b\={\i}}{\={}{\biggestBM}\fi
            \advance \numb by -\biggestBA
            \ifnum 0<\numb \space n\'{a}
                \ifnum 1=\numb gw\={\textepsilon}
                \else v\'{\textepsilon} \fi
            \fi
        \fi
        \ifnum 0<\numb \biron {}{\'}{\numb}\fi
    \fi
}
```

Figure 2: Macro definitions for generating Birom numerals.

is `\birom` (e. g., `\birom{117}` produces *bākūrū ʃāābītāt ná vɛ̀ ʃāātàt* $(12-3)\times 12+(12-3)$). The auxiliary `\biron` is somewhat more complex than its counterpart for Bulgarian: it implements the expression of 9–11 by subtraction and the prefixing of *bī-* and the tone change in coefficients in the names of dozens.

A part of the text of the problem on Birom numerals prepared with the use of this method (as LaTeX source and typeset) is in Figure 3.

```
\newcommand \numB [1]{\mbox {\bomfont {\birom {#1}}}}

\begin{enumerate}
 \item $\numB{5}^2 + \numB{3} + \numB{4} = \numB{32}$
 \item $\numB{3}\totheB{4} = \numB{81}$
 \item $\numB{7}^2 + \numB{9} + \numB{1} = \numB{59}$
 \item $\numB{9}\totheB{1} = \numB{9}$
 \item $\numB{8}^2 + \numB{2} + \numB{5} = \numB{71}$
 \item $\numB{2}\totheB{5} = \numB{32}$
 \item $\numB{9}^2 + \numB{4} + \numB{3} = \numB{88}$
 \item $\numB{4}\totheB{3} = \numB{64}$
 \item $\numB{16} + \numB{21} = \numB{18} + \numB{2} + \numB{17}$
\end{enumerate}

...
\item Write the numbers \numB{36}, \numB{11}, \numB{12}
and the equalities (A) and (B) in numerals.
%
\begin{enumerate}
 \item[A.] $\numB{108} - \numB{3} - \numB{13} = \numB{92}$
 \item[B.] $\numB{49} - \numB{14} - \numB{15} = \numB{20}$
\end{enumerate}
```

1. **tùŋūn^2 + tàt + nààs = bākūrū bībā ná vɛ̀ rwīīt**
2. **tàt$^{\text{nààs}}$ = bākūrū bītīīmìn ná vɛ̀ ʃāātàt**
3. **tàāmà^2 + ʃāātàt + gwīnìŋ = bākūrū bīnāās ná vɛ̀ ʃāāgwīnìŋ**
4. **ʃāātàt$^{\text{gwīnìŋ}}$ = ʃāātàt**
5. **rwīīt^2 + bà + tùŋūn = bākūrū bītūŋūn ná vɛ̀ ʃāāgwīnìŋ**
6. **bà$^{\text{tùŋūn}}$ = bākūrū bībā ná vɛ̀ rwīīt**
7. **ʃāātàt^2 + nààs + tàt = bākūrū bītāāmà ná vɛ̀ nààs**
8. **nààs$^{\text{tàt}}$ = bākūrū bītūŋūn ná vɛ̀ nààs**
9. **kūrū ná vɛ̀ nààs + kūrū ná vɛ̀ ʃāātàt = kūrū ná vɛ̀ tìímìn + bà + kūrū ná vɛ̀ tùŋūn**

(b) Write the numbers **bākūrū bītāt**, **ʃāāgwīnìŋ**, **kūrū** and the equalities (A) and (B) in numerals.

A. **bākūrū ʃāābītāt − tàt − kūrū ná gwɛ̄ gwīnìŋ = bākūrū bītāāmà ná vɛ̀ rwīīt**

B. **bākūrū bīnāās ná gwɛ̄ gwīnìŋ − kūrū ná vɛ̀ bà − kūrū ná vɛ̀ tàt = kūrū ná vɛ̀ rwīīt**

Figure 3: An excerpt from the statement of the Birom problem.

5.3. Yoruba

Yoruba operates a decimal–vigesimal system; its most peculiar feature is that subtraction (of 10 from a whole twenty and of 1 to 5 from a whole ten) is liberally used where most other languages use addition. The rules produce the numbers up to 184 except for the range $[25; 34]$, because 30 has a suppletive name, which was not featured in the problem for which the macros shown here were made.[1]

Rule no.		Lines
1	*ọkan* 1, *eji* 2, *ẹta* 3, *ẹrin* 4, *arun* 5, *ẹfa* 6, *eje* 7, *ẹjọ* 8, *ẹsan* 9	#6
2	*ẹwa* 10	#19
3	*ogun* 20	#24
4	*ogun* $\beta = \beta \times 20$ $(2 \leq \beta \leq 9)$	##24–25
5	*ẹwa din ogun* $\beta = \beta \times 20 - 10$ $(3 \leq \beta \leq 9)$	##23–25
6	α *din* $\gamma = \gamma - \alpha$ $(1 \leq \alpha \leq 5; \gamma = k \cdot 10, 4 \leq k \leq 19)$	##16–17
7	α *l-*$\gamma = \gamma + \alpha$ $(1 \leq \alpha \leq 4; \gamma = k \cdot 10, 1 \leq k \leq 19)$	##16–17

Table 4: Rules for Yoruba

The TEX macros for generating these numerals in Yoruba and an excerpt from the text of the problem produced with their use (as LATEX source and typeset) are shown in Figures 4 and 5, respectively.

```
\newcommand \yorn [1]{%
    \ifcase #1\or\d okan\or eji\or\d eta\or\d erin\or arun \or\d efa
    \or eje\or\d ej\d o\or\d esan\fi}
\newcount \scor \newcount \tens \newcount \ones \newcount \absones
\newcommand \yorr [1]{%
    \ifnum #1<10 \yorn #1%
    \else
        \tens =#1\divide \tens by 10
        \ones =-\tens \multiply \ones by 10\advance \ones by #1
        \ifnum 4<\ones
            \advance \tens by 1\advance \ones by-10%
            \absones =0\advance \absones by-\ones
        \else \absones =\ones
        \fi
        \ifnum 0=\ones
        \else \yorn \absones \space
            \ifnum 0<\ones l-\else din \fi
        \fi
        \ifnum 1=\tens \d ewa%
        \else
            \scor =\tens \divide \scor by 2%
            \ones =-\scor \multiply \ones by 2\advance \ones by \tens
            \ifnum 1=\ones \advance \scor by 1\d ewa din\fi
            ogun%
            \ifnum 1<\scor \space \yorn \scor \fi
        \fi
    \fi
}
```

Figure 4: Macro definitions for generating Yoruba numerals.

[1]The forms given here are in fact reconstructions which reveal the internal structure of the numerals but conceal the complex morphophonological processes which produce the surface forms of the contemporary living language.

```
\newcommand \yorl [1]{$#1$ & \textit {}\yorr{#1}}}

\begin{tabular}{rl}
\yorl 3 \\
\yorl{11} \\
\yorl{22} \\
\yorl{37} \\
\yorl{66} \\
\yorl{93} \\
\yorl{135}
\end{tabular}
...
\item[(a)]Identify the numbers: \yorr{144}; \yorr{45}.
```

3	*ẹta*
11	*ọkan l-ẹwa*
22	*eji l-ogun*
37	*ẹta din ogun eji*
66	*ẹrin din ẹwa din ogun ẹrin*
93	*ẹta l-ẹwa din ogun arun*
135	*arun din ogun eje*

(a) Identify the numbers: *ẹrin l-ogun eje*; *arun din ẹwa din ogun ẹta*.

Figure 5: An excerpt from the statement of the Yoruba problem.

5.4. Some other noteworthy issues

The other occasions in which the method has been used for writing number names in linguistic problems will not be considered in detail here, for want of space, but a few notes on various interesting issues that come up are in order.

Generating large numerals in a non-decimal system is error-prone. The Arammba number system is base-6 and goes up to $6^7 = 279\,936$, so there is much to be gained by leaving the number crunching to the computer. This passage encodes the fact that a number greater than or equal to $6^5 = 7776$ (and presumed less than $2 \times 6^5 = 15552$ because of the parameters of the linguistic problem for whose typesetting the macros were composed) is named *weremeke* '6^5' followed by the difference:

```
\ifnum 7775<\numb
    \ifstarted \space \fi
    weremeke\advance \numb by -7776
    \startedtrue
\fi
```

Likewise for the lower degrees of 6, with coefficients where necessary. The same technique is applied, in the text of the same linguistic problem, for the base-20 system of Nahuatl (if the number is greater than 7999, then 8000 is subtracted, etc.).

Umbu-Ungu is base-24, with 4 as a secondary base, but has special (unanalysable) names for all multiples of 4 up to 32, which means that, although 48 is expressed as 24×2 *tokapu talu*, the following two fours are $52 = 24 + 28$ *tokapu alapu* and $56 = 24 + 32$ *tokapu polangipu*, and 24×2 only comes up in $60 = 24 \times 2 + 12$ *tokapu talu rurepo*. The macro `\tuu{`k`}` (where $3 \leq k \leq 26$; k is not divisible by 6) generates the name of the kth multiple of 4.

```
\newcommand \tuu [1]{\uux=#1\relax
  \ifnum 20<\uux tokapu yepoko \advance\uux by-18
  \else \ifnum 14<\uux tokapu talu \advance\uux by-12
  \else \ifnum 8<\uux tokapu \advance\uux by-6
  \fi\fi\fi
  \advance\uux by-2
  \ifcase \uux \or rurepo\or malapu\or
    supu\or tokapu\or alapu\or polangipu\fi}
```

Also the language uses overcounting, so 57 is *tokapu talu rurepo·nga telu* $24 \times 2 + 12 \neg 1 = 60 \neg 1$ (‘1 from the 4 that completes 60’). This is implemented by checking if the number of ones is zero, and if not, adding 1 to the number of fours before generating their name.

```
\ones=#1
\uuy=#1\divide \uuy by4\fours=\uuy
\multiply \uuy by4\advance \ones by-\uuy
\ifnum 0=\ones \tuu {\fours}%
\else \advance \fours by1\tuu {\fours}nga
    \ifcase \ones \or telu\or talu\or yepoko\fi%
\fi
```

Drehu, which has a vigesimal system but uses 5, 10 and 15 as supplementary bases, calls these three numbers β*-pi*, where β is the quantity of fives, but has a suffix for each of them when a number of the range $[1; 4]$ is to be added: 15 is *köni·pi* 3×5 but 18 is *köni·qaihano* $3 + 15$.

```
\ifnum 0=\ones
    \Drehun \fems pi%
\else \ifcase \fems \Drehun \ones
    \or \Drehun \ones ngömen%
    \or \ifcase \ones \or caa\or lua\or köni\or eka\fi ko%
    \or \Drehun \ones qaihano\fi
```

The macro `\Drehun` produces the numbers from 1 to 4; the forms they assume before the suffix *-ko* ‘$+ 10$’ are simply listed because of various opaque morphophonological changes.

Generating whole noun phrases containing numerals as quantifiers can present additional challenges. The Sulka language has three number systems (for counting coconuts, breadfruit, and everything else). Some of the nouns have suppletive singular and plural forms (e. g., sg. *tu*, pl. *sngu* ‘yam’). There is also a dual number (marked by *lo* preposed to the singular), although it does not preclude the use of a numeral. So generating an expression combining a noun and a number involves choosing the appropriate system as well as putting the noun in the appropriate grammatical form (*a tu a tgiang* ‘1 yam’, *a lo tu a lomin* ‘2 yams’, *o sngu a korlotge* ‘3 yams’).

6. Conclusions

It is hoped that this brief exposition has sufficed to demonstrate both the advantages of leaving the construction of complex number names to the computer whilst creating a text – in essence, a minor exercise in automatic natural language generation – and the difficulties one may encounter when doing so. The last example that was mentioned here touched upon the possibility of expanding the method beyond the numeral, which hints at the great potential of the approach.

References

Derzhanski, I. and Veneva, M. (2018). Linguistic Problems on Number Names. In *Proceedings of the Third International Conference Computational Linguistics in Bulgaria*, pages 169–176. https://dcl.bas.bg/clib/wp-content/uploads/2018/07/CLIB_2018_Proceedings_v2_final.pdf.

Derzhanski, I. (2009). *Linguistic Magic and Mystery*. Sofia: Union of Bulgarian Mathematics.

Derzhanski, I. (2013). Multilingual Editing of Linguistic Problems. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 27–34.

Dreyfus, M. and Gangloff, C. (1975). *La pratique du Fortran: Exercises commentés*. Paris: Dunod.

Hanke, T. (2005). *Sum: Overcounting in Numerals*. *LINGUIST List* **16**.2448. https://linguistlist.org/issues/16/16-2448.html.

Knuth, D. E. (1986). *The TeXbook*. Addison-Wesley Professional.

LaTeX 2_ε. (2018). *LaTeX2e Unofficial Reference Manual*. http://tug.org/texinfohtml/latex2e.html.

Menninger, K. (1969). *Number Words and Number Symbols: A Cultural History of Numbers*. Cambridge: MIT Press.

Todorova, M., Armyanov, P., Petkova, D., and Georgiev, K. (2008). *Sbornik ot zadachi po programirane na C++: Chast 1*. Sofia: Tehnologika.

A Natural Language for Bulgarian Primary and Secondary Education

Iglika Nikolova-Stoupak
Staffordshire University
iglika.nikolova.stoupak@gmail.com

Abstract

This paper examines the qualities and applicability of a provisional programming language, especially designed for use by beginner-level students in Bulgarian primary and secondary schools. The necessity for such a language is investigated. Then, relevant features are defined, as inspired by various programming languages (notably, languages used in education and characterised with non-English syntax) and by general trends related to the achievement of natural language in software development. A survey is conducted to test young students' interaction with the language, and the latter's advantages and limitations are listed and discussed.

Keywords: natural-language-programming, elementary education, Bulgarian education, cultural and social implications

1. Context

1.1. Computer Education in Bulgaria

Generally speaking, Bulgarian people take pride in their country's relationship with computer science. They are ready to point out that John Atanasoff, the inventor of the first electronic digital computer, was of Bulgarian origin. In terms of contemporary context, many are happy to discover that Bulgaria's Internet speed and accessibility rank within the world's top lists ("Bulgaria Ranks World's 20th in Internet Speed, Accessibility", 2014).

Computer studies were first introduced in Bulgarian education as early as 1959, in the University of Sofia and under the course name "Computational Mathematics and Cybernetics" (Kaltinska, 2018). The then utilised Romanian ECM CIFA machine was replaced by the Bulgarian computer Vitosha in 1963 (2018). They both worked solely with machine code. Computer Science or Informatics entered specialised high school education in 1972 for grade levels 10 and 11 (i.e. student age 16-18) (2018). The subject became mandatory for Bulgarian education in 1986, starting from grade 6 (age 12-13), and the language Logo was most frequently used during the next few years (Zamfirov, 2016: 47).

In 2008, a survey was carried out among high school teachers of Mathematics and Science, revealing a common opinion that computer studies were integrated too late within the curriculum, especially given students' experience with technology as part of contemporary everyday life ("Popitahme uchitelite!":2). In order to fill this hole, and partly motivated by recent growth in the IT sector in the country, computer education is currently being redefined and popularized (Stoyanova, 2013). Thematic extracurricular activities are increasingly on offer for young children, such as via established institutions,

newly opened internationally affiliated organisms and subscription-based websites. Many of these courses invite even preschoolers, who are not expected to be able to read or write.

Public education has been attempting to keep up with the described trend. Over the past decade, various Bulgarian schools have participated in facultative initiatives related to computer studies (and associated teacher training) which have been highly successful (Ayvaz, 2018). “It is a fact that the driving force behind the maintenance of interest and the improvement of instruction is the entirely personal initiative of teachers and students,” “Popitahme uchitelite!” asserts (2008: 2). As of the academic year 2018/2019, a major development occurred with the subject becoming obligatory in the context of primary education (namely, grade 3), under the title “Computer Modelling”. The current plan is to gradually include this subject in the curriculum for grade 1 (Dyulgerova, 2018). The basic topics discussed within the course include: what a computer is and basic hardware components; the generation of user profiles; risks and precautions; and basic algorithms.

It does not come as a surprise that the described young computer curriculum is still largely imperfect. Firstly, teacher training takes place in solely a single day and specialises strictly in the material covered by the course (Regional Educational Management–Sofia, 2018), thus generating an issue in terms of teaching competence. Computer equipment is, unfortunately, extremely scarce. As of 2013, the country’s average is one computer for eleven students as opposed to a European average of one to five (Nikolov, 2013: 10). Negative consequences of the problem are already perceivable, as whilst good theoretical knowledge is objectively demonstrated by Bulgarian IT students, their practical skills remain far from satisfactory (10). Finally and very importantly, no programming language has been established that reflects the Bulgarian cultural and educational context. The international visual language Scratch is most commonly used; specifically, in its imperfect and partial Bulgarian translation.

Given the described gaps within the Computer Studies curriculum as present within Bulgarian education, it is relevant to undergo the current project. Especially following the unexpected and unprecedented necessity for online education during the academic year 2019/2020, computer education and computer literacy within the entire school curriculum have come to acquire a key role. Consequently, the lack of a defined, optimally usable and appealing programming language is to be becoming increasingly obvious.

1.2. International Trends

An important general trend in relation to contemporary programming languages is the quest for “natural” language; in other words, a programming language aims to be as close as possible to the programmer’s “human” language and, consequently, as removed as possible from the machine code that hides behind the offered interface. Examples of so-called “high-level” or natural languages include COBOL, Pegasus and Jaa (a Java dialect). Contemporary research is highly centered on similar languages’ development and optimisation, and they are especially praised when child or beginner programmers are concerned. Stefik and Siebert show through an experiment that users, notably inexperienced ones, find established languages like Python highly more intuitive to use than a made-up language named “Randomo”, whose commands are not designed to resemble human language (2013). Myers et al, stating as their goal “to make it possible for people to express their ideas in the same way they think about them”, conduct a detailed study with non-programmers to define and test the language and environment Human-centered Advances for the Novice Development of Software (HANDS), where animations and simulations are utilised to express meaning (2004: 47).

To go further, researchers tend to agree that the naturalness and intuitiveness of a programming language meant for beginners can benefit greatly if native rather than English-based syntax is used where applicable. For instance, according to the designer of Russian educational language KuMir, Dr. Leonov, it was mandatory that the language have native syntax, as introducing a foreign one would add to the already unavoidable initial confusion that students experience (2013: 137). On a similar note, Baron et al claim in relation to French-syntax educational language LSE that the use of French can both facilitate the learning

process and avoid the potential negative interference that a non-native language might have on learners (1985: 10).

Let us also note that upon reaching a particular target audience, software needs to be both translated and "culturalised." It is therefore important that national programming languages, whether adapted or specially developed, take into account the very culture at hand. An example of optimal development in this regard is Japanese educational language "Kotodama on Squeak", which not only uses local syntax but also seeks to appeal to an audience that values literary and esthetic language, such as by removing abbreviations and ensuring that sentences feature correct and varied grammar.

The last feature specific to educational languages that is going to be underlined is their involvement in a particular school curriculum as opposed to isolated use. The skills developed in a computer course can naturally be applied to other aspects of academic life, including general research, the completion of homework, and the assimilation of mathematical skills and notions. French-language LSE illustrates this interdisciplinary quality especially vividly with the large associated library of educational software made available by the French National Center for Educational Documentation, which was launched soon after the language's popularisation in the 1980s and covered all school disciplines (Baudé, 2016: 48). Interestingly, this plurality would not be unprecedented within Bulgarian education itself. In 1984, a Computer Science textbook by Nikolov and Sendova claimed to be instructive simultaneously in Logo programming, Mathematics, English and Russian (Zamfirov, 2016: 48). It would thus be relevant to point out that the previously emphasised importance of native language programming is not to say that English should be avoided per se, especially taking into consideration its major role in school curriculums and the contemporary global world in general; rather, it is the *availability* of one's native language that should be a key concern.

2. Research Design

2.1. Methods

Secondary research for the current project encompasses sources related to the educational and cultural context at hand as well as to applicable international practices (particularly focusing on educational programming languages with national syntaxes) and to global trends concerning natural language programming. Primary research analyses the development and application of a programming language as designed for use within primary and secondary education in Bulgaria. The provisional language, Monoglossia (1Gl), is tested by way of a survey issued among students, and its discussion aims at constructive conclusions pointing at further work. Taking into account the importance of potential use of the language even without the presence of a computer (as mandated by current restrictions of equipment in Bulgarian public classrooms), the survey was printed out and completed in pen. For the language's full Syntax specification, please refer to Appendix D (Nikolova-Stoupak, 2020).

2.2. Participants

The survey was distributed to 20 children from a Bulgarian primary school, aged from 9 to 12. Their prior experience with programming was minimal, and their level of English was basic (with the exception of one child, bilingual in English and Bulgarian).

3. Primary Research

3.1. Data Collection: Survey

The survey (Nikolova-Stoupak, 2020 Appendix A) was completed by 20 children from a Bulgarian primary school, aged 9-12. They either had no formal experience with programming or had been following a Computer Basics course at school for no more than a few months. With the exception of one bilingual child, the participants' knowledge of English was beginner.

The survey consists of three programming exercises. Exercise 1 involves writing code in Bulgarian in the language prototype 1Gl, exercise 2 requires coding in English in 1Gl, and exercise 3 is composed in Logo, the English-based educational language that has historically been used in Bulgarian education. No

prior knowledge is required on the side of the students, as there are explanations and examples of all utilised constructions. Each exercise includes basic visual commands in its first part and the use of a simple loop in its second part.

The students were asked to record the time it took them to complete each exercise (in minutes). The exercises were followed by two "yes/no" questions: whether it was easier to program in Bulgarian; and whether writing code in English feels like English language practice to them. Finally, they were asked to identify the exercise that they deemed most difficult.

Via quantitative analysis, the survey seeks to achieve the following goals:

- Examine the kinds of language-related mistakes committed by students and their occurrence.
- Compare the number of non-language-related mistakes in 1Gl and Logo.
- Determine whether programming in Bulgarian was easier for the students.
- Correlate the timing of completion of the three exercises.

For all raw data involved in the discussion, please refer to Appendix B (Nikolova-Stoupak, 2020).

3.2. Results and Data Analysis

- Figure 1 presents the types of language-related mistakes committed by students along with percentage. Almost half of all mistakes are related to spelling (46.5 %), mistakes involving non-existent syntax following at 38.6%. Mistakes in punctuation and other undefined mistakes come together at 14.8%.

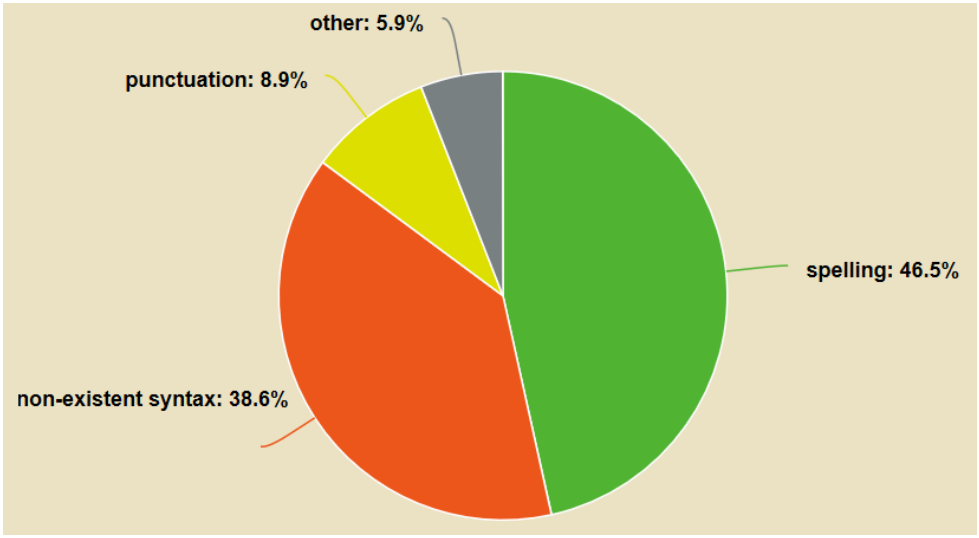


Figure 1: Types of language-related mistakes committed by students

It is important to note that, as should be expected, the vast majority of language-related mistakes (81%) were committed in the more natural-language-like programming language, 1Gl. The presence of spelling mistakes (especially in English) is explainable given the young age of students and the fact that most of them completed the survey on paper and without access to language tools. Mistakes linked to syntax were mostly based on wrong assumptions about the programming language's syntax (such as the writing of non-existing commands), inspired by students' experience with human language.

- As Figure 2 shows, the number of non-language-related mistakes (including use of symbols, use of loops, missing commands, unnecessary commands, misunderstood instructions, wrong calculation and non-optimal programming practice) is significantly higher in the Logo exercise (68) as compared with 1Gl (50 for Bulgarian-based and 55 for English-based). It may thus be suggested that the natural quality achieved by 1Gl aids in the prevention of mistakes of mathematical and logical nature.

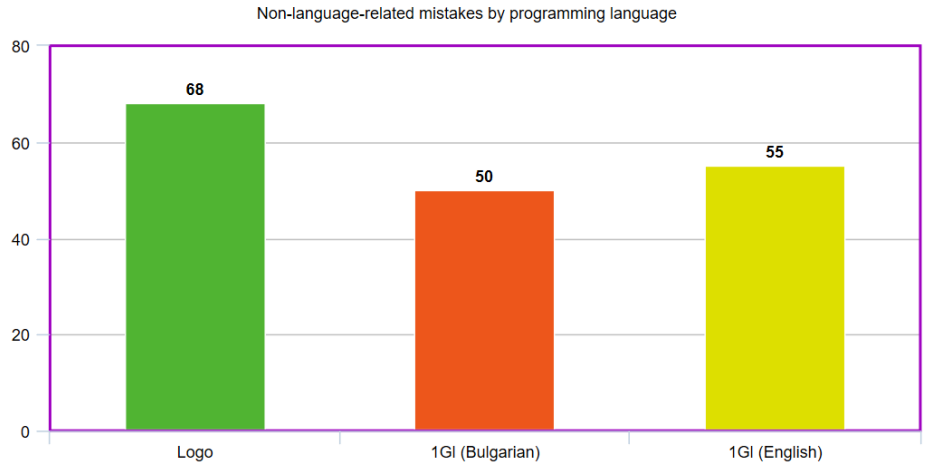


Figure 2: Number of non-language-related mistakes by programming language

- When asked about the most difficult exercise in their judgment, 7 students selected exercise 1, 8 students - exercise 2, and 5 students - exercise 3. Whilst this distribution is too even to welcome generalisations, one may consider the possibility that the fact that English-based 1Gl was voted as most difficult by one student more than Bulgarian-based 1Gl, even though the English-language exercise had the privilege of being very similar to the previous one, implies that Bulgarian-syntax programming tends to be perceived as easier.

  The explicit question of whether students found programming in Bulgarian easier than programming in English was met with 16 (or 80% of) positive answers. This result further supports the previously presented suggestion. Yet, possible external influences behind answers should be noted; for instance, that "yes" is generally a more ready answer than "no" for young students and that the word "Bulgarian" may by default be associated with a feeling of ease as compared with "English". To further test students' honesty and validity of judgment at answering the question, cross tabulation was performed, associating the answers provided with the number of language-based mistakes actually committed in the two versions of 1Gl (see Table 1).

	Higher No. of Mistakes			
Answer	**Bulgarian**	**English**	**Equal**	**Total**
No	2	2	0	4
Yes	3	11	2	16
Total	5	13	2	20

Table 1: Cross tabulation between the number of language-based mistakes in the two versions of 1Gl and students' answers to the question of whether they found programming in Bulgarian easier.

Only 19 % of respondents who answered in the affirmative made a higher number of linguistic mistakes in Bulgarian, pointing to general validity of the answers. Also, 50% of respondents who answered

"no" actually made more mistakes in Bulgarian (despite the overall lower number of mistakes in the language), showing that judgment was mostly adequate. In fact, these two students make up for a whole 40% of all respondents who made more mistakes in Bulgarian.

When asked the next question i.e. whether they deemed that programming in English could enhance their knowledge in the language, the majority of respondents (75%) answered positively. As thorough analysis of the truthfulness of this statement can only be achieved through a detailed temporal study, an assumption will need to be made that the general correctness of students' judgment in relation to the previous questions is also applicable here.

- Table 2 shows that only two students were faster in the English version of 1Gl than in the Bulgarian one. Notably, one of them was the single bilingual child (this result is explainable through the fact that English syntax is significantly shorter and, therefore, naturally easier to use in the presence of identical knowledge of the two languages).

Student No.	Exercise 1	Exercise 2	Exercise 3
1	11	9	8
2	15	15	10
3	5	8	5
4	8	8	8
5	13	15	12
6	10	15	10
7	5	5	5
8	18	20	15
9	14	16	11
10	6	8	10
11	3	3	3
12	9	8	7
13	10	12	10
14	20	25	15
15	4	5	3
16	7	9	10
17	17	18	13
18	6	6	5
19	11	19	12
20	8	10	10

Table 2: Time (in minutes) of completion for each exercise, as provided by students

The Logo exercise took longest to complete for only 2 (10%) of the participants. It shared the smallest number of minutes with one or both of the other exercises in another 7 of cases, and took shortest to complete

for a whole 10 or 50% of cases. It seems therefore safe to assume that the language Logo is faster to compose code in than 1Gl. However, such a difference in timing is to be expected given that one of the main ideas behind 1Gl is, as explained, its multidisciplinary nature. In other words, at using 1Gl, a student may take longer than at working with a classical beginner programming language, but they are using this time to simultaneously build skills in an increased number of academic disciplines (notably, linguistics and ESL).

3.3. Discussion

Coding in 1Gl may introduce types of mistakes that are not readily committed in languages with less natural and/or non-native syntax; notably, spelling mistakes and wrong assumptions concerning syntax. The former could be reduced with the involvement of language editing software and even simply through continuous practice with the language, coincidently leading to an improvement of students' general spelling skills. Wrongly assumed syntax is more difficult to address, and it comes as an established problem in relation to natural programming languages. For instance, within his study of natural-language programming in English, Buckman states that "[i]n most cases, such errors involve a child guessing at a command's name or the syntax of its arguments" (1999: 211). The problem may be addressed through an efficient system of error messaging as well as additional instruction concerning computer logic and limitations (notably, computers' inability to understand unedited everyday speech). It is also important to note that, partly making up for the newly introduced language-related mistakes, more traditional ones (such as skipped symbols or commands) are likely to be reduced due to the language's intuitiveness.

As seen, students are likely to have an initial preference for coding in Bulgarian. This tendency works toward proving the highly supported opinion that it is psychologically and practically beneficial for students to be able to use their native language within the field. It can, however, be expected that 1Gl's English version is to become increasingly more intuitive following regular practice with and association of the two syntax varieties.

The survey results show that the general time for the composition of code is higher in 1Gl than in a classical programming language. As mentioned, the time "lost" can be viewed as largely made up for, given the undelined expected acquisition of interdisciplinary skills during the programming process. Also, in the context of beginner computer studies, students are not to be encouraged to focus on their speed but to firstly ensure that code is correct and optimally structured.

3.4. Sources of Error and Suggestions for Further Work

Very importantly, not all major features of the proposed language, 1Gl, have been tested by the current survey. The main focus of the study is on natural language programming and multilingualism, and many syntax elements are left unaccounted for in order for exercises to remain simple enough for students with no programming background.

Another limitation of the study comes in the face of difficulties to differentiate between mistake types. For instance, if a student fails to colour the path taken by the actor, does this imply missing commands or a prior misunderstanding of instructions?

The study would be enriched by future involvement of a larger sample of respondents as well as by a wider age range (which would in turn imply more varied ESL skills, for instance). In the presence of a greater number of respondents, more elaborate statistical analyses such as a chi-square test would be applicable in the evaluation of relationships.

Finally, it should be noted that the Logo task has been perceived as rather simple by participants as compared with the other tasks. While this simplicity is not mainly due to the programming language in use (but to, for instance, similarity to the other two tasks and shorter commands), erroneous assumptions could be made in this direction. The selection of a slightly more complex task or of one slightly larger in size (for instance, the drawing of several adjacent figures instead of a single one) could improve the survey's accuracy.

4. Conclusion

Students are the main target group of users to rely on natural, human-like programming as well as to utlilise native (non-English) syntax. In particular, Bulgarian national education in its extending and increasingly early focus on computer studies is in need of unified and systematised programming practice. This project evaluated the potential benefits of a proposed programming language for beginner students with dual Bulgarian and English syntax. Ideally, the study will proceed to further development of the language's characteristics, accompanied with gradual analyses of its reception by Bulgarian students.

Ethical Consideration

Ethical approval has been granted by the associated higher education institution prior to the project's completion. All participants in the utilised survey are anonymous, and parents have agreed to the participation of their children in the project.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Ayvaz, H. (2018). Kompyutarno Modelirane v Uchilishte za Po-razchupeno Obrazovanie, *Bloomerang TV*. https://www.bloombergtv.bg/biznes-start/2018-02-21/kompyutarno-modelirane-v-uchilishte-za-po-razchupeno-obrazovanie.

Baron, G. et al (1985). Dix ans d'informatique dans l'enseignement secondaire : 1970-1980, *Institut national de recherche pédagogique.* http://lara.inist.fr/bitstream/handle/2332/1250/INRP_RP_81_113op.pdf?sequence=2.

Baudé, J. (2016). 'Le système LSE.' *EpiNet* (182): 41-56. http://www.societe-informatique-de-france.fr/wp-content/uploads/2015/12/1024-no7-Baude.pdf.

Bruckman, A. and Edwards, K. (1999). Should We Leverage Natural-Language Knowledge? An Analysis of User Errors in a Natural-Language-Style Programming Language, *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, Pittsburgh, Pennsylvania, USA, 15-20 May, pp. 207-214. https://doi.org/10.1145/302979.303040.

Bulgaria Ranks World's 20th in Internet Speed, Accessibility (2014) *Sofia News Agency*. https://www.novinite.com/articles/164134/

Dyulgerova, D. (2018). 'Predmeti kato Kompyutarno Modelirane i Informatsionni Tehnologii shte Razvivat Digitalnoto Obrazovanie v Stranata', *Focus News Agency*. http://www.focus-news.net/news/0000/00/00/2574674/.

Kaltinska, R. Nachalo na Informatikata v Balgariya (1959-1980). *Bulgarian Museum of Mathematics and Computer Science*. http://mmib.math.bas.bg/?page_id=5612.

Leonov, A. G. (2013). The Logical Design of Pedagogical Programming Systems, *Yaroslavskiy Pedagogicheskiy Vestnik*, 3(4):134-141.

Milanova, A. N. et al (2018). *Kompyutarno Modelirane za 3 Klas*. Sofia: Prosveta Plus.

Popitahme uchitelite! (2008). *Infoman.bg*. infoman.musala.com/files/view/static/anniversary/articles/AskTeachers.pdf/.

Myers, B. A. et al (2004). Natural Programming Languages and Environments. *Communications of the ACM*, 47(9):47-52. https://doi.org/10.1145/1015864.1015888.

Nikolov, A. (2013) Uchilishtnoto obrazovanie v Balgariya: Sastoyanie i Tendentsii. *Institute for Market Economics*. https://ime.bg/var/images/secondary_education_Adrian.pdf.

Nikolova-Stoupak, I. (2020) Appendices. In "A Natural Language for Bulgarian Primary and Secondary Education". *Computational Linguistics in Bulgaria 2020*. Sofia, Bulgaria, 25-26 June. https://www.kaggle.com/iglikastoupak/natural-language-for-bulgarian-education?select=Appendices.docx.

Regional Educational Management - Sofia. (2018) *Obuchenie po Kompyutarno Modelirane za Uchiteli* [Press release]. 6 November. http://ruo-sofia-grad.com/новини/обучение-по-компютърно-моделиране-за-учители.

Stefik, A. and Siebert, S. (2013). An Emprical Investigation into Programming Langauge Syntax, *Transactions on Computing Education (TOCE)*, 13(4):1-40. https://doi.org/10.1145/2534973.

Stoyanova, S. (2013). Mahat informatikata ot chasovete v uchilishte? *Dnes.bg*. https://www.dnes.bg/obshtestvo/2013/12/13/mahat-informatikata-ot-chasovete-v-uchilishte.209454.

Zamfirov, M. (2016). Sitoricheski predpostavki za vnedryavaneto I razvivaneto na obuchenieto po informatika v balgarskite uchilishta. *3rd Congress of Physical Sciences*. University of Sofia, Sofia, Bulgaria, 29 September -2 October, pp 47-50. www.trioiskar.com/hp/2016-12mzamfirov.pdf.

Digital Edition of the *Life of St. Petka*

Ivan Šimko
Slavic Seminary
University of Zurich
ivan.simko@uzh.ch

Abstract

This paper presents the construction of a digital edition of multiple versions of the hagiography of St. Petka of Tarnovo. Two related versions are uploaded at first: a Church Slavonic print edition and its later damaskini redaction. Both texts are adapted for user-friendly reading with side-by-side facsimiles. Translations and additional data concerning separate tokens and sentences can be shown up by the cursor on fly. Further metadata will be available for search. Annotation has been adapted for the transitionary status of the language of the texts: it allows us to compare similar morphological forms with various functions. The edition has already been published online and can be used for both teaching and studying.

The texts have been digitalized as a part of a larger project concerning the development of the Balkan areal features.

Keywords: linguistic annotation, corpus linguistics, text analysis

1. Introduction

Resources for quantitative research of the phenomena distinguishing Middle and Modern Bulgarian, which have been considered Balkan areal features (like e.g. postpositional definiteness marking, analytic infinitive or renarrative mood), are limited. The existing digital resources for study of Bulgarian contain primarily modern standardized variety (e.g. BNC, link), while Church Slavonic literature is online represented mostly by older sources (TITUS, link; Obdurodon, link). Furthermore, the requirements of their genre prevent us to observe the contemporary language shifts: medieval scribes did their best to preserve both the contents and the form of the original. The earliest texts, which show the phenomena mentioned above, can be dated back to the 17th century. They have attracted considerable attention by slavists (e.g. Petkanova-Toteva 1965, Demina 1968) and balkanists (e.g. Sonnenhauser 2016) as well. Although some of these texts have been digitalized (e.g. the damaskin of Loveč: link; cf. Mladenova & Velčeva 2013), users have often struggled with basic problems such as character encoding. No edition so far could satisfyingly combine the searchability of a digital corpus, philological exactness of a critical edition, and user-friendliness of a webpage.

To address these needs, we are developing a method to create a unified Balkan Slavic diachronic corpus, which will contain texts from various areas and epochs. Such a corpus will enable us not only to compare these texts with older varieties, but also with present-day dialectal data[1]. To demonstrate the capabilities of this method, we have created a webpage with two model texts processed in this way. This paper will explain the selection of sources and their processing into a simply accessible website.

[1] The paper has been written within the framework of the project *'Ill-bred sons', family and friends: tracing the multiple affiliations of Balkan Slavic*, led by Prof. Barbara Sonnenhauser of the University of Zurich, funded by the Swiss National Science Foundation (SNSF project grant IZRPZ0_177557/1), to whom I would like to express my thanks for support.

2. Source Texts

Both texts are based on the hagiography of St. Petka of Tarnovo (Parascheva of Epibates), written in the second half of the 14th century by Patriarch Euthymius of Bulgaria. This text is not only a precious artefact of the Middle Bulgarian literature, but it also played an important political role as well, being a crucial part of the cult of this ascetic saint, whose relics resided in the royal capital of Tarnovo. For a 17th-century reader (or audience), it preserved the memory of the former glory, kindling early sparks of national consciousness. In the Christian topology, St. Petka presented a figure of a universal ascetic role model; she personalized the idea that the sainthood can be attained by anyone.

The choice of text offers multiple advantages from the philological point of view. First, the differences between Church Slavonic and "simple Bulgarian"[2] of the damaskini are smaller than between Slavic and Greek. The translator could concentrate more on the styllistics than on translation itself. Thus, unusual grammatical constructions were less likely to appear because of the influence of a foreign language: they would more likely appear due to innovations in the target language. Second, there is a relatively large number of editions (so far 19 versions from the 16th to 19th century are known to the author) from various dialectal areas and epochs due to the popularity of the text, allowing us to use it as a reference, without interferences of the differences in content and genre.

The first version processed for our website is the shortened print edition by the monk Moses. The text was adapted to the Church Slavonic of the Resava redaction and published by Božidar Vuković in his *Traveller's miscellany*, first in 1521 in Venice. An online copy of the 1536 edition provided by the Matica Srpska in Novi Sad has been used for our purposes. As few pages of the copy were damaged, the missing words were completed according to other sources, including the related manuscript NBKM 665 of the National Library in Sofia[3], the modern critical edition based on Vuković's prints by Novaković (1877) or the edition based on multiple manuscript versions of Euthymius' original hagiography by Kałużniacki (1901). Vuković's print editions of the text likely reached the early damaskini circles.

First translations of the hagiography into "simple Bulgarian" appear in the 17th century (Kenanov 2009:59). These were often transcribed along the Church Slavonic version of the story. Curiously, the damaskin NBKM 709 from Sliven begins with the "simple Bulgarian" version, switching to Church Slavonic at the end. Our edition is from the Berlin damaskin, which was likely composed in 1803 in Pleven or Svištov (Ciaramella 1996). It is based on a later, 18th-century edition of the hagiography, reflecting a Moesian dialect. Unlike the earlier editions, this version does not end at the translation of St. Petka's relics by King John Asen II, but continues with their fate after the Ottoman conquest of Tarnovo[4], and in the end adds an original exegesis. An earlier fragment of this edition can also be found in the damaskin of the Church Archive in Sofia (CIAI) 133 from Pleven. The source copy was provided by the Jagiellon University of Cracow (signature Slav. fol. 36).

3. Processing

The texts were manually transcribed by the method developed for the diachronic corpus. As they are quite short (Vuković's edition: 2222 tokens, Berlin edition: 4852), use of automatic tools (like

[2] The term is based on the headings of the texts authored by Damaskēnos Stouditēs: *metaphrastheis eis tēn koinēn glōssan* 'translated into the common language' (e.g. Stouditēs 1751:5). Church Slavonic editions translate the phrase literally *ob'štymь ęzykomь*, while their translations into early modern Bulgarian use adjectives *prostymь* 'simple' or *bolgarskymь* 'Bulgarian'.

[3] NBKM 665 from Serbia includes both a synaxar- and the shortened panegyric (few pages are missing) version of the hagiography, as well as the liturgical service for the saint. Vuković's edition reflects the shortened panegyric *Life* from NBKM 665 or a similar handwritten source. Conev (1923:176) and recently Mineva (2005:5) date the manuscript already to the second half of the 15th century, at least a century before the emergence of the damaskini tradition.

[4] The source for this information was likely the printed *Menaion* of Demetrius of Rostov (1689), which was known among damaskini circles (cf. Kenanov 2009:59). This source included a new shortened edition, based on the original panegyric hagiography by Patriarch Euthymius (according to personal communication with Jürgen Fuchsbauer).

Transkribus, link) was unnecessary. For the sake of cross-platform compatibility, a set of Latin-based characters has been used, which is compatible with the UTF-8 format and which can easily be converted to the Cyrillic alphabet. The digital transcripts do not reflect graphemes serving rather ornamental functions (e.g. spirits). Broad initials (e.g. <ѻ> for /o/) and space-saving variants (e.g. the <*T*>-like character for /t/, adopted from the Greek cursive) were not distinguished; only *paerčik* (reflected as apostrophe <'>) and double gravis (<">, often representing a word-final /i/) are reflected in the transcript. Accent markers are mostly written with vowel characters together (e.g. *á*). If the accents are not fully compatible with the vowel character, they are written as separate characters following the vowel (e.g. *ę́* for a <ѧ> with an acute).

Additionally, the transcripts include auxiliary markers. The character <+> at the end of a token marks orthographic words: clusters of monosyllabics written together due to orthography, including articles and the negative particle *ne*. For example: ᾽иѳпѳстинїе́то 'and in the desert' is rendered as four separate tokens (*i+ u+ pustinïé+ to*). The marker <_> reflects the separation of a token over two lines or pages (e.g. *mnó_go* 'much'), or the separation of verbal or adjectival prefixes, which is common in the damaskini (e.g *zlató_juzdnïĭ konè* 'golden-bridled horses'). If the line break is marked in the text with a hyphen, the marker <-> is used instead of <_> (e.g. *pro-lét'nĭ* '[of the] spring'). Cyrillic numbers are marked with two asterisks <*> (e.g **a** '1'), as their actual marking in the source varies, but they do not occur in the two selected texts.

The texts are split into a table of tokens, which can be converted into an .xml file. A token possesses the following structure in the source table:

token	diplomatic	lemma	PoS tag	sentence id	syntactic position	syntactic dependency	dependency type
pétky	*petki*	*Petka*	NFSGY	1	5	4	NMOD

A diplomatic transcript is created for each token to simplify search queries, using a smaller set of Latin characters. In this layer, sets of graphemes representing the same phoneme are removed. For example, letters <и й і ї ы> are all reflected by specific characters in the first column, but diplomatized as *i* in the second. Accentuation, punctuation, and auxiliary markers are also removed (e.g. *i+u+pustinïé+to* > *i u pustinie to*). This layer makes the lemmatization easier, and helps us to train automatic recognition of morphological markers, which is relevant for further work on the diachronic corpus.

Lemmatization itself is based on various sources. The most specific dictionary, based on the Tixonravov damaskin (Demina et al. 2012), was supplemented by Church Slavonic (Miklosich 1865, Cejtlin 1994) and dialectal Bulgarian dictionaries (*Etymological Dictionary of BAN* 1972-2006). Lemmatization helps us to cope with orthographic differences, especially in texts radically adhering to phonetic principles and using non-Cyrillic scripts (e.g. *ζιδβέενιτδ* for *živenie+ to* 'the Life' in NBKM 1064), as well as in texts, which follow the orthographic norms only loosely (such as the writing of <ы> in the damaskini).

Individual tokens are annotated of morphological and syntactic features. The tag set has been customized not only to reflect the "simple Bulgarian" of the damaskini era, but also Church Slavonic. Morphological annotation (marked as "PoS tag" on the page) contains in most cases a single tag based on to the MultextEast standard. The tag set corresponds closely to the variant developed for Croatian (link). The tags include a marker for the part-of-speech category of the token (e.g. an N for a noun, A for an adjective etc.), which is followed by further information on number, tense, and other morphological categories.

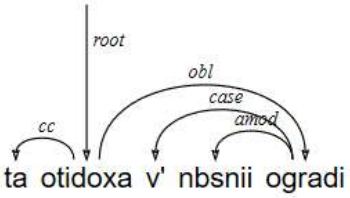
The new tag set was designed to limit the number of semantic and syntactic features. Thus a *da* would be tagged always as a "conjunction" (i.e. C), even if it serves as a marker of the infinitive or future tense, for such compounds are reflected by the syntactic annotation well enough. If the token bears features of two part-of-speech categories, it can receive an additional PoS tag ("alt.PoS tag" on the

page). For example, Church Slavonic verbal participles show features of both verbs (tense) and adjectives (gender, case). Thus the clause *kъ crstvoujuštomou prišьd'ši grádou* 'as [she] came to Tsarigrad' (Vuković) receives tags as follows:

diplomatic	lemma	PoS tag	alt. PoS tag
kъ	*k*	SD	
crstvujuštomou	*carstvuvati*	AMSDY	VMPP-S
prišьdьši	*priiti*	VMPA-S	AFSNN
gradou	*grad*	NMSDY	

Morphological ambiguities are resolved according to Church Slavonic inflection paradigms and phonological shifts. For example, an *-i* in *ja*-inflected words like *dši* 'soul'.DAT.SG is handled as a proper dative, and the token is tagged as NFSDY. In words of the *a*-inflection it is handled as a genitive marker, i.e. [*Žitie*] *Petki* 'Petka'.GEN.SG is tagged as NFSGY. If the same marker is used for multiple options within the same paradigm, the first option in the traditional order (e.g. N-G-D-A-V-L-I for nominal cases) is used. Thus, an *-i* in a phrase like [*ou*] *kašti* 'house'.LOC.SG (NBKM 328) is tagged as a "dative" (NFSDN). This procedure allows us to mark the shape of morphemes with ambiguous functions: e.g. a MASC.SG (*o-*/*jo*-inflection) word ending in an *-a* is always tagged as a "genitive" (e.g. *bga* 'God'.OBL.SG is tagged as NMSGY), even if the ending reflects a shortened article (e.g. *diavola se prestruvaše* 'the Devil was changing himself', Xrulev 1856) or a nominal count form (e.g. *četýri pъ́prišta* 'four shots', Tixon. d.). The PoS tag can be viewed on the website in a hover box, when moving the cursor over a word.

Syntactic annotation is based on the Universal Dependencies (link) scheme, which was designed in order to be applied to any human language. It works well enough for the transitionary varieties of early modern Bulgarian. The scheme is based on marking dependency relations (e.g. NMOD reflects a nominal modifier of the head: e.g. *pétky* in *pámétь pétky* 'remembrance of Petka'), using numbered syntactic positions within the range of a single sentence (denoted by the sentence id number). An extension layer was further added to provide closer information on the position of articles and demonstratives (e.g. P_NOM if they follow a noun), on the spatial relations denoted by an oblique modifier (e.g. LAT for the lative relation, LOC for locative etc.). For example, the dependency relations in the sentence *ta wtídoxa v' nbsny" ográdĭ* 'and they went into the heavenly gardens' (Berlin d.) can be visualized in the following way (using Arborator; link):



The scheme can be represented in an .xml table:

diplomatic	sentence	position	dep.	dep. type	dep. ext.
ta	19	1	2	CC	
otidoxa	19	2	0	ROOT	
vъ	19	3	5	CASE	
nbsnii	19	4	5	AMOD	
ogradi	19	5	2	OBL	LAT

While this layer of annotation is already used in the analyses of our corpus, it has not yet been integrated to the webpage. Only sentence numbers are shown in the text. The first token of the sentence also shows the translation of the whole sentence or of the following subordinate clause. Last

tokens of the line or page were marked with respective signs as well. First tokens of each page also contain the link to the original scan of the page.

An .xml table for both texts was generated by Excel. The .xml is transformed into an .html webpage by the Oxygen editor (link), using a customized stylesheet. Two versions were produced: one using the Cyrillic script, another one with the Latin transcription. After minor manual modifications, the webpages and scans of the originals were provisionally uploaded to the webspace provided by the University of Vienna. The website with both scans and the transcripts can be accessed via the following link - https://homepage.univie.ac.at/ivan.simko/. The description page (link) also contains source files (both in Excel and .xml format), as well as further information about the tag set and the stylesheet

4. Perspective

The existing webpage is very basic: it does not include any scripts, nor source data can be accessed by now. The next step will be the implementation of a visual representation of the syntactic annotation (e.g. using exported images from Arborator or a tabular diagram) and the development of the search widget, which would be able to process the annotation data, provided at the website. It is possible the website will be supplemented by additional pages reflecting other versions of the hagiography, given the facsimile are available to the author. These may include the print editions by Demetrius of Rostov (1689) and in Sophronius' *Nedělnik* (1806), as well as the well-preserved damaskini of Sliven (NBKM 709), Drjanovo (NBKM 711), Pop Punčo (NBKM 697) or Hadži Gendo (NBKM 1064) and other sources. In this way the website could find use both for teachers of Bulgarian literature to illustrate diachronic developments to students in a modern way, and for scholars studying these developments themselves.

References

Cejtlin, R.M., et al. (1994). *Staroslavjanskij slovar': po rukopisjam X-XI vekov*. Moskva: Russkij jazyk.

Ciaramella, R. (1996). Novi danni za Berlinskija damaskin. *Palaeobulgarica* XX (3), 120-129.

Conev, B. (1923). *Opis na slavjanskite răkopisi v sofijskata narodna biblioteka. Tom II*. Sofia: Narodna biblioteka.

Damaskēnos Stouditēs (1751). *Thēsauros Damaskinou tou Hypodiakonou kai Stouditou tou Thessalonikeōs*. Enetia: Nikolaos Glykeos.

Demetrius: Dmitrij ep. Rostovskij (1689) *Kniga žitij svjatyx*, 1. Kiev: Lavra Pečerskaja.

Demina, E.I. et al. (2012). *Rečnik na knižovnija bălgarski ezik na narodna osnova ot XVII vek*. Sofia: Valentin Trajanov.

Etymological Dictionary of BAN: Georgiev, V.I. ed., (1972-2006). *Bălgarski etimologičen rečnik, tom I-V*. Sofia: BAN.

Kałużniacki E. (hrsg., 1901). *Werke des Patriarchen von Bulgarien Euthymius (1375-1393)*. Wien: Carl Gerold's Sohn.

Kenanov, D.V. (2009). Žitija i služby na sv. Petka Tărnovska v staropečatnite slavjanski knigi. *Biblioteki. Četene. Komunikacii. Sedma nacionalna konferencija V. Tărnovo 21.-22.11.2008*. Veliko Tărnovo: UI, 57-65.

Miklosich, Fr. (ed., 1865). *Lexicon palaeoslovenico-graeco-latinum.* Vindobona: Guilelmus Braumueller.

Mineva E. (2005). *Pet ximnografski tvorbi za sv. Petka Tărnovska (XIII-XV v.)*. Academia.edu (link), 18.06.2020.

Novaković, St. (1877). *Život sv. Petke patrijarha bugarskoga Jeftimija*. Predano u sjednici filologičko-historičkoga razreda jug. ak. 30.5.1877. 48-59.

Petkanova-Toteva, D. (1965). *Damaskinite v bălgarskata literatura*. Sofia: BAN.

Velčeva, B. & O.M. Mladenova (2013). *Loveški damaskin: novobălgarski pametnik ot XVII vek*. Sofia: Nacionalna biblioteka "Sv.Sv. Kiril i Metodij".

Vuković, Božidar (1536). [*Traveller's miscellany*]. Venice. (link)

Sophronius: Sofronij ep. Vračanskij (1806). *Kyriakodromion sireč Nedělnik - Poučenie*. Râmnic: ep. Nektarij.

Sonnenhauser, B. (2016). "The Balkan Manner of Narration": Narrative Functions of the *l*-Periphrasis in Pre-Standardized Balkan Slavic. *Balkanistica* 29, 1-42.

Tixon.d.: Demina, E.I. (1968). *Tixonravovskij damaskin. Bolgarskij pamjatnik XVII v. Issledvanie i tekst, 1*. Sofia: BAN.

Used Manuscripts of the National Library in Sofia: NBKM 328, NBKM 665, NBKM 697 (Pop Punčo's miscellany), NBKM 709, NBKM 711 (Drjanovo B damaskin), NBKM 1064

Used Manuscripts of the Church Historical and Archive Institute in Sofia: CIAI 133.

Used Manuscripts of the Jagiellon University Library in Cracow: Slav. fol. 36 (Berlin damaskin).

Transcriptions from Cyrillic are based on Church Slavonic ISO 9 standard (link) with minor customizations (link).

All hyperlinks used in the text refer to the state of 18.06.2020.

SPECIAL SESSION ON WORDNETS AND ONTOLOGIES

A Bilingual Lexicosemantic Network of Bread Based on a Parallel Corpus

Ivan Derzhanski **Olena Siruk**
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
iad58g@gmail.com olebosi@gmail.com

Abstract

We present an experiment in using a corpus of Bulgarian and Ukrainian parallel texts for the automatised construction of a bilingual lexicosemantic network representing the semantic field of BREAD. We discuss the extraction of the relevant material from the corpus, the production of networks with varying parameters, some issues of the interpretation of these networks, and possible ways of making them more accurate and informative.

Keywords: bread, lexical semantics, semantic net, wordnet, parallel corpus, Bulgarian language, Ukrainian language

1. Introduction

Systems for thesaurus representation of vocabulary such as WordNet (Fellbaum, 1998) are widely used for storing lexical and semantic (including ontological) data. They typically are hierarchical networks which reflect synonymy by grouping lexemes into synsets, and other lexical and semantic relations by labelled directed edges. The knowledge encoded in them can come from various sources. Resources of this type have been developed for many languages.

In this study we present an experiment in the automatised construction of a bilingual lexicosemantic network on the basis of a corpus of parallel texts. Our working languages are Bulgarian and Ukrainian. They are related, though not nearest of kin, spoken in regions neither adjacent nor really distant, moderately close typologically, and with similar history of substantial lexical borrowing from Western European and other languages.

We chose to focus our attention on the field of BREAD. The concept of bread is extremely important in Western civilisation: as bread and other bread-like products have been baked and eaten for millennia, bread vocabulary is highly developed everywhere; on the other hand, for many centuries it has been developing separately, making for very complex and interesting relationships between words of different languages. Besides, bread and bread-like goods are produced by people and (mostly) for people; as such this semantic and lexical field is part of the anthropocentric image of the world and the anthropocentric vocabulary, which forms an ideal basis for setting and solving any general linguistic problems. The apparatus of theoretical notions and technical approaches developed on such material has the best chances of being extrapolated to other nominal lexicosemantic fields.

2. The Corpus

The bilingual Bulgarian–Ukrainian corpus (CUB) (Derzhanski and Siruk, 2019) consists of parallel texts available in electronic libraries or obtained by us from paper editions through scanning, optical character recognition and error correction by *ad hoc* software tools and by hand. For this reason the corpus is composed of fictional works, mostly of novels, which dominate in such sources.

Because original and translated parallel texts for Ukrainian and Bulgarian are hard to come by, especially in online-accessible computer-readable form, we also use Bulgarian and Ukrainian literary translations from other languages as corpus material. The current version of CUB includes eleven sectors, each of which covers parallel Bulgarian and Ukrainian texts with the same original language:

- original Bulgarian and Ukrainian texts, as well as translations from English-1 (by authors from the British Isles), English-2 (by authors from the United States), French, German, Italian, Russian-1 (stories about the past and present), Russian-2 (stories about the future), and French—approx. 2 million words in each of the ten sectors (in the two corpus languages counted together; for various reasons the ratio tends to be about 53:47);
- the Bible, in canonical translations from Hebrew, Aramaic and Greek into Bulgarian and Ukrainian —1⅓ million words.

The total size of CUB is 10 million words in Ukrainian (and 11½ million in Bulgarian). The Bible is aligned by verse, and the other texts (mostly) by sentence.

3. Data Collection and Preliminary Analysis

The collection of the material took place in the following way. We started with the two languages' principal 'bread' words, Bg *хляб* and Uk *хліб*, and their near-synonym derivatives Bg *хлебец* '(a little) bread (hypocoristic)', *хлебче* 'small bread, roll', Uk *хлібець* 'little bread (dimin. or hypocor.)', *хлібина*[1] 'loaf of bread'. All occurrences of these words in CUB were located. The numbers are summarised in Table 1, separating the singular and the plural (and, in Bulgarian, count) forms of the words *хляб*/*хліб*, as well as the cases when they cooccur with a specification of quantity such as 'a loaf', 'a piece' or '200g'. The label 'others' means other words for bread-like substances, on which more anon:

Bg \ Uk	*хліб* (sg.)	(Q) *хліба*	*хліб* (pl.)	*хлібина*	*хлібець*	others	∑
хляб (sg.)	1111	53	6	13	5	61	1249
(Q) *хляб*	22	153	0	0	0	5	180
хляб (count)	8	3	18	5	0	8	42
хляб (pl.)	22	1	18	3	0	18	62
хлебец	2	1	0	1	4	0	8
хлебче	0	0	0	0	8	41	49
others	22	3	0	16	3		44
∑	1187	214	42	38	20	133	1634

Table 1: Occurrences of correspondences of *хляб*/*хліб* and cognate words

We see that Bg *хляб* is somewhat more readily used as a count noun and Uk *хліб* as a mass one, that it is more common for Bg *хляб* than for Uk *хліб* to correspond to a word with a different root, and that for Bg *хлебче* and Uk *хлібина* this is even more common.

At the next stage of the research the 'other' corresponding words were similarly sought out, then the correspondences of their correspondences, and so on until no new words were found. Only words denoting kinds of cooked dough (baked, boiled or fried) or their parts, products or subproducts – but not just dough, flour, grain or gruel – were considered. Words meaning 'piece (of anything)' were included only when it was clear that bread or another relevant substance was meant (and usually mentioned in the same sentence). Where in serious doubt, we preferred to err on the side of inclusion. For example, while both Bg *пастет* and Uk *паштет* normally mean 'meat or fish paste', they do occur in the meaning of 'pie, pâté (with a crust)', especially in translations from languages where the related word (De *Pastete*, Fr *pâté*, Ru *паштет*) has this as one of its regular meanings; accordingly we took all cases where Bg *пастет* or Uk *паштет* corresponds to a word for a bakery product in the other language, as well as the occurrences of the pair *пастет* : *паштет* from *The Black Obelisk* by E. M. Remarque, where the *Leberpastete* 'liver pâté' is cut in slices and eaten as a dish in its own right, which suggests a crust. Uk *товченик* 'fish or meat dumpling' (Hrinchenko, 1958: v. 4, 270), where flour is an ingredient but not a major one, is on the fringe; its entry in *The Geese and Swans Are Flying* by M. Stelmakh was taken because the Bulgarian translation is *тиганица* 'pancake'. But we stopped short of including Uk *зефір*

[1] A Bulgarian counterpart of this word (*хлъбина*, *хлъбинка*) was in use as late as the 19th c. and is registered in N. Gerov's dictionary (Panchev, 1904: 501; Panchev, 1908: 320), but has been lost in the contemporary language.

'marshmallow', which is what is left of *marshmallow cookies* in I. Shaw's *Bread Upon the Waters* (contrariwise, in the Bulgarian translation of the book only *бисквити* 'cookies' is preserved).

Several cases of idiomatic use of bread words whose literal meaning was not very close were discarded: the sentence pair *Със сух комат да се задавиш дано* || *Бодай ти немащеним млинцем удавився* 'May you choke on a dry chunk of bread || an unbuttered pancake' (V. Shishkov, *Gloomy River*) was not considered an occasion to postulate a word pair *комат* : *млинець*. Exceptionally a word was counted twice if it had two counterparts on the other side: the pair of sentences *Има ли — попитал той — такъв човек под слънцето, който не би предпочел пшеничната питка пред ечемичната?…* || *Чи ж є, — спитав він, — така людина під сонцем, яка б, маючи ячний коржс*$_1$*, не хотіла б пшеничної паляниці*$_2$*?* "Is there,' he asked, 'any man under the sun who would not prefer a wheat roll$_2$ over of a barley cake$_1$?" (B. Prus, *Pharaoh*; tr. into English by Christopher Kasparek) yielded the two word pairs *питка* : *корж* and *питка* : *паляниця*. Adjectives derived from the words *хляб* and *хліб* were not taken,[2] but from other words they were: the pair of phrases *тоя козуначен Казанова* || *цей містечковий женолюб* 'old Pastry-Casanova' (E. M. Remarque, *Three Comrades*; tr. into English by A.W. Wheen) produced the word pair *козунак* : *містечко*.

The total number of word pairs thus found was 3240.

As is typical in Slavic languages, many Bulgarian and Ukrainian bread words have diminutives (and rediminutives), which sometimes acquire different meanings—an extreme case is Ukrainian *батончик* 'stick of confectionery, candy bar'[3] from *батон* 'long bread loaf'. To avoid making a judgement in each case, we considered lexemes that only differ in diminution as separate if each appeared three or more times (e. g., Bg *кора* 2, *корица* 1 and *коричка* 16 'crust' were all counted jointly).

In Bulgarian the words *кора*$_1$ 'crust (of bread)' and *кора*$_2$ 'sheet of phyllo pastry', as well as *пита*$_1$ 'round bread' and *пита*$_2$ 'fruit pie', which have diverged semantically to a considerable degree, were considered different lexemes. In Ukrainian *сандвич* and *сендвич* 'sandwich' were counted together, as were *баніца* and *баниця* < Bg *баница*[4] '(Bulgarian) layered pastry', being simply different adaptations of the same foreign (and infrequent) words.

In all 91 Bulgarian and 110 Ukrainian lexemes were found.[5] They are listed in Appendices A and B, along with the number of their occurrences in our data, all in citation form, which is the plural in two cases (Bg *равиоли* 'ravioli' and Uk *потапці* 'dunked bread'). The vast majority denote baker's goods specified for shape (elongated, round, crescent), size, grain (oats, rye), presence of leaven, taste (savoury, sweet), presence of a topping or filling, etc.

A few words are exceedingly rare or appear in rare forms. Bg *книш* can hardly be called a regular word of the language, but appears as a rendering of the identical Ukrainian word in P. Zahrebelnyi's novel *Let's Come to Love*, where it couldn't have been translated because it happens to be a character's surname. Bg *тако* 'taco' appears in R. Bradbury's *Death is a Lonely Business* in the un-Bulgarian plural form *такос*, which indicates that it has not (yet) been adopted. Uk *гостія* 'host, Catholic Eucharist wafer' is not registered by SUM, though found in some dictionaries of loanwords (e. g., Bojkiv et al., 1955: 114). Uk *ощіпок* 'flat unleavened bread' (Hrinchenko, 1958: v. 3, 84) is an uncommon variant of *ощипок* (SUM, 1974: v. 5, 840). Uk *пляцинда* 'sweet pie' is found in Hrinchenko (1958: v. 3, 199), but SUM (1975: v. 6, 572) knows only the variant *плачинда*. Uk *попряничок*, variant of *пряник* 'gingerbread', is a *hapax legomenon* in M. Lukash's translation of *The Decameron* by G. Boccaccio, but still registered by SUM-20. Uk *метерка* appears in our data as a diminutive of *метеря* 'bread soup, sop', but SUM (1979: v. 10, 102) only registers the homonym *метерка* 'greyhen'.

Let us formulate several semantic categories with the aid of the two languages' main interpreting dictionaries, RBE and SUM, and illustrations chosen among the words found in CUB:

I. bread proper: *хляб* and *хліб* themselves, as well as the hypocoristic Bg *хлебец*;

II. a whole unit (i. e., loaf), able to appear in combination with *хляб*/*хліб* (Bg *пита* and *самун*; Uk only *буханець*);

[2] On two occasions an exception is made for the Ukrainian adjective, which obviously corresponds to the Bulgarian noun in *коричка хляб* : *хлібна скоринка* 'bread crust' and *трошичка хляб* : *хлібна крихта* 'bread crumb'.

[3] Used for 'Nestlé's Crunch bar' in the translation of R. Bradbury's *Death Is a Lonely Business*. We discard this word when its match is Bg *шоколадче* 'chocolate bar', but when it is *вафла* 'wafer', we take the pair for want of certainty that we should not: a Nestlé Crunch (despite containing crisped rice) is not really a bread product, but a *батончик* in general can be.

[4] This word actually has a Ukrainian version in the person of *баник*, interpreted as *родватрушки* 'a kind of cheesecake' (Hrinchenko, 1958: v. 1, 26), but absent from SUM.

[5] The idea that the set of words had to be fully connected had the corollary that some semantically appropriate words were not taken: such were Bg *аньолоти* 'agnolotti' and *пелмени* '(Russian) dumplings' and Uk *пельмені*, which only correspond to one another in the corpus.

III. a piece of bread or other baker's goods (Bg *комат*, *кора*$_1$, *крайче*, *крайщник*, *къшей*, *порязаница*, *среда*, *троха*, *филия*[6]; Uk *м'якуш*[7], *м'якушка*, *окраєць*, *скорина*, *шкуринка*);

IV. a piece of food, or of anything but usually bread (Bg *залък*, *резен*, *хапка*; Uk *крихта*, *кришка*, *скиба*);

V. a piece of anything (Bg *къс*, *парче*; Uk *кавалок*, *кусень*, *кус(ок)*, *луста*, *шмат(ок)*);

VI. a whole item, except for those accounted for under II.

The correspondences look as shown in Table 1, in which, when two values symmetrical with respect to the main diagonal differ significantly (e. g., Bg words for pieces of bread correspond to Uk words for pieces of food 74 times, whilst the opposite happens only 3 times), the greater one is in boldface and the smaller one in italics:

Bg \ Uk	I	II	III	IV	V	VI	∑
I	1418	**13**	*3*	0	0	**107**	**1541**
II	*3*	15	0	1	0	**49**	**68**
III	**6**	1	36	**74**	**44**	**44**	**205**
IV	2	0	*3*	11	**31**	**4**	*51*
V	0	3	*6*	*13*	70	1	*93*
VI	*13*	*6*	*1*	*0*	3	1259	*1282*
∑	*1442*	*38*	*49*	**99**	**148**	**1464**	3240

Table 2: Correspondences between the semantic categories

From this table one sees that Bulgarian has a greater fondness of the word *хляб* and the terms for units and pieces of bread big and small, and Ukrainian of special terms for baker's goods and general words for pieces of food or of any stuff.[8]

Finally, let us look at the individual lexemes. Table 3 shows the upper left corner of the correspondence table sorted by the overall frequency of the words in our data, that is, featuring the most frequent words. The values which are greatest in both their rows and their columns are in boldface:

	хліб	*бутерброд*	*торт*	*пиріг*	*корж*	*шматок*	*тістечко*	*булочка*	*сухар*	*печиво*	…	∑
хляб	**1415**	6	——	——	8	——	——	——	1	1	…	1533
сандвич	1	**117**	——	——	——	——	1	1	——	——	…	171
питка	2	——	——	1	**62**	1	——	——	——	1	…	139
торта	——	——	**104**	6	——	——	3	——	——	2	…	116
сладкиш	——	——	10	19	——	——	**31**	2	——	16	…	96
парче	——	——	——	——	——	**44**	——	——	——	——	…	71
сухар	——	——	——	——	2	——	——	——	**54**	——	…	63
баница	——	——	——	**39**	——	——	——	——	——	7	…	60
филия	4	4	——	——	——	3	——	——	——	——	…	53
кифла	——	1	——	1	——	——	——	**38**	——	——	…	51
…	…	…	…	…	…	…	…	…	…	…	…	…
∑	1442	137	118	116	83	81	73	64	60	56	…	3240

Table 3: Correspondences between the most frequent lexemes

[6] This word is not in fact restricted to bread (there is *филийка салам* 'slice of sausage', for example), but bread is always understood unless something else is specified.

[7] The dictionary interprets this word as "*Підшкірна частина плодів, ягід тощо*" 'The part under the skin of fruit, berries etc.' (SUM 1973: 838), but in the corpus it behaves in the same way as *м'якушка*.

[8] There is another circumstance which is outwith our scope here but still relevant: Uk *хліб* is more readily used in the meaning of 'grain'.

Uk *хліб* is found 27 times outside of the most frequent pair *хляб* : *хліб* (and no other Bulgarian word corresponds to it more than 4 times), whereas Bg *хляб* occurs 118 times (among its Ukrainian correspondences are *хлібина* 'loaf of bread' 23, *опріснок* 'unleavened bread' 20, *буханець* 'loaf of bread' 13, *грінка* 'toast' 10). Also, Bg *филия* 'slice of bread' is the 9th most frequent word in the list, whilst its most frequent Uk counterpart *скибка* is only the 18th.

4. A Bilingual Lexicosemantic Network

We can build a lexicosemantic network on the basis of the correspondence table, ignoring the numbers and drawing an edge between a Bulgarian and a Ukrainian node if there is at least one match. Figure 1 presents such a network.[9] The triangles and the stars are the Bulgarian and Ukrainian words, respectively. In the centre is Bg *питка* 'bread roll'; from this word any other can be reached in five moves at most (the farthest ones being Uk *батончик*, *луста*, *м'якушка*, *тортик*, *трубочка* and *шкуринка*). Other choices might have been Bg *хляб* and Uk *калач*. The number of edges is 395.

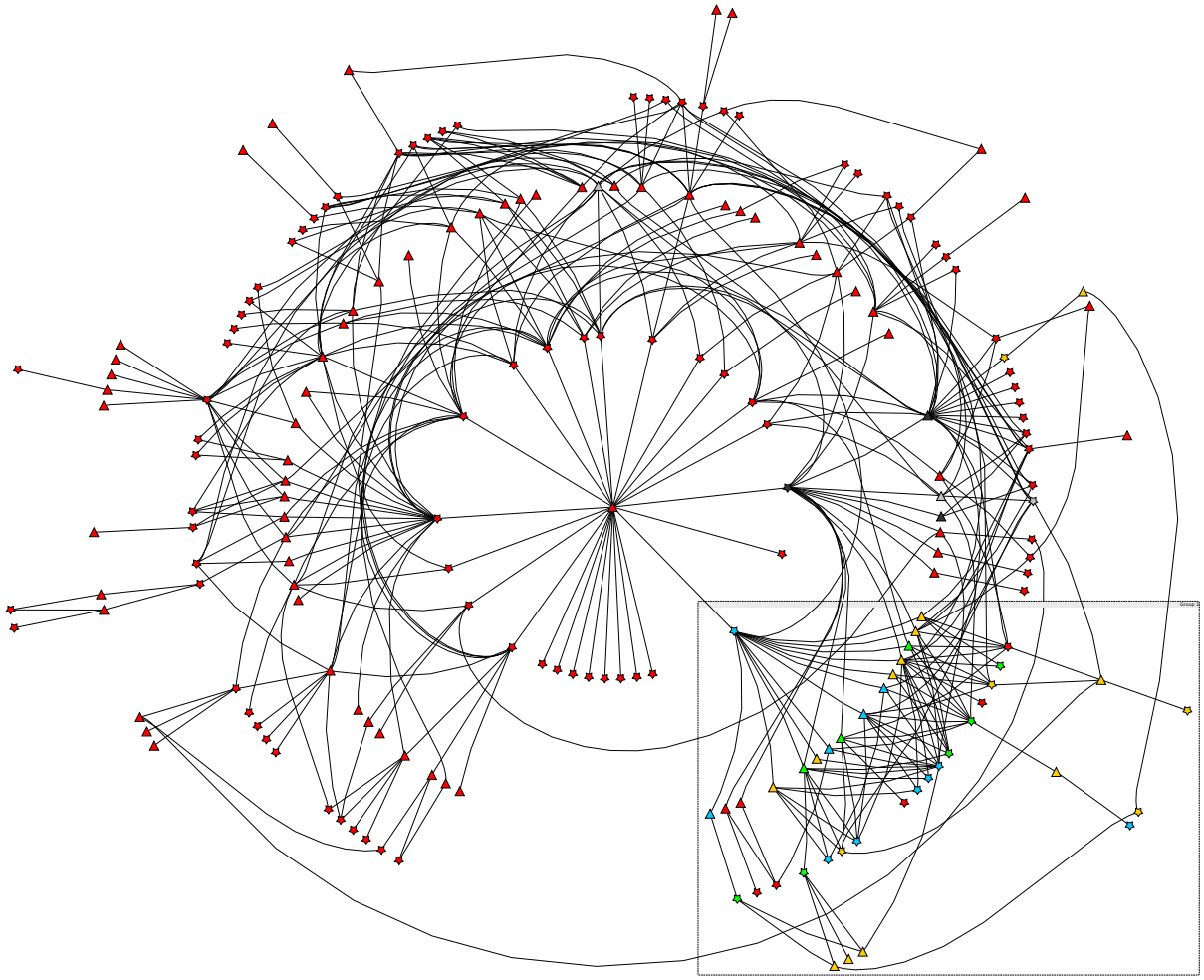


Figure 1: The unabridged lexicosemantic network of bread

The greatest distance between two nodes is 9 edges; such a one separates Bg *кекс* 'sweet cake, cupcake' and Uk *луста* 'piece', or Bg *маршал* 'Marshall cake' and Uk *батончик* 'stick of confectionery, candy bar' and *трубочка* 'puff', or Bg *троха*, *трохица/трошица* and *трохичка/трошичка* 'crumb' and Uk *батончик* and *трубочка* again. Uk *батон* and *батончик* are separated by 6 edges; Uk *калач* and *калачик*, *сухарик* and *сухарець*, *торт* and *тортик* by 4; the remaining Ukrainian pairs of cognate

[9] The figures are drawn with the graph editor yEd (`https://www.yworks.com/products/yed`).

words, as well as all Bulgarian ones, are 2 edges apart, which here means that there is a word in the other corpus language to which each corresponds at least once.

We would draw attention to the lower right corner of the image, connected to the rest of the graph through Uk *шматок*, where almost all piece words are located, with the exception of Uk *м'якуш* 'soft inner part of bread' (adjacent to Bg *хляб* in the graph) and Bg *среда* 'middle, inner part (incl. of bread)', but with the addition – for reasons of geometry – of Bg *нафора* and *просфора* and Uk *гостія* and *облатка* 'communion wafer', as well as Uk *глевтяк* 'underbaked bread' (an infrequent word which only occurs once in CUB, in Oles Honchar's *Guide-on Bearers*, in the plural and with the meaning 'chunks of ~', and this has prompted its being translated into Bulgarian as *клисави хапки* 'sodden mouthfuls'), *грінка* (in Bulgarian mostly *препечена филийка* 'toasted slice of bread') and *тост* 'toast'.

Conceivably some edges are an artefact of translation from third languages and actually connect words with substantially different meanings. The network in Figure 2 is limited to the 764 word pairs found in texts where Bulgarian or Ukrainian is the original language (somewhat less than ¼ of the 3240 in the whole corpus). It features 58 Bulgarian and 69 Ukrainian lexemes (just over ½ of all) and has 141 edges. (The most frequent pair from the big corpus which is missing here is *сладкиш* : *тістечко*, followed by *хляб* : *опріснок* from the Bible.) In the centre is Uk *калач*, which is 8 edges away from the farthest nodes, Uk *скоринка* and *шкуринка* (at the right) and *торт* (in the lower left). The system is no longer fully interconnected; 13 Bulgarian and 10 Ukrainian terms are not linked to the rest. They are in the upper right corner of the figure, from right to left: Bg *кифла* : (*рогалик* : (*кифличка*, *кроасан*), *книш* : *книш*, *булочка*), Uk *крихта* : (*троха*, *трошиця*), Uk *марципан* : (*бадемовка*, *марципан*), Bg *трохичка* : *кришка*, Bg *попара* : *тетеря*, Bg *порязаница* : *луста*, Bg *кекс* : *кекс*, Bg *сухар* : *сухар*.

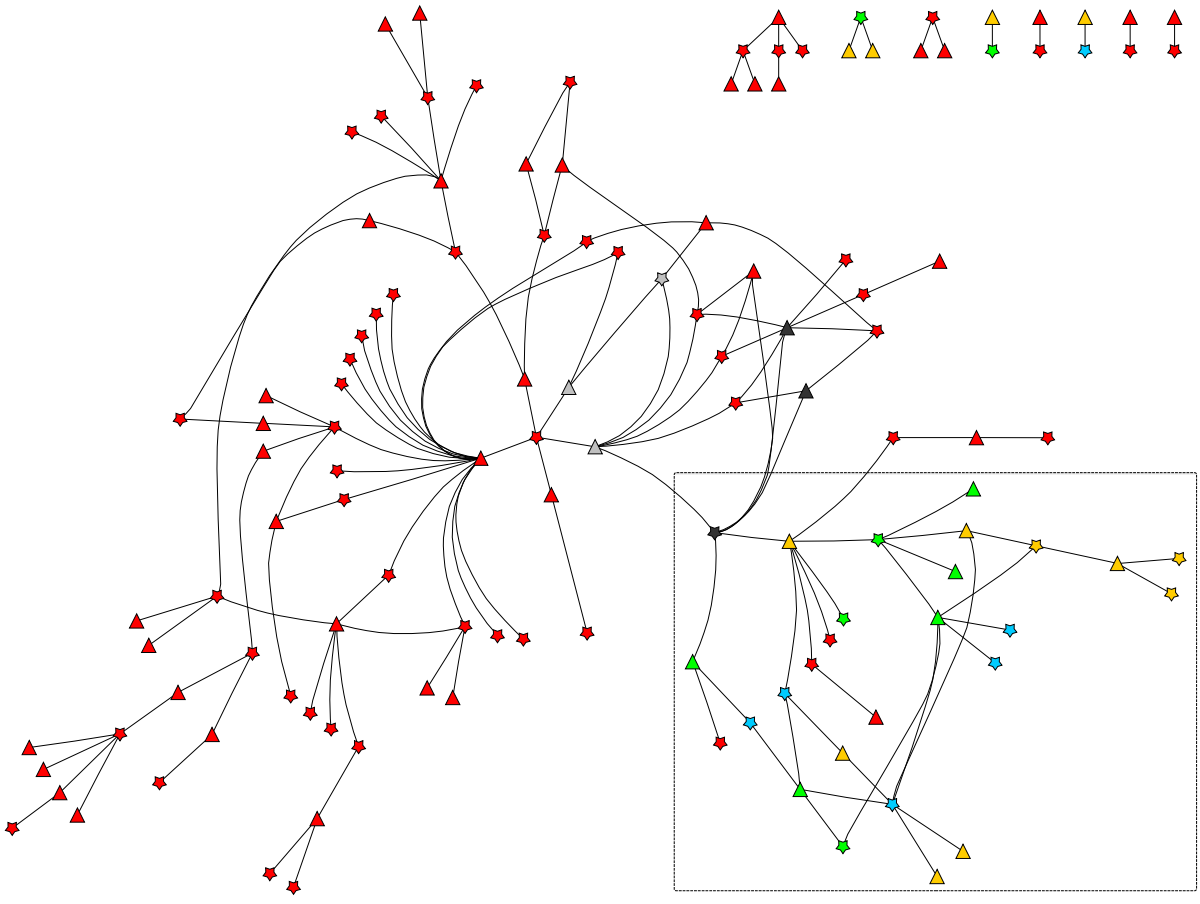


Figure 2: A network based on texts with Bulgarian or Ukrainian originals

Among the pairs in CUB some are extremely rare, such as *сандвич* : *хліб* or *питка* : *пиріг*, which only occur one time each and thus obviously have little to tell us about the highly frequent words that

compose them. Other pairs are more telling, e. g., *хамбъргър*[10] : *гамбургер*: it appears 6 times, accounting for all occurrences of the Bulgarian word and all but one (a sole *сандвич* : *гамбургер*, which ensures the contact of this pair to the rest of the system) of the Ukrainian. This suggests that a way of making the network more informative would involve labelling every edge by the number of occurrences of the corresponding word pair or another numeric value reflecting the relevance of the edge. Another way would be to make the edges directed by having them point from the less to the more frequent lexeme, e. g., from Uk *гамбургер* (7) to Bg *сандвич* (171) and from the latter to Uk *хліб* (1442); such edges will often be interpretable as indicating an 'is a kind of' relation.

The network can also be made more observable by excluding part of the correspondences, choosing them among the ones that are least well supported by corpus data. The one in Figure 3 has been made with the condition that if two word pairs share a word and one is more frequent than the other, we may not drop the first and keep the second.

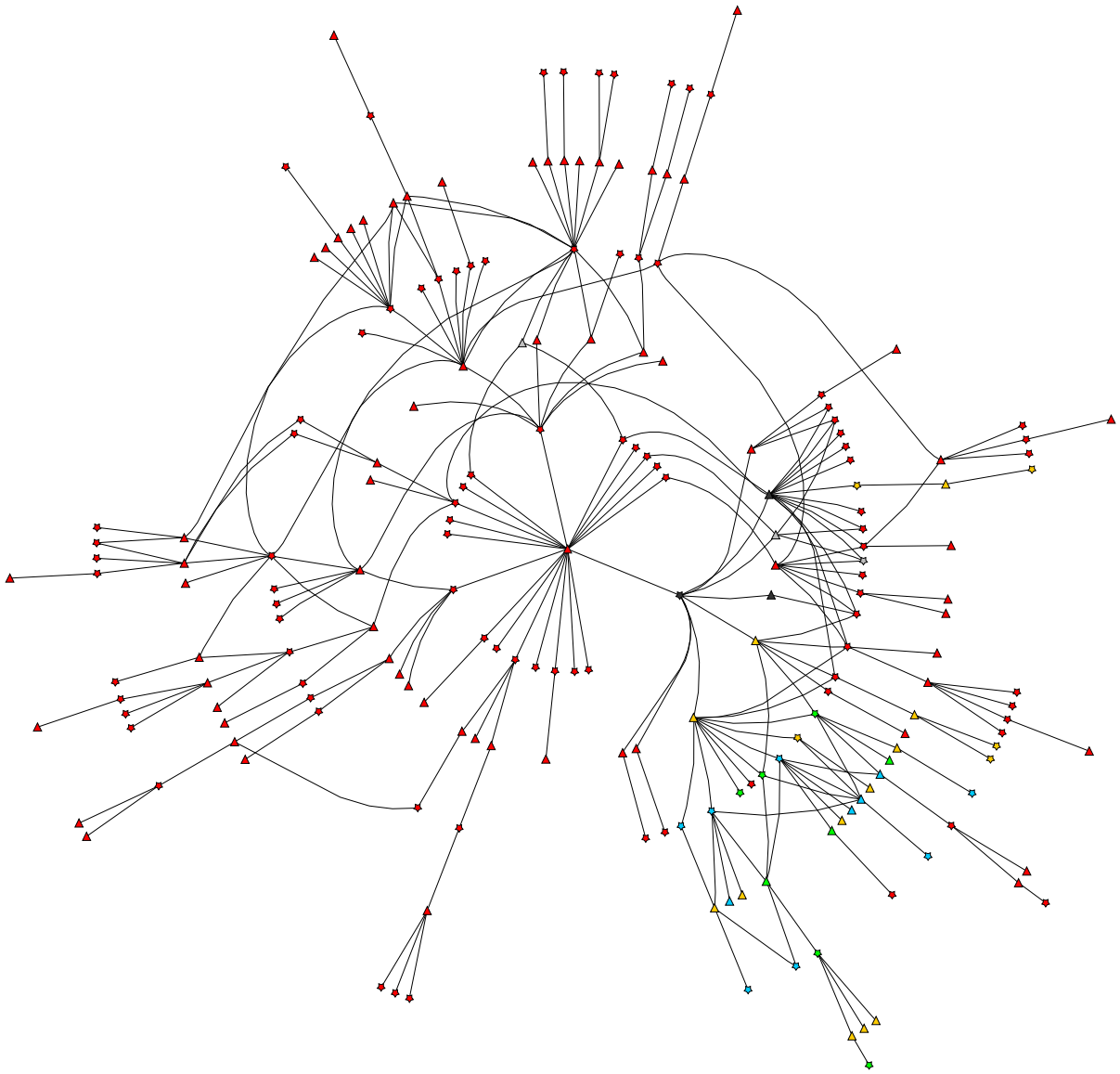


Figure 3: An abridged lexicosemantic network of bread

[10] Thus in the source (a 1983 translation of E. M. Remarque's novel *Shadows in Paradise*); remarkably, the contemporary language has settled for the form *хамбургер*, which is a less precise rendering of the English pronunciation.

The graph has 251 edges, and no further edge can be removed while preserving both this feature and the integrity of the system. Here, too, the central position is held by Bg *питка*, which is seven edges away from the most distant nodes, Uk *гостія* and *кришка*. (Another option would have been Uk *хліб*.) The greatest distance here is 13 edges, as between Uk *гостія* or *кришка* and Bg *вареник*, *геврек*, *кекс*, *равиоли* or *соленка*.

We can also relax the requirement that the edges from each node must be chosen among the most frequent ones, and take only as many edges as are needed to keep the system fully connected (namely 200, one fewer than the nodes). Such a minimal network is shown in Figure 4. The graph is drawn as a tree, but this should not be taken as implying a hierarchy, since no edges are directed; Bg *питка* is the top node not because it is a root, but because in this graph, as in the other two, it is central: from it all other nodes can be reached in no more than 14 moves, the most distant ones (lowermost in the picture) being Bg *вареник*, *равиоли* and *бухта*. (The other candidate for the central position is Uk *корж*, 14 edges away from Uk *гостія*. In the picture these are the 'root' of the right-hand subtree and the lowermost 'leaf' of the subtree on the left, respectively.)

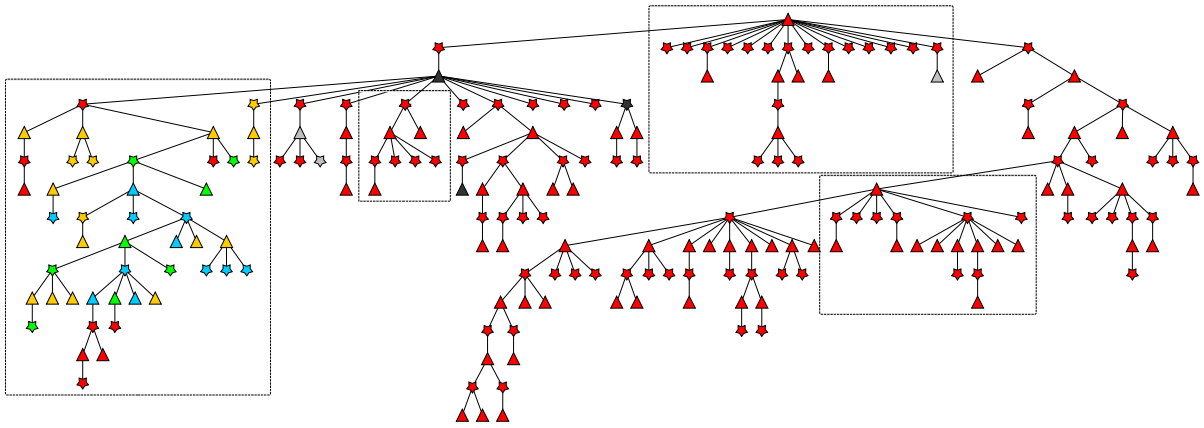


Figure 4: A minimised lexicosemantic network of bread

The largest distance, 27 edges, separates Bg *вареник*, *равиоли* and *бухта* from Uk *гостія*. The pair of cognate words that are farthest apart are Bg *кифла* 'bun' and *кифличка*; there are 16 edges between them. Between Uk *батон* and *батончик* there are 14 edges, as between Bg *сухар* 'rusk' and *сухарче*, and there are 12 between Uk *сухар* or *сухарець* on one hand and *сухарик* on the other.

What is especially intriguing about this tree is that its subtrees contain semantically well-formed subsets of the bread lexicon of the two languages. There are two big subtrees, one on the left headed by Uk *калач* (91 nodes) and one on the right headed by Uk *корж* (85 nodes), and as a general rule the words for types and quantities of bread are found in the former and for more complex products of bakery and confectionery in the latter. Here again there is a domain where the piece words are concentrated; another, including the centre, with words for types of bread by content (25 nodes); and in the subtrees smaller semantic areas can be recognised, for example sandwiches (in the left-hand tree, headed by Uk *бутерброд*) or desserts (in the right-hand one, headed by Bg *сладкиш*). In the figure these areas are highlighted by boxes.

A further refinement of the networks may involve a correction to the weight of an edge based on the number of sectors or texts in which it is encountered. This would reduce the impact of pairs which occur multiple times but are the handiwork of a single translator, such as Bg *курабия* : Uk *корж*, seen 7 times in Mykhailo Stelmakh's novel *The Four Fords* and nowhere else.[11]

5. Conclusions

We would highlight some valuable traits of the proposed approach:

[11] Such cases are rare. Their origin – oversight, *faux ami*, influence of a third language, pursuit of pragmatic rather than semantic equivalence, etc. – is of definite interest and merits special study; but in the context of constructing an adequate bilingual lexicosemantic network they demonstrate the desirability of manual edition involving expertise in both languages.

1. formalisation: the sequence of actions is precisely outlined, logically substantiated and carried out;
2. universality: the procedure can be applied to any field defined by a certain concept;
3. relevance: the research base is grounded in both the generalised translation experience embodied in the parallel texts collected in CUB and the fruits of the interpretative lexicography of the two languages embodied in the respective dictionaries;
4. objectivity: the exact rules of action (the formalisation and automation of the procedure) contribute to reducing the subjective component in the linguistic research as much as possible;
5. comprehensive coverage: the multilinguality of the sources (the presence of parallel translations from third languages in CUB) increases the diversity of detected entries;
6. 'double hit': the use of a parallel corpus allows building a network on the basis of two languages and for both languages simultaneously, transgressing the boundaries of translation from a source language to a target language.

It should be kept in mind that this method will not always be able to reveal all entries of a particular field on its own. For this task it is better to use a stepwise method of vocabulary identification with substantial use of dictionary interpretations.[12] But when it comes to finding new meanings or even entries not registered by explanatory dictionaries, the use of the parallel corpus method can give interesting results.

Further steps and directions in the development of the research can include an overlay of the obtained network of meanings on similar networks built by other (deductive and inductive) methods in order to compare their coverage, as well as combining the results, so as to obtain a more complete and structured overall system of meanings for each language.

References

Bojkiv, I., Izjumov, O., Kalyshevs'kyj, H., and Trokhymenko, M. (1955). *Slovnyk chuzhomovnykh sliv*. New York: M. Borets'kyj.

Derzhanski, I. and Siruk, O. (2019). The Intensifying Prefix *pre-* in a Corpus of Bulgarian and Ukrainian Parallel Texts. In *Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 9.* Sofia: Institute of Mathematics and Informatics—BAS, pages 177–188. http://dipp.math.bas.bg/images/2019/177-188_12_2.10_fDiPP2019-67_f_v.1a.F_20190908.pdf.

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

Hrinchenko, B. (1958). *Slovar' ukrajins'koji movy*. Kyjiv: Vyd–vo Akademiji nauk Ukrajins'koji RSR. http://hrinchenko.com.

Panchev, T., Ed. (1904). *Rechnik na bălgarskij ezyk s tălkuvanie rechite na bălgarski i na ruski. Săbral i iztălkuval Najden Gerov. Chast 5: R–Ja.* Plovdiv: "Săglasie".

Panchev, T., Ed. (1908). *Dopălnenie na bălgarskija rechnik ot N. Gerov.* Plovdiv: "Trud".

RBE (1977—). *Rechnik na bălgarskija ezik*. Sofiya: Izdatelstvo na BAN "Prof. Marin Drinov". http://ibl.bas.bg/rbe/.

SUM (1970–1980). *Slovnyk ukrajins'koji movy v 11 tomax*. Kyjiv: Naukova dumka. http://sum.in.ua.

SUM-20 (2010—). *Slovnyk ukrajins'koji movy u 20 tomax*. Kyjiv: Naukova dumka. https://slovnyk.me/dict/newsum.

Turchyn, Je. (1985). Leksychna realizacija mikropolja "chastyny khlibyny" v ukrajins'kykh hovorakh, *Strukturni rivni ukrajins'kykh hovoriv*, pages 146–165. Kyjiv: Naukova dumka.

[12] The inclusion of dictionaries of dialects is especially advisable, as they are likely to present greater lexical variety than standard explanatory dictionaries, witness the wealth of words for pieces of bread in Ukrainian dialects summarised and classified in (Turchyn, 1985).

Appendix A. Bulgarian Wordlist

1. *бадемовка* 2
2. *баница* 60
3. *баничка* 15
4. *безквасник* 2
5. *бисквита* 33, *бисквитка* 1
6. *блин* 13
7. *бутерброд* 3
8. *бухта* 8, *бухтичка* 1
9. *вареник* 25
10. *вафла* 9
11. *геврек* 25, *гевречен* 1
12. *геврече* 8
13. *еклер* 2
14. *залък* 33, *залче* 1
15. *кадаиф* 1
16. *канола* 4
17. *кейк* 11
18. *кекс* 2
19. *кифла* 51
20. *кифличка* 6
21. *кнедла* 8
22. *кнши* 2
23. *козунак* 24, *козуначен* 1
24. *комат* 10
25. *коричка* 16, *кора*$_1$ 2, *корица* 1
26. *кора*$_2$ 1
27. *кравай* 10
28. *кравайче* 15
29. *крайче* 2
30. *крашник* 6, *крайшник* 2, *крайщник* 2, *краещник* 1
31. *кроасан* 3
32. *курабийка* 14
33. *курабия* 27
34. *къс* 3
35. *късче* 10
36. *къшей* 22
37. *листо* 2
38. *марципан* 7, *марципанен* 1
39. *маршал* 1
40. *маца* 3, *мац* 1
41. *меденка* 1
42. *мекица* 2
43. *нафора* 27
44. *пай* 3
45. *палачинка* 17
46. *парче* 71
47. *парченце* 9
48. *паста* 16
49. *пастет* 8
50. *пирог* 15
51. *пирожка* 39
52. *пита*$_1$ 21
53. *пита*$_2$ 7
54. *питка* 139
55. *погача* 14
56. *попара* 7
57. *порязаница* 2
58. *просеник* 1
59. *просфора* 3
60. *пудинг* 16
61. *пърженка* 2
62. *равиоли* 1
63. *реване* 1
64. *резен* 8, *резенче* 1
65. *самун* 46, *самунче* 1
66. *сандвич* 171
67. *симид* 1
68. *сладка* 19
69. *сладкиш* 96
70. *соленка* 2
71. *среда* 6
72. *сухар* 63
73. *сухарче* 3
74. *тако* 2
75. *тиганица* 9, *тиганичка* 1
76. *торта* 116
77. *тортичка* 7
78. *точено* 1
79. *тригуна* 1
80. *троха* 41
81. *трошица* 2, *трохица* 1
82. *трохичка* 3, *трошичка* 1
83. *филийка* 32
84. *филия* 53
85. *франзела* 5, *франзелка* 1
86. *хамбъргър* 6
87. *хапка* 8
88. *хлебец* 8
89. *хлебче* 49
90. *хляб* 1533
91. *щрудел* 9

Appendix B. Ukrainian Wordlist

1. *бабка* 9
2. *балабуха* 1
3. *баніца* 2, *баниця* 1
4. *батон* 2
5. *батончик* 3
6. *бісквіт* 13
7. *бріош* 2
8. *бублик* 26, *бубликовий* 1, *бубличний* 1
9. *бубличок* 3
10. *булка* 28
11. *булочка* 63, *півбулочки* 1
12. *бутерброд* 133, *бутербродний* 3, *бутербродик* 1
13. *буханець* 34, *буханка* 2, *боханець* 1, *півбухана* 1
14. *вареник* 28, *вареничок* 1
15. *варениця* 1
16. *ватрушка* 4
17. *вафля* 5, *вафельний* 2
18. *випічка* 3
19. *вівсяник* 1
20. *галета* 6
21. *гамбургер* 7
22. *глевтяк* 1
23. *горохвяник* 2
24. *гостія* 1
25. *грінка* 43, *гріночка* 1
26. *душеник* 2
27. *житник* 1
28. *завиванець* 1
29. *кавалок* 3
30. *калач* 28
31. *калачик* 11
32. *канапка* 2
33. *картопляник* 6
34. *квашене* 2
35. *кекс* 4
36. *кнши* 5
37. *колобок* 4
38. *коржс* 83
39. *коржик* 31
40. *коровай* 11
41. *крекер* 2
42. *крендель* 4, *крендельок* 1
43. *крихта* 39, *крихітка* 1, *крихточка* 1
44. *кришка* 8
45. *кукурудзяник* 1
46. *кулеб'яка* 1
47. *кулич* 1
48. *кусок* 14, *кус* 1
49. *кусень* 12
50. *лигун* 1
51. *луста* 1
52. *маківник* 2
53. *малай* 1
54. *марципан* 9, *марципановий* 2
55. *маца* 3
56. *медівник* 1
57. *млинець* 35, *млинчик* 2
58. *м'якуш* 5
59. *м'якушка* 4
60. *налисник* 3
61. *облатка* 28
62. *окраєць* 22, *окрайчик* 1
63. *оладка* 7
64. *опріснок* 25
65. *ощіпок* 1
66. *паляниця* 13, *паляничка* 1
67. *пампушка* 9
68. *паска* 19, *пасочка* 1
69. *паштет* 12
70. *перепічка* 13
71. *печиво* 56
72. *пиріг* 116
73. *пиріжечок* 5
74. *пиріжок* 42, *пиріжковий* 1
75. *підпалок* 9
76. *пляцинда* 1
77. *пляцок* 1
78. *потапці* 1
79. *прісне* 3
80. *пряник* 18, *пірник* 1, *попряничок* 1
81. *пудинг* 14
82. *пундик* 4
83. *рогалик* 7
84. *сандвіч* 39, *сендвіч* 6
85. *скиба* 3
86. *скибка* 34
87. *скибочка* 13
88. *скоринка* 11
89. *слойка* 1
90. *струдель* 9
91. *сухар* 60
92. *сухарець* 3
93. *сухарик* 4
94. *тартинка* 4
95. *тетеря* 5, *тетерка* 1
96. *тістечко* 71, *тістечковий* 2
97. *товченик* 1
98. *торт* 118
99. *тортик* 5
100. *тост* 1
101. *трубочка* 4
102. *хліб* 1440, *хлібний* 2
103. *хлібець* 20
104. *хлібина* 37, *півхлібини* 2, *півхлібинки* 1, *хлібинка* 1
105. *чурек* 1
106. *шкуринка* 6
107. *шмат* 6
108. *шматок* 81
109. *шматочок* 30
110. *шуляк* 1

A Customizable WordNet Editor

Andrei-Marius Avram
Research Institute for Artificial Intelligence, Romanian Academy,
University POLITEHNICA of Bucharest
avram.andreimarius@gmail.com

Verginica Barbu Mititelu
Research Institute for Artificial Intelligence, Romanian Academy
vergi@racai.ro

Abstract

This paper presents an open-source wordnet editor that has been developed to ensure further expansion of the Romanian wordnet. It comes with a web interface that offers capabilities in selecting new synsets to be implemented, editing the list of literals and their sense numbers and adding these new synsets to the existing network, by importing from Princeton WordNet (and adjusting, when necessary) all the relations in which the newly created synsets and their literals are involved. The application also comes with an authorization mechanism that ensures control of the new synsets added in novice or lexicographer accounts. Although created to serve the current (more or less specific) needs in the development of the Romanian wordnet, it can be customized to fulfill new requirements from developers, either of the same wordnet or of a different one for which a similar approach is adopted.

Keywords: wordnet editor, Romanian, synset creation, semantic relations, non-lexicalized synsets

1. Introduction

The importance of wordnets as linguistic knowledge repositories and as resources exploitable by language applications has been widely acknowledged and at present the wordnet community is still an active one, where:

- mature wordnets are:
  - exploited with state of the art technologies (Kafe, 2019), and
  - enriched (Dziob et al., 2019), even voluntarily (McCrae et al., 2019),
- projects for building wordnets still continue, either:
  - in the traditional way (Dziob et al., 2019) or
  - adding new dimensions to the lexico-semantic network: audio (Kashyap et al., 2019), visual (Deng et al., 2009) or another modality (Lualdi et al., 2019),
- wordnet development projects are debuting (Sio and da Costa, 2019),
- existing wordnets are converted from their various formats to a common format (Bond and Foster, 2013) in order to ensure their interlinking, as well as linking to other resources (Simov et al., 2019).

Against this effervescent background, we present the work for developing a tool for continuing the enrichment of the Romanian wordnet (RoWN, henceforth) with new synsets and relations. The development of the RoWN started in the BalkaNet project (Tufiș et al., 2004). That was when the implementation principles were established, the necessary tools were developed (Barbu and Tufiș, 2004)

and a core RoWN was created. Its enrichment continued in other projects (Tufiş et al., 2013), but had to stop abruptly when technical problems occurred and no support could be offered to lexicographers. However, enrichment with new relations and new types of information relevant to various applications was possible: further derivational relations between literals were added (Barbu Mititelu, 2013), verbal multiword expressions were annotated with four types (Barbu Mititelu and Mitrofan, 2019) established within the PARSEME initiative (Ramisch et al., 2018).

When developing a wordnet manually, an editor customized to the steps to take in this process is a must and many teams have developed their own (see section 2. below). For RoWN the tools, WNBuilder and WNCorrect (Barbu and Tufiș, 2004), installed locally, ensured a two-phase process, and several scripts complemented them: the synsets to be implemented were selected from Princeton WordNet (Fellbaum, 1998) (PWN, henceforth) with the help of a script, following criteria of interest at different moments in the project evolution, WNBuilder was used by lexicographers to create the synsets (i.e., enumerate the literals and specify their sense number) and their glosses, these synsets were automatically added to the RoWN file together with all the respective relations imported from PWN (thus ensuring the alignment of RoWN with PWN), ensuring their structural correctness at the same time (i.e., all literals must have a sense number, a literal could not occur twice in the same synset irrespective of its sense number, etc.). All semantic errors (i.e., the same literal with the same sense number occurring in at least two different synsets) were automatically identified with the help of a script and they were uploaded in the WNCorrect interface, where lexicographers manually corrected them. The semantic correction phase was iterative until no more errors were found. As mentioned above, the tools are no longer usable, given technical limitations (their dependence on older version of operating systems and of other software).

Continuing the development of the RoWN required a tool that would allow for synchronous development by different lexicographers and that would ensure both the syntactic and the semantic correctness of the new synsets. We present here the tool we have created to serve these needs.

The paper is organized as follows: section 2. briefly presents the wordnet editing tools available, in section 3. we explain the specificities of RoWN that required a new editing tool, which is described in details in section 4. The possibilities for customizing this editor for other wordnets are presented in section 5., before concluding the paper.

2. Other Wordnet Editors

As mentioned above, almost each team developing a wordnet has created a tool to help in this process. Some of them are project-specific, others are designed to serve multiple project, as well as multiple tasks: development, validation, visualization. We review below several such tools that we considered for further enrichment of RoWN.

Hydra (Rizov, 2014) is an open-source system allowing for PWN synsets cloning and their further modification (in any way: adding literals, deleting literals, deleting the whole synset, undoing any action, as well as redoing it), relations import, creation of language-specific synsets and their linking to existing synsets in the wordnet. It allows for simultaneous access and use by multiple users, with modifications becoming accessible to all of them instantly. Hydra can be used in browser, does not require local installation. RoWN can be queried in the 'Hydra for web' tool[1], built on top of Hydra, either in a single view mode or in parallel with other wordnets uploaded in the tool and which RoWN is aligned with at the synset level. Given the logic model behind Hydra (Rizov, 2008), the sense numbering of homograph literals ignores their part of speech, as well as the sense numbers in the original wordnets. Given the almost parallel development of the Bulgarian and Romanian wordnets (Barbu Mititelu et al., 2019), Hydra would have made an adequate solution for continuing the RoWN development, provided that it had offered a solution to the sense numbering system specific to RoWN (see section 3.) and to allowing access to language resources used by lexicographers when implementing new synsets or ensuring their quality.

DEBVisDic (Horák et al., 2006) is another tool for creating and further editing wordnets, as well as for querying them and visualizing the results even in several wordnets at a time, in a synchronous mode

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

(the autolookup function). It also offers the possibility of multiple users working simultaneously, without interfering with the work of each other. However, some relations have to be manually added, it is quite restrictive in access (noncommercial, nonprofit internal research purposes only), runs only on Mozilla's Firefox, cannot cope with the RoWN sense numbering system. At the beginning of the BalkaNet project, our team used a previous version of this tool, namely VisDic (Horák and Smrz, 2004), but only for visualization of wordnets content, as well as of our in-house Romanian explanatory dictionary that was in XML format, compliant with the requirements of the VisDic tool.

OMW editor (da Costa and Bond, 2015) is web-based, but requires local installation. It can be used for various languages (thus being advertised as a "multilingual editing environment"), by any number of users simultaneously, with their work becoming available for the others immediately. The tool also allows for checks of the work done, i.e. ensuring the structural validity of synsets, with no sense numbers, definitions, etc. missing.

Other wordnet editors were mainly tailored on the wordnet to be developed, on the respective creation process, on the resources used in this process, and many others. Their adaptation to the specificity of other wordnets and to the needs of other wordnets developers are not trivial and also lengthy: see the adaptation of WordnetLoom, created for the development of the Polish wordnet, to the requirements of the Portuguese wordnet development, as reported by Naskret et al. (2018).

As such, we faced the challenge of finding a new solution for continuing the development of our wordnet with new synsets.

3. The Romanian Wordnet

The method for creating RoWN consists in finding the right equivalent(s) of the PWN literals in a set of synsets chosen to be implemented and the transfer of the PWN semantic relations between the implemented synsets. This is the expand method (Rodriguez et al., 1998) in wordnets building. Two principles were always observed when selecting a set of synsets to implement: the Hierarchy Preservation Principle (semantic relations were imported automatically from PWN) and the Conceptual Density Principle (ensuring that no orphan synsets, i.e. lower-level synsets without direct ancestors, are created) (Tufiș et al., 2004).

One of the resources exploited for creating the RoWN was an in-house explanatory dictionary of Romanian. Its senses are organized in nests, which means that the main senses of a lexical entry are identified and for each such sense all its subsenses are defined; even subsenses can have sub-subsenses. The subsenses are clearly semantically related to the respective main sense and the semantic similarity between a subsense and its main sense is more obvious than the semantic similarity between main senses. The same applies to sub-subsenses and subsenses. While main senses are distinguished by integer sense numbers (e.g., literal:*1*, literal:*2*, etc.), subsenses are assigned decimal sense numbers (e.g., literal:*1.1*, literal:*1.2*, literal:*2.1*, literal:*2.2*, and even literal:*1.1.1*, literal:*1.1.2*, etc. for sub-subsenses). This sense numbering system was preserved for the literals included in RoWN (Tufiş et al., 2013), as semantically valuable and relevant information for cases when RoWN is used for calculating semantic distances between words or word senses. Moreover, these sense numbers can also be suffixed with ".x" whenever a subsense of a sense existing in the dictionary is necessary for rendering the PWN equivalent. Whenever for a literal existent in the Romanian dictionary a sense was needed that was not recorded in the dictionary (but lexicographers found it attested in Romanian corpora), the respective literal got the sense number "x" (this time, not a suffix to a sense number, but a sense number itself). The same sense number was used for any sense of a literal nonexistent in the dictionary. We can thus notice the multiple values this "x" has. When two different PWN synsets are difficult to understand as semantically different by our lexicographers (after analysing the glosses, the examples, the sets of synonyms, other lexicographic resources or corpora), they are considered artificially distinct synsets and their equivalent Romanian synsets contain the same literal with the same sense number, this time suffixed with ".c", which is further semantic information in our wordnet, signaling the possibility of semantically clustering together the respective synsets (Tufiş et al., 2013).

Sense numbering starts from 1 for the homonyms belonging to a given part of speech, which is, in

fact, also the case in PWN. As such, it becomes mandatory to specify the part of speech for a combination literal:sense number, as the same combination can apply to two or even more parts of speech, thus being ambiguous.

4. Implementation and Functionalities

This section further presents the decisions that were made during the development of RoWN Editor and the technical details of its implementation. The main aims in its development were portability, so it can easily be deployed on any operating system, and ease of implementation and maintenance, so a reliable and easy to use application was to be developed. We found out that Python was a perfect fit because it is an interpreted, general-purpose programming language that offers a lot of useful packages.

In order to make RoWN Editor a web-based application, we had to choose a web framework. Flask[2] proved to be the best candidate because of its lightweight nature and because it was designed to make getting started quick and easy, with the ability to scale up to complex applications. We also used packages that offer support for wordnets: RoWN API (Dumitrescu et al., 2018) for the RoWN and Natural Language Toolkit (NLTK) (Loper and Bird, 2002) for PWN. The rest of Python dependencies are listed in the `requirements.txt` file and can easily be installed with The Python Package Index (PyPI) via the `pip install -r requirements.txt` command. RoWN Editor has no system dependencies.

From the architectural point of view, the application is composed of three modules: synset selection, synset creation and authorization mechanisms. Each of these modules will be described in the following subsections and a workflow diagram that shows the interactions of these modules is depicted in Figure 1.

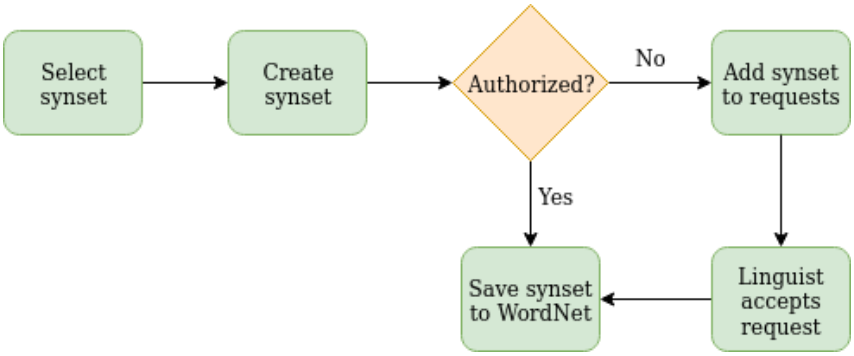


Figure 1: The work flow diagram of RoWN Editor. The user selects a new synset and implements it. The newly created synset will be directly added to the Romanian WordNet if the user is authorized (i.e. (s)he is a lexicographer). Otherwise, the synset will be saved in a list of requested synsets and will not be added until a lexicographer accepts the request.

4.1. Synset Selection

The RoWN Editor comes with an automatic way of suggesting new synsets that can be further added to the structure of wordnet. It does this by exploiting the partial mapping of synset IDs between RoWN and PWN. It selects the nodes (synsets) that exist in PWN, but not in RoWN, and have at least one edge (relationship) with a node that is implemented in RoWN. Figure 2 depicts the role of the synsets from a part of the hypernym tree when the selection algorithm is run. The green nodes represent the common synsets between RoWN and PWN. The red nodes represent the selected synsets - that are implemented in PWN, but not implemented in RoWN, and have at least one relation with one of the common synsets (the green nodes). The white nodes represent the English synsets that are neither implemented in RoWN, nor have a relation with one of the common synsets.

It must be noted that the existence of common edges between the selected synsets and the already implemented synsets in RoWN is a necessary condition, because otherwise it would imply the creation of a disconnected graph, thus disobeying the Conceptual Density Principle (see section 3.). Another

[2]The official documentation of Flask can be found at `https://flask.palletsprojects.com/en/1.1.x/`

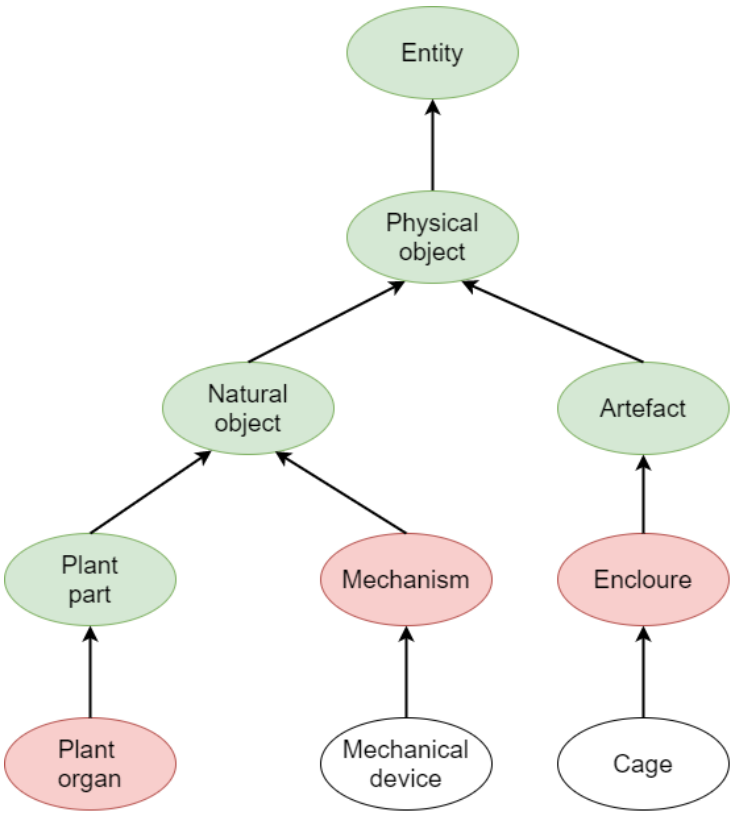


Figure 2: The synset selection algorithm.

important observation is that the algorithm does not return all the possible synsets, but only the first N nodes found, simply because of time constraints. N is a configurable parameter, that can also be specified as "all" to return every possible synset. However, this is not recommended because the selection algorithm might take a long time to find all the synsets. When needed, further restrictions can be imposed on the selected synsets: e.g., synsets belonging to a certain domain or involved in a certain type of relation, etc.

4.2. Synset Creation

Another important feature of the RoWN Editor is the ease of creating new synsets in the web interface. Once a user has selected an unimplemented synset, (s)he will be redirected to a new page where (s)he will have to complete the wordnet entry with the defining fields of a synset (definition, nonlexicalized, stamp, lemmas and lemmas sense number). The user can freely add and remove any number of lemmas, using the buttons "Add lemma" and "Remove lemma", respectively. If the Nonlexicalized box is checked, this option will be disabled and the added lemmas will not be taken into consideration. Once the synset is considered complete, the user can create it by clicking the button "Create synset". RoWN Editor will automatically import its identification number (ID) and all the relations between Romanian synsets from PWN, and, depending on user privileges, it will be either added directly to the wordnet or to a request list. To facilitate the implementation, the top of the page contains the English description of the selected synset. Figure 3 depicts the creation interface for the PWN synset {four-stroke engine:1, four-stroke internal-combustion engine:1}.

4.3. Authorization

The application has two types of users: the novice user and the lexicographer. One of the roles of the latter is to supervise the former by accepting, editing and rejecting the synsets proposed by him/her. RoWN Editor implements this feature by saving the synsets of the novice user in a list of requests, that will be added to the wordnet only after a lexicographer approves it. This mechanism ensures consistency

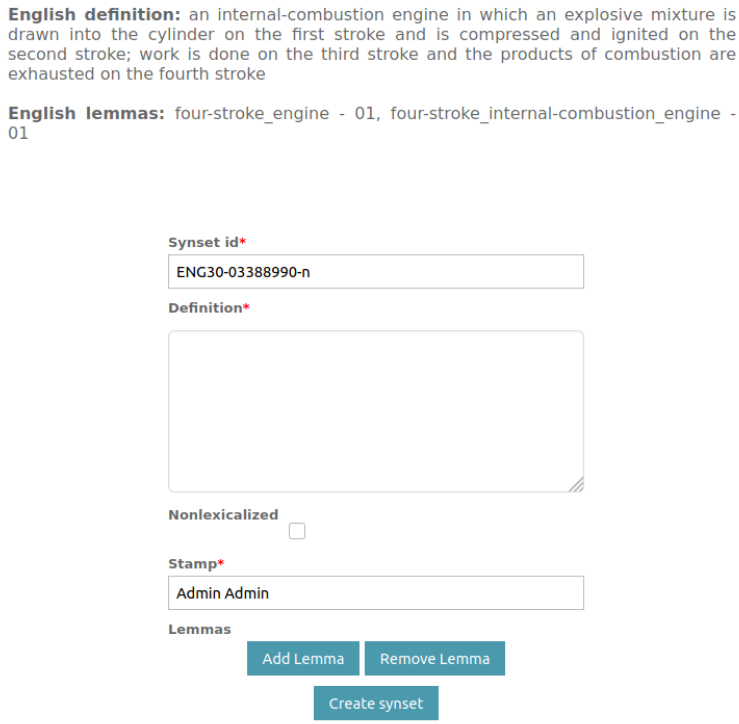


Figure 3: Synset creation interface for the synset with id ENG-30-03388990-n

in the working methodology.

5. Compatibility with Other Wordnets

The RoWN used in the application can be replaced with any wordnet by simply replacing the file `RoWordNet.xml` from the `rowordnet` directory with another wordnet in the XML format. However, the new wordnet must be aligned to PWN so that the synset selection algorithm can work. Also the new wordned XML must observe the following format:

- the root tag must be named `WORDNET`;
- each synset must be inside the `WORDNET` tag and must be marked by the `SYNSET` tag;
- a synset must contain the following tags:
  - `ID` - the identification number of the synset: this is identical to the PWN corresponding synset;
  - `POS` - the part of speech of the literals in the synset;
  - `SYNSONYM` - this is where the literals of the synset are listed. Each literal will be marked by the `LITERAL` tag, and each literal must contain the `SENSE` tag, denoting its sense number;
  - `STAMP` - it contains the name of the creator of the synset; this is an optional tag;
  - `DEF` - the gloss of the synset;
  - `ILR` - an inbound relation of the synset. It must contain the `TYPE` tag that denotes the type of relation with the synset it refers to. This tag can repeat as many times as many relations the synset has with other synsets.

Figure 4 depicts the synset with ID ENG30-00006269 from the RoWN XML. It can be observed that besides the above mentioned tags, it also contains the `DOMAIN`, `SUMO` and `SENTIWN` tags. These were imported into RoWN, but their presence is not manadatory in other wordnets and, thus, they can be omitted.

```
<SYNSET>
  <ID>ENG30-00006269-n</ID>
  <POS>n</POS>
  <SYNONYM>
    <LITERAL>viață<SENSE>1</SENSE></LITERAL>
  </SYNONYM>
  <STAMP>Dan Cristea</STAMP>
  <ILR>ENG30-00004258-n<TYPE>hypernym</TYPE></ILR>
  <ILR>ENG30-07993776-n<TYPE>hyponym</TYPE></ILR>
  <DEF>forme de viață, văzute în mod global; " Nu există viață pe Marte "</DEF>

  <DOMAIN>factotum</DOMAIN>
  <SUMO>Organism<TYPE>=</TYPE></SUMO>
  <SENTIWN>
    <P>0.0</P>
    <N>0.0</N>
    <O>1.0</O>
  </SENTIWN>
</SYNSET>
```

Figure 4: XML format for a synset in the RoWN.

6. Conclusion

This paper introduces the RoWN Editor, an application that facilitates the extension of the RoWN by offering an intuitive graphical interface through which a user can create new synsets. It comes with an algorithm that automatically suggests new synsets by comparing the RoWN with the PWN, and also an authorization mechanism that ensures its consistency by allowing only a lexicographer to add new synsets. Furthermore, the application has been developed for RoWN, but it can be customized to allow editing any other wordnet that respects the presented XML format.

RoWN Editor is a web application developed entirely in Python using Flask. It has no system dependencies, has an open MIT license and can be installed by following the steps at `https://github.com/avramandrei/RoWordNet-Editor`.

References

Barbu, E. and Tufiș, D. (2004). A Methodology and Associated Tools for Building Interlingual Wordnets. In *Proceedings of LREC2004*, pages 1067–1070.

Barbu Mititelu, V. and Mitrofan, M. (2019). Leaving No Stone Unturned When Identifying and Classifying Verbal Multiword Expressions in the Romanian Wordnet. In *Proceedings of Global WordNet Conference*, pages 10–15.

Barbu Mititelu, V., Stoyanova, I., Leseva, S., Mitrofan, M., Dimitrova, T., and Todorova, M. (2019). Hear about Verbal Multiword Expressions in the Bulgarian and the Romanian Wordnets Straight from the Horse's Mouth. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 2–12.

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

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352—-1362.

da Costa, L. M. and Bond, F. (2015). OMWEdit - The Integrated Open Multilingual Wordnet Editing System. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 73–78. ACL and AFNLP.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Dumitrescu, S. D., Avram, A. M., Morogan, L., and Toma, S.-A. (2018). RoWordNet–A Python API for the Romanian WordNet. In *2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–6. IEEE.

Dziob, A., Piasecki, M., and Rudnicka, E. (2019). plWordNet 4.1 – a Linguistically Motivated, Corpus-based Bilingual Resource. In *Proceedings of the 10th Global WordNet Conference*, pages 353–362.

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

Horák, A. and Smrz, P. (2004). VisDic - Wordnet Browsing and Editing Tool. In *Proceedings of GWC2004*, pages 136–141.

Horák, A., Pala, K., Rambousek, A., and Povoln, M. (2006). Debvisdic–first version of new client-server wordnet browsing and editing tool. In *Proceedings of the Third International Wordnet Conference (GWC-06)*.

Kafe, E. (2019). Fitting Semantic Relations to Word Embeddings. In *Proceedings of the 10th Global WordNet Conference*, pages 228–237.

Kashyap, K., Sarma, S. K., and Sweta, K. (2019). Spoken WordNet. In *Proceedings of the 10th Global WordNet Conference*, pages 260–263.

Loper, E. and Bird, S. (2002). Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

Lualdi, C., Hudson, J., Fellbaum, C., and Buchholz, N. (2019). Building ASLNet, a Wordnet for American Sign Language. In *Proceedings of the 10th Global WordNet Conference*, pages 315–322.

McCrae, J. P., Rademaker, A., Bond, F., Rudnicka, E., and Fellbaum, C. (2019). English WordNet 2019 – An Open-Source WordNet for English. In *Proceedings of the 10th Global WordNet Conference*, pages 245–252.

Naskret, T., Dziob, A., Piasecki, M., Saedi, C., and Branco, A. (2018). WordnetLoom – a Multilingual Wordnet Editing System Focused onGraph-based Presentation. In *Proceedings of the 9th Global Wordnet Conference*.

Ramisch, C., Cordeiro, S., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Jolanta Kovalevskaite, and Simon Krek, a. T. L., Liebeskind, C., Monti, J., Parra Escartín, C., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A., and Walsh, A. (2018). Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multi-word Expressions and Constructions*, pages 222–240.

Rizov, B. (2008). Hydra: A Modal Logic Tool for Wordnet Development, Validation and Exploration. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1523–1528.

Rizov, B. (2014). Hydra: A Software System for Wordnet. In *Proceedings of the Global WordNet Conference*, pages 142–147.

Rodriguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertagna, F., and Roventini, A. (1998). The top-down strategy for building eurowordnet: Vocabulary coverage, base concepts and top ontology. *Computers and the Humanities*, 32(2-3):117–152.

Simov, K., Osenova, P., Laskova, L., Radev, I., and Kancheva, Z. (2019). Aligning the Bulgarian BTB WordNet with the Bulgarian Wikipedia. In *Proceedings of the 10th Global WordNet Conference*, pages 290–297.

Sio, J. U.-S. and da Costa, L. M. (2019). Building the Cantonese Wordnet. In *Proceedings of the 10th Global WordNet Conference*, pages 206–215.

Tufiş, D., Barbu Mititelu, V., Ştefănescu, D., and Ion, R. (2013). The Romanian Wordnet in a Nutshell. *Language Resources and Evaluation*, 47(4):1305–1314.

Tufiș, D., Cristea, D., and Stamou, S. (2004). BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. *Romanian Journal of Information Science and Technology*, 7(1-2):9–43.

Comparison of Genres in Word Sense Disambiguation using Automatically Generated Text Collections

Angelina Bolshina
Lomonosov Moscow State University, Moscow, Russia
angelina_ku@mail.ru

Natalia Loukachevitch
Lomonosov Moscow State University, Moscow, Russia,
Kazan Federal University,
Kazan, Russia
louk_nat@mail.ru

Abstract

The best approaches in Word Sense Disambiguation (WSD) are supervised and rely on large amounts of hand-labelled data, which is not always available and costly to create. In our work we describe an approach that is used to create an automatically labelled collection based on the monosemous relatives (related unambiguous entries) for Russian. The main contribution of our work is that we extracted monosemous relatives that can be located at relatively long distances from a target ambiguous word and ranked them according to the similarity measure to the target sense. We evaluated word sense disambiguation models based on a nearest neighbour classification on BERT and ELMo embeddings and two text collections. Our work relies on the Russian wordnet RuWordNet.

Keywords: Word sense disambiguation, Russian dataset, Monosemous relatives.

1. Introduction

Word sense disambiguation (WSD) is one of the major challenges of computational semantics and it addresses the issue of lexical ambiguity. The aim of a WSD system is to identify the correct sense of a polysemous word in a context. This task has a wide range of potential applications including information retrieval, machine translation, and a knowledge graph construction. The training of well-performing supervised WSD algorithms involves a vast number of sense-labelled samples for each polysemous word in a language. There exist several hand-crafted sense-annotated datasets for English (Miller et al., 1993; Taghipour and Ng, 2015). However, this requirement is currently beyond reach in many languages and Russian is among them.

In this paper we present a knowledge-driven method based on the concept of monosemous relatives for the automatic generation of a training collection. We exploit a set of unambiguous words (or phrases) related to particular senses of a polysemous word. However, as it was noted in (Martinez et al., 2006), some senses of target words do not have monosemous relatives, and the noise can be introduced by some distant relatives. In our research we tried to address these issues.

In this work we proposed an extended and modified algorithm of training data generation based on monosemous relatives approach. The main contribution of this study is that we have expanded a set of monosemous relatives under consideration: in comparison with earlier approaches now they can be situated at greater distance from a target ambiguous word in a graph. Moreover, we have introduced a numerical estimation of a similarity between a monosemous relative and a particular sense of a target word which is further used in the development of the training collection. In order to evaluate the created training collections, we used contextualized word representations – ELMo (Peters et al., 2018)

and BERT (Devlin et al., 2019). We investigated the application of our algorithm to the training and test collections of different genres and their impact on the resulting performance of the WSD system[1].

The paper is organized as follows. In section two we review the related work. Section three is devoted to the data description. The fourth section describes the method applied to automatically generate and annotate training collections. The procedure of creating the collections is explained in the fifth section. In the sixth section we describe a supervised word sense disambiguation algorithm trained on our collected material and demonstrate the results obtained by four different models. In this section we also present a comparative analysis of the models trained on different kinds of train collections. Concluding remarks are provided in the seventh section.

2. Related Work

To overcome the limitations, that are caused by the lack of annotated data, several methods of generating and harvesting large train sets have been developed. There exist many techniques based on different kinds of replacements, which do not require human resources for tagging. The most popular method is that of monosemous relatives (Leacock et al., 1998). Usually WordNet (Miller, 1995) is used as a source for such relatives. WordNet is a lexical-semantic resource for the English language that contains a description of nouns, verbs, adjectives, and adverbs in the form of semantic graphs. All words in those networks are grouped into sets of synonyms that are called synsets.

Monosemous relatives are those words or collocations that are related to the target ambiguous word through some connection in WordNet, but they have only one sense, i.e. belong only to one synset. Usually, synonyms are selected as relatives but in some works hypernyms and hyponyms are chosen (Przybyła, 2017). Some researchers replace the target word with named entities (Mihalcea and Moldovan, 2000), some researchers substitute it with meronyms and holonyms (Seo et al., 2004). In the article (Yuret, 2007) a special algorithm was created in order to select the best replacement out of all words contained within synsets of the target word and neighbouring synsets. The algorithm described in (Mihalcea, 2002) to construct an annotated training set is a combination of different approaches: monosemous relatives, glosses and bootstrapping. Monosemous relatives can be also used in other tasks, for example, for finding the most frequent word senses in Russian (Loukachevitch and Chetviorkin, 2015). Other methods of automatic generation of training collections for WSD exploit parallel corpora (Taghipour and Ng, 2015), Wikipedia and Wiktionary (Henrich et al., 2012), topic signatures (Agirre and De Lacalle, 2004). (Pasini and Navigli, 2017) created large training corpora exploiting a graph-based method that took an unannotated corpus and a semantic network as an input.

Various supervised methods including kNN, Naive Bayes, SVM, neural networks were applied to word sense disambiguation (Navigli, 2009). Recent studies have shown the effectiveness of contextualized word representations for the WSD task (Wiedemann et al., 2019; Kutuzov and Kuzmenko, 2019). The most widely used deep contextualized embeddings are ELMo and BERT.

In ELMo (Embeddings from language models) (Peters et al., 2018) context vectors are computed in an unsupervised way by two layers of bidirectional LSTM, that take character embeddings from convolutional layer as an input. Character-based token representations help to tackle the problems with out-of-vocabulary words and rich morphology. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) has a different type of architecture, namely a multi-layer bidirectional Transformer encoder. During the pre-training procedure, the model is "jointly conditioning on both left and right context in all layers" (Devlin et al., 2019: 1). Since these contextualized word embeddings imply capturing polysemy better than any other representations and, thus, we employ them in our investigation.

3. Data

In our research as an underlying semantic network, we exploit Russian wordnet RuWordNet (Loukachevitch et al., 2016). It is a semantic network for Russian that has a WordNet-like structure. In total it contains 111.5 thousand of words and word combinations for the Russian language. RuWordNet was used to extract semantic relations (e.g. synonymy, hyponymy, etc.) between a target sense of a polysemous word and all the words (or phrases) connected to it, including those linked via

[1] The source code of our algorithm is publicly available at: https://github.com/loenmac/russian_wsd_data

distant paths. The sense inventory was also taken from this resource. RuWordNet contains 29297 synsets for nouns. There are 63014 monosemous and 5892 polysemous nouns in RuWordNet. Table 1 presents a summary of the number of senses per noun:

Number of senses of a polysemous noun	**Number of nouns in RuWordNet**
2 senses	4271
3 senses	997
4 senses	399
5 senses	149
> 5 senses	76
Total number of senses	14 357

Table 1: Quantitative characteristics of polysemous nouns in RuWordNet

We utilized two corpora in the research. A news corpus consists of news articles harvested from various news sources. The texts have been cleaned from HTML-elements or any markup. Another corpus is Proza.ru, a segment of Taiga corpus (Shavrina and Shapovalova, 2017), which is compiled of works of prose fiction. We exploit these two corpora because we want to investigate whether the genre of the training corpus has an impact on the performance on the test dataset.

For evaluation of our algorithm of training data generation, we used three distinct RUSSE'18 datasets for Russian (Panchenko et al., 2018) that were created for the shared task on word sense induction for the Russian language. The first dataset is compiled from the contexts of the Russian National Corpus[2]. The second dataset consists of the contexts from Wikipedia articles. And the last dataset is based on the Active Dictionary of the Russian Language (Apresyan et al., 2017) and contains contexts taken from the examples and illustration sections from this dictionary. All the polysemous words are nouns.

Explanation	**Number of words**	**Example**
A word has only one sense in RuWordNet	34	The word двойник (*dvojnik*, "doppelganger") has only one sense in RuWordNet whereas in RUSSE'18 it has 4.
A word is missing in the RuWordNet vocabulary	9	The word гипербола (*giperbola*, "hyperbole").
The senses from RuWordNet and RUSSE'18 dataset have only one sense in common	4	The word мандарин (*mandarin*) has two senses in RUSSE'18: its sense "tangerine" is included in the thesaurus, whereas its sense "mandarin, bureaucrat" is absent.
Controversial cases of sense mapping	29	The word демократ (*democrat*, "democrat") has 2 senses: "supporter of democracy" and "a member of the Democratic Party". But there's another one in RUSSE'18: "a person of a democratic way of life, views".
Not enough examples for senses in the corpora	2	Words карьер (*kar`er*, "quarry/a very fast gallop") and шах (*shax*, "shah/check").
Words with morphological homonymy	1	The word суда (*suda*, "court (Gen, Sg)/ship (Nom, Pl)") can have two distinct lemmas.

Table 2: Cases when a word from RUSSE'18 dataset was not included in the final test dataset

[2] http://www.ruscorpora.ru/new/index.html

From the RUSSE dataset we excluded some polysemous words, and in Table 2 we overview the common reasons why it was done. The final list of the target ambiguous words contains 30 words in total, each having two different senses. All the texts with the target ambiguous nouns in this dataset have sense annotation. We will call the resulting test dataset RUSSE-RuWordNet because it is a projection of RUSSE'18 sense inventory on the RuWordNet data.

We also created a small training dataset, that consists of the word sense definitions and examples of uses from Ozhegov dictionary (Ozhegov, 2014) for every target polysemous word. This training data is utilized as a baseline for the WSD task. In this set each sense of an ambiguous word has one definition and between 1 and 3 usage examples.

Table 3 demonstrates quantitative characteristics of all of the above-mentioned corpora:

	Taiga-Proza.ru	**News Corpus**	**RUSSE-RuWordNet**	**Dictionary Corpus (Baseline)**
Number of sentences	32,8 million	24,2 million	2 103	144
Number of lemmas	246,8 million	288,1 million	39 311	657
Number of unique lemmas	2,1 million	1,4 million	12 110	475

Table 3: Quantitative characteristics of the corpora and datasets used in the experiments

4. Candidate Selection and Ranking Algorithm

The central idea of our method is based on the assumption that a training collection can be built not only with the direct relations like synonymy, hypernymy and hyponymy but also with far more distant words, such as co-hyponyms. For example, most contexts for the word крона (*krona*) in the sense "krona, currency" match the contexts of the other words denoting currency like английский фунт (*anglijskij funt*, "pound sterling") as they have common hypernym валюта (*valyuta*, "currency").

The principal features of our approach are as follows:

1. We take into consideration not only the closest relatives to a target word sense, as it was done in previous works, but also more distant relatives.
2. We utilize similarity scores between a candidate monosemous relative and synsets close to a sense of a target polysemous word in order to evaluate how well this candidate can represent the sense of an ambiguous word.
3. We introduce the notion of ***a nest*** that is used to assess the potential of a candidate's usage contexts for displaying target sense of a polysemous word. In order to measure the relevance and suitability of a monosemous candidate, we exploit a thesaurus set of words similar to a target sense. The group of synonyms to a target sense and all the words from directly related synsets within 2 steps from a target word comprise ***the nest*** for a target sense.
4. We check similarity scores to the nest for both closest and further located monosemous relatives because a word described as monosemous in the thesaurus can actually have polysemous usage in a corpus. For example, Russian word ириска (*iriska*, "toffee") can also denote a nickname of Everton Football Club (The Toffees) (Loukachevitch, 2019). Thus, all candidate monosemous relatives should be further checked on the source corpus.
5. We propose two distinct methods of compiling a training collection based on the monosemous relatives rating.

A target word sense is a sense of a polysemous word that we want to disambiguate. Candidate monosemous relatives are unambiguous words (or phrases), that can be located in up to four-step relation paths to a polysemous word and include co-hyponyms, two-step (or more) hyponyms and hypernyms. We consider the words (or phrases), that have more than 50 occurrences in the corpus.

A fragment of the nest for the word *такса* (*taksa*, "dachshund") is given below:

(1) охотничий пёс (*oxontichij pyos*, "hunting dog"), пёсик (*pyosik*, "doggie"), четвероногий друг (*chetveronogij drug*, "four-legged friend"), собака (*sobaka*, "dog"), терьер (*ter`er*, "terrier") … etc.

The choice of the distance constant for the nest was motivated by the fact that the senses of the relatives located at the 2-step relation path are close to the target sense of the polysemous word and, thus, these relatives are more reliable and do not require sophisticated additional verification. As for the distance used to extract candidate monosemous relatives, we decided to stick to the maximum distance of 4, because usually the words located at 5 or more steps from the target sense are too generic. For example, the monosemous candidate for the word такса (*taksa*, "dachshund") located at 4-step path is животное (*zhivotnoe*, "animal") and the candidate at the 5-step path is биологический организм (*biologicheskij organizm*, "biological organism"). We can see that the second word is more general and can be used in a wide variety of contexts, and many of them may not at all be related to animals and dogs in particular. Another similar example is гвоздика (*gvozdika*, "clove"): its 4-step relative is продовольственные продукты (*prodovolstvenny`je producty* "food products") and 5-step relative is вещество (*veshhestvo* "substance").

Our method of extracting monosemous relatives is based on comparison of distributional and thesaurus similarities. Embedding models are utilized to select the most appropriate monosemous relatives whose contexts serve as a good representation of a target word sense. We used the word2vec models to extract 100 most similar words to each monosemous word from the candidates list. In that way, we collected the words that represent a distributional set of close words with the respective cosine similarities measures. Our selection and ranking method, thus, consists of the following steps:

1. We extract all the candidate monosemous relatives within 4 steps from a target polysemous word sense $\boldsymbol{s_j}$.
2. We compile the nest $\boldsymbol{ns_j}$ which consists of synonyms to a target sense and all the words from the synsets within 2 steps from a target word $\boldsymbol{s_j}$. The nest $\boldsymbol{ns_j}$ consists of $\boldsymbol{N_k}$ synsets.
3. For each candidate monosemous relative $\boldsymbol{r_j}$, we find 100 most similar words according to the word2vec model trained on a reference corpus.
4. We intersect these top-100 words with the words included in the nest $\boldsymbol{ns_j}$ of the target sense $\boldsymbol{s_j}$.
5. For each word in the intersection, we take its cosine similarity weight calculated with the word2vec model and assign it to the synset it belongs to. The final weight of the synset in the nest $\boldsymbol{ns_j}$ is determined by the maximum weight among the words $\boldsymbol{w^j_{k_1}, \dots, w^j_{k_i}}$ representing this synset in the intersection.
6. The total score of the monosemous candidate $\boldsymbol{r_j}$ is the sum of the weights of all synsets from the nest $\boldsymbol{ns_j}$. In such a way more scores are assigned to those candidates, that resemble a greater number of synsets from the nest close the target sense of the ambiguous target word. Thus, the final weight of the candidate can be defined as follows:

$$Weight_{r_j} = \sum_{k=1}^{N_k} \max\left[cos\left(r_j, w^j_{k_1}\right), \dots, cos\left(r_j, w^j_{k_i}\right)\right]$$

The following fragment of list of monosemous relatives with similarity scores (given in brackets) was obtained for the noun гвоздика (*gvozdika*, "clove"):

(2) мускатный орех (*muskatny`j orex*, "nutmeg") (6), имбирь (*imbir`*, "ginger") (6.4), корица (*korica*, "cinnamon") (6.5), кардамон (*kardamon*, "cardamom") (6.8), чёрный перец (*cherny`j perec*, "black pepper") (7.5)… etc.

We have also found some examples where a monosemous word is connected to a sense of a target word but got zero similarity weight. For example, the word марля (*marlya*, "gauze") is a cohyponym to the word байка in the sense (*bajka*, "thick flannelette") but was not included in the monosemous relatives list because its distributional set of close words did not have any intersection with the nest.

As a result of this procedure, all monosemous relatives are sorted by the weight they obtained. The higher-rated monosemous relatives are supposed to be better candidates to represent the sense of the target word and, consequently, their contexts of use are best suited as the training examples in the WSD task. The candidate ranking algorithm identifies which monosemous relatives are most similar to the target ambiguous word's sense. Once we have detected the monosemous candidates, we can extract from the corpus the contexts in which they occur. Then in these texts we substitute the monosemous relatives with the target ambiguous word and add the texts with the respective sense labels to a training collection.

In order to verify the applicability of our method to the RuWordNet material, we found candidate monosemous relatives for the ambiguous words in the thesaurus using our algorithm but without word2vec filter. Only two words out of 5895 do not have monosemous relatives within four-step relation path in RuWordNet graph. The quantitative characteristics of the candidate monosemous relatives are presented in Table 4. As it was mentioned in (Taghipour and Ng, 2015: 339), 500 samples per sense is enough for training data. Table 5 demonstrates how many target senses have at least 500 samples of their monosemous relatives in a reference corpus. We also take into consideration the case when word2vec filter was applied to the candidate monosemous relatives. These tables show that by applying our approach to the RuWordNet data we would be able to find monosemous relatives to almost all the polysemous words in the thesaurus and create a training collection for a WSD system.

Distance to a candidate monosemous relative	**Number of target senses, that have at least one relative at this distance**
0 (synset)	9 818
1	13 095
2	14 129
3	14 021
4	13 768

Table 4: Quantitative characteristics of candidate monosemous relatives for RuWordNet target senses

	Number of target senses when word2vec filter was not applied	**Number of target senses when word2vec filter was applied**
Taiga-Proza.ru	13 738	12 797
News Corpus	14 017	13 099

Table 5: Target senses with more than 500 occurrences of monosemous relatives in the corpora.

5. Generating Training Data using Monosemous Relatives

For comparison, we decided to create two separate training collections compiled from the news and Proza.ru corpora, and we also exploited two distinct approaches to a collection generation. According to the first method, we compiled the collection only with a monosemous relative from the top of the candidate rating. We wanted to obtain 1000 examples for each of the target words, but sometimes it was not possible to extract so many contexts with one particular candidate. That is why in some cases we also took examples with words next on the candidates' list. For simplicity, we call this collection Corpus-1000 because we obtained exactly 1000 examples for each sense.

The second approach enables to harvest more representative collection with regard to the variety of contexts. The training examples for the target ambiguous words were collected with the help of all respective unambiguous relatives with non-zero weight. The number of extracted contexts per a monosemous candidate is in direct proportion to its weight. We name this collection a balanced one because the selection of training examples was not restricted to the contexts which have only one particular monosemous relative.

In Table 6 we present the quantitative characteristics of the two collections, such as the relations connecting the target senses and their monosemous relatives, distances between them, and a proportion of monosemous relatives expressed as a phrase.

Feature	Proportion of occurrences in the news collection	Proportion of occurrences in Proza.ru collection
Distance to a target sense		
0 (synset)	2%	4%
1	13%	9%
2	38%	37%
3	31%	34%
4	16%	16%
Relation between a target sense and a monosemous relative		
Synonyms	2%	4%
Hyponyms	13%	8%
Hypernyms	11%	9%
Cohyponyms	28%	28%
Cohyponyms situated at three-step path	24%	28%
Cohyponyms situated at four-step path	19%	22%
Other	3%	1%
Word combinations	48%	29%

Table 6: Quantitative characteristics of monosemous relatives

Two word2vec embedding models that we used in our experiments were trained separately on the news and Proza.ru corpora with the window size of 3. As a preprocessing step, we split the corpora into separate sentences, tokenized them, removed all the stop words, and lemmatized the words with pymorphy2 tool (Korobov, 2015). The words obtained from the word2vec model were filtered out – we removed the ones not included in the thesaurus.

6. Experiments

We conducted several experiments to determine which text collection used as training data for a WSD model gives the best performance on the test dataset. Following (Wiedemann et al., 2019), in our research we used an easily interpretable classification algorithm – non-parametric nearest neighbor classification (kNN) based on the contextualized word embeddings ELMo and BERT.

In our experiments we exploited two distinct ELMo models ‑ the one trained by DeepPavlov on Russian WMT News and the other is RusVectōrēs (Kutuzov and Kuzmenko, 2017) lemmatized ELMo model trained on Taiga Corpus (Shavrina and Shapovalova, 2017). The difference between these two models is that from the first model we extracted a vector for a whole sentence with a target word, whereas from the second model we extracted a single vector for a target ambiguous word. As for BERT, we used two models: BERT-base-multilingual-cased released by Google Research and RuBERT, which was trained on the Russian part of Wikipedia and news data by DeepPavlov (Kuratov and Arkhipov, 2019). To extract BERT contextual representations, we followed the method described by (Devlin et al., 2019) and (Wiedemann et al., 2019) and concatenated “the token representations from the top four hidden layers of the pre-trained Transformer” (Devlin et al., 2019: 9).

The Tables 7 and 8 demonstrate the results obtained by different types of contextualized word embeddings, the training collections, and model parameters.

Model	ELMo RusVectōrēs (target word)		ELMo DeepPavlov (whole sentence)		RuBERT DeepPavlov		Multilingual BERT	
k	Proza.ru	News collection	Proza.ru	News collection	Proza.ru	News collection	Proza.ru	News collection
1	0.809	0.794	0.765	**0.752**	0.751	0.735	0.668	0.67
3	0.826	0.811	**0.773**	0.749	0.781	0.756	0.684	0.673
5	0.834	**0.819**	0.77	0.748	0.793	0.771	0.694	0.667
7	**0.841**	**0.819**	0.767	0.746	**0.804**	**0.774**	0.699	0.673
9	0.84	0.816	0.762	0.747	0.802	0.769	**0.7**	**0.677**
Baseline	0.772		0.716		0.667		0.672	

Table 7: F1 scores for ELMo- and BERT-based WSD models, Corpus-1000 collections

Model	ELMo RusVectōrēs (target word)		ELMo DeepPavlov (whole sentence)		RuBERT DeepPavlov		Multilingual BERT	
k	Proza.ru	News collection	Proza.ru	News collection	Proza.ru	News collection	Proza.ru	News collection
1	0.812	0.797	0.745	0.758	0.746	0.75	0.669	0.662
3	0.833	0.81	0.775	0.753	0.778	0.755	0.707	0.681
5	0.845	0.81	0.776	0.756	0.792	0.769	0.717	0.682
7	**0.857**	0.815	**0.793**	**0.759**	0.802	0.768	0.723	0.683
9	0.856	**0.821**	0.791	0.753	**0.812**	**0.774**	**0.729**	**0.688**
Baseline	0.772		0.716		0.667		0.672	

Table 8: F1 scores for ELMo- and BERT-based WSD models, balanced collections

As it can clearly be seen, all the systems surpassed the quality level of the baseline solution trained on the dataset of the dictionary definitions and usage examples. This means that we have managed not only to collect training data sufficient to train the WSD model but also to show a good performance on the RUSSE-RuWordNet dataset.

The Proza.ru model achieves better results and outperforms the news model. The qualitative analysis of the classification errors caused by the model trained on the news collection showed that the main cause of mistakes were lexical and structural differences between training and test sets. The examples from the test dataset were from the Russian National Corpus and Wikipedia, whereas the training collections were composed of news articles. On the contrary, Proza.ru collection consists of various works of fiction, so, the training samples have more similar representations to the test ones. We thus conclude that similar genres of train and test collections give higher results in the WSD task.

The algorithm based on the ELMo pre-trained embeddings by RusVectōrēs outperformed all other models achieving 0.857 F1 score. The second-best model in the WSD task is RuBERT by DeepPavlov, followed by ELMo model by DeepPavlov. The lowest F1 score belongs to Multilingual BERT. As for the difference in F1 scores between the Corpus-1000 and the balanced collection, we can observe the performance drop for the Corpus-1000 for all the models, which means that the approach used to generate the balanced collection is better suited for the task. Corpus-1000 does not include all possible monosemous relatives, so the collection lacks contextual diversity, the balanced collection, on the contrary, is more representative with regard to the variety of contexts.

7. Conclusion

The issue that we addressed in this article is the lack of sense-annotated training data for supervised WSD systems in Russian. In this paper we have described our algorithm of automatic collection and

annotation of training data for the Russian language. The main contribution of the paper is that we have utilized in the selection algorithm not only close monosemous relatives but also more distant ones. Moreover, we implemented the procedure of ranking monosemous relatives' candidates. Our training collections consist of the texts extracted from the news and Proza.ru corpora. The candidate scores were obtained from two word2vec models trained separately on each corpus.

In order to evaluate the training collections, we applied kNN classifier to the contextualized word embeddings extracted for target polysemous words and measured its performance on the RUSSE-RuWordNet test dataset. We have investigated the capability of different deep contextualized word representations to model polysemy. The best result was obtained with RusVectōrēs ELMo model and amounted to 0.857 F1 score. We have also found out that the training collection harvested from the Proza.ru corpus gave higher F1 scores on the RUSSE-RuWordNet test dataset than the collection from the news corpus.

Acknowledgements

The work of Loukachevitch N. in the current study concerns formulation of the disambiguation approach for RuWordNet data, calculation of paths between synsets, criteria for selecting contexts; this work is supported by the Russian Science Foundation grant no. 19-71-10056 financed through Kazan Federal University.

References

Agirre, E., De Lacalle, O. L. (2004). *Publicly Available Topic Signatures for all WordNet Nominal Senses*. In LREC.

Apresyan, V. Yu., Apresyan, Yu. D., Babaeva, E. E., Boguslavsaya, O. Yu., Glovinskaya, M. Ya., Iomdin, B. L., Krylova, T. V., Levontina, I. B., Lopukhina, A. A., Ptentsova, A. V., Sannikov, A. V., Uryson, E. V. (2017). *Active Dictionary of the Russian Language* [*Aktivny`j slovar` russkogo yazyka*]. Publishing House Nestor-Istoria, Moscow, Vol. 3.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. In Proc. of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, pp. 4171‑4186.

Henrich, V., Hinrichs, E., Vodolazova, T. (2012). *Webcage: A Web-harvested Corpus Annotated With GermaNet Senses*. In Proc. of the 13th Conf. of the European Chapter of the ACL, pp. 387-396.

GOST 7.79-2000: *System of Standards for Information, Library Services and Publishing. Rules for Transliteration of Cyrillic Letters into the Latin Alphabet.*

Korobov, M. (2015). *Morphological Analyzer and Generator for Russian and Ukrainian Languages*. In Analysis of Images, Social Networks and Texts, pp. 320-332.

Kuratov, Y., Arkhipov, M. (2019). *Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language*. arXiv preprint arXiv:1905.07213.

Kutuzov, A., Kuzmenko, E. (2017). *WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models*. In Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661. Springer, Cham.

`https://rusvectores.org/ru/`

Kutuzov, A., Kuzmenko, E. (2019). *To Lemmatize or not to Lemmatize: How Word Normalisation Affects ELMo Performance in Word Sense Disambiguation*. In Proc. of the First NLPL Workshop on Deep Learning for Natural Language Processing, pp. 22-28.

Leacock, C., Miller, G. A., Chodorow, M. (1998). *Using Corpus Statistics and WordNet Relations for Sense Identification*. Computational Linguistics, vol. 24(1), pp. 147-165.

Loukachevitch, N. (2019). *Corpus-Based Check-Up for Thesaurus*. In Proc. of the 57th Annual Meeting of the ACL, pp. 5773-5779.

Loukachevitch, N. V., Lashevich, G., Gerasimova, A. A., Ivanov, V. V., Dobrov, B. V. (2016). *Creating Russian WordNet by Conversion*. In Proc. of Conference on Computational linguistics and Intellectual technologies Dialog-2016, pp. 405-415.

Loukachevitch, N., Chetviorkin, I. (2015). *Determining the Most Frequent Senses Using Russian Linguistic Ontology RuThes*. In Proc. of the workshop on Semantic resources and semantic annotation for NLP and the Digital Humanities at NODALIDA 2015, pp. 21-27.

Martinez, D., Agirre, E., Wang, X. (2006). *Word Relatives in Context for Word Sense Disambiguation*. In Proc. of the Australasian Language Technology Workshop 2006, pp. 42-50.

Mihalcea, R. (2002). *Bootstrapping Large Sense Tagged Corpora*. In Proc. of the Third International Conference on Language Resources and Evaluation (LREC-2002), vol. 1999.

Mihalcea, R., Moldovan, D. I. (2000). *An Iterative Approach to Word Sense Disambiguation*. In FLAIRS Conference, pp. 219-223.

Miller, G. (1995). *WordNet: A Lexical Database for English*. In Communications of the ACM, vol.38(11), pp. 39-41.

Miller, G. A., Leacock, C., Tengi, R., Bunker, R. T. (1993). *A Semantic Concordance*. In Proc. of the workshop on Human Language Technology, pp. 303-308.

Navigli, R. (2009). *Word Sense Disambiguation: A survey*. ACM computing surveys (CSUR), vol. 41(2), 10.

Ozhegov, S.I. (2014). *Explanatory Dictionary of the Russian Language* [*Tolkovy`j Slovar` Russkogo Yazyka*]. Edited by Skvortsova S.I., 8, pp. 1376.

Panchenko, A., Lopukhina, A., Ustalov, D., Lopukhin, K., Arefyev, N., Leontyev, A., Loukachevitch, N. (2018). *RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language*. In Computational Linguistics and Intellectual Technologies: Dialogue-2018, pp. 547‑564.

Pasini, T., Navigli, R. (2017). *Train-O-Matic: Large-Scale Supervised Word Sense Disambiguation in Multiple Languages Without Manual Training Data*. In Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 78-88.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. (2018). *Deep Contextualized Word Representations*. In Proc. of the 2018 Conference of the North American Chapter of the ACL: Human Language Technologies, pp. 2227‑2237.

Przybyła, P. (2017). *How Big is Big Enough? Unsupervised Word Sense Disambiguation Using a Very Large Corpus*. arXiv preprint arXiv:1710.07960.

Seo, H. C., Chung, H., Rim, H. C., Myaeng, S. H., Kim, S. H. (2004). *Unsupervised Word Sense Disambiguation Using WordNet Relatives*. Computer Speech & Language SPEC. ISS., vol. 18, no. 3, pp. 253-273.

Shavrina, T., Shapovalova, O. (2017). *To the Methodology of Corpus Construction for Machine Learning: «Taiga» Syntax Tree Corpus and Parser.* In Proc. of "CORPORA2017", international conference, Saint-Petersburg.

Taghipour, K., Ng, H. T. (2015). *One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction*. In Proc. of the 19th Conf. on computational natural language learning, pp. 338-344

Wiedemann, G., Remus, S., Chawla, A., Biemann, C. (2019). *Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings*. arXiv preprint arXiv:1909.10430.

Yuret, D. (2007). *KU: Word Sense Disambiguation by Substitution*. In Proc. of the 4th International Workshop on Semantic Evaluations, pp. 207-213.

Consistency Evaluation towards Enhancing the Conceptual Representation of Verbs in WordNet

Svetlozara Leseva and **Ivelina Stoyanova**
Department of Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences
zarka@dcl.bas.bg, iva@dcl.bas.bg

Abstract

This paper outlines the process of enhancing the conceptual description of verb synsets in WordNet using FrameNet frames. On the one hand we expand the coverage of the mapping between WordNet and FrameNet, while on the other – we improve the quality of the mapping using a set of consistency checks and verification procedures. The procedures include an automatic identification of potential inconsistencies and imbalanced relations, as well as suggestions for a more precise frame assignment followed by manual validation. We perform an evaluation of the procedures in terms of the quality of the suggestions measured as the potential improvement in precision and coverage, the relevance of the result and the efficiency of the procedure.

Keywords: FrameNet, WordNet, Frame semantics, consistency evaluation

1. Introduction and Motivation

Our work aims at enhancing the conceptual description of verb synsets in WordNet through integrating frame semantics as represented in FrameNet. Below we briefly discuss the resources used in the study and the methodology we apply. We first overview existing alignments between WordNet and FrameNet which make an impact on the adopted methodology for the mapping of the two resources. The mapping proposed herein elaborates on and expands previous alignments by employing the inheritance of conceptual and lexical information (Section 2.2). In addition, we devise a set of consistency checks and frame suggestion procedures in order to further improve the quality and coverage of the resulting resource; these procedures rely on lexical and semantic properties, similarity and the relational structure of WordNet and FrameNet (Section 3). The evaluation of the procedures is based on a manually validated dataset (Section 4) and is presented in Section 5, which is followed by a brief discussion of the results in Section 6. The conclusions are summed up in Section 7.

1.1. Resources

We employ two lexical semantic resources – WordNet and FrameNet. WordNet (Miller, 1995; Fellbaum, 1998) is a large lexical database that represents comprehensively conceptual and lexical knowledge in the form of a network whose nodes denote sets of cognitive synonyms (synsets) interconnected through a number of conceptual-semantic and lexical relations such as synonymy, hypernymy, meronymy, etc. The main relation that determines WordNet's taxonomical structure is the relation of hypernymy.

FrameNet (Baker et al., 1998; Baker, 2008) represents lexical and conceptual knowledge couched in the apparatus of frame semantics. Frames are conceptual structures describing particular types of objects, situations, or events along with their components, called frame elements, or FEs (Baker et al., 1998; Baker and Ruppenhofer, 2002; Ruppenhofer et al., 2016). Depending on their semantic obligatoriness and contribution to the conceptual description, FEs may be core, peripheral or extra-thematic(Ruppenhofer et al., 2016); core FEs are the most essential as their configuration makes a frame unique, which is why our focus is on them. Frames are instantiated by lexical units (LUs) which are

included as part of the description of the relevant frame. In addition, frames are related by means of frame-to-frame relations, some of which are discussed below.

The combination of the two resources is expected to strengthen their individual advantages, in particular the great lexical coverage and the branched and rich relational structure of WordNet with the detailed conceptual description of the combinatorial potential of lexical units supplied by FrameNet. The contribution of our research is directed to the expansion of the mapping of the two resources (with the prospect of integrating others, such as VerbNet, Propbank, etc.) to the end of overcoming the sparsity of the overlap between synset members (literals) and LUs in FrameNet.

1.2. Methodology

The work presented here is a continuation of our previous work on enhancing the conceptual description of verbs in WordNet that goes in two directions – expanding the coverage and improving the quality of the mappings (Leseva and Stoyanova, 2019; Stoyanova and Leseva, 2019). To this end, we propose a set of methods for expanding the mapping between the resources based on the relations of inheritance (cf. Section 2.2), which are further enhanced by means of automatic procedures for validation and improvement that focus on minimising manual work (cf. Section 3.1).

Most of the procedures that have been proposed rely on: (i) the notion of inheritance and the hierarchical relational structure of FrameNet and WordNet – the internal structure of the two resources is determined to a great extent by the notion of inheritance: in WordNet it is realised by the hypernymy relation, whereas in FrameNet it is represented mainly by the relations of *Inheritance* (strong inheritance) (Ruppenhofer et al., 2016), *Using* (weak inheritance) (Petruck, 2015) , as well as by relations such as *Subframe*, and *Perspective on*, although in a very limited way; (ii) semantic and lexical analysis of the components of the description in FrameNet and WordNet.

2. Mapping WordNet and FrameNet

2.1. Compilation of Existing Mappings between WordNet and FrameNet

Earlier research aimed at maximising the advantages and the richness of the conceptual and lexical information encoded in WordNet, FrameNet and other resources has led to a number of proposals, including the mapping of WordNet, FrameNet and VerbNet by Shi and Mihalcea (2005), the elaboration of WordFrameNet[1] by Laparra and Rigau (2010) and MapNet[2] by Tonelli and Pighin (2009), the implementation of other FrameNet-to-WordNet mappings, e.g. by Ferrandez et al. (2010). More enhanced proposals have been made too, such as Semlink[3] (Palmer, 2009), which unifies WordNet, FrameNet and VerbNet with PropBank, and its follow-up Semlink+ that brings in a mapping to Ontonotes (Palmer et al., 2014).

In the domain of verbs, out of 14,103 verb synsets, only 4,306 (30.5%) have been mapped through finding existing equivalents (using existing mappings) (Leseva and Stoyanova, 2019). A number of consistency checks were also implemented on the result of the initial mapping before the application of the frame assignment procedure described in 2.2. These checks led to improved connectivity between synsets, which in turn made the hypernym-to-hyponym frame assignment more efficient; part of the checks involved correction of already assigned frames so as to avoid the propagation of errors in the course of hypernym-to-hyponym frame assignment (Leseva and Stoyanova, 2019).

2.2. Inheritance-based mapping

After the implementation of the initial mapping and preliminary validation procedures mentioned above, we undertake expansion of the mapping by incorporating procedures aimed at 'digging up' non-explicit information about the frame membership of WordNet literals on the basis of the relational information in the two resources. Along these lines Burchardt et al. (2005) propose the expansion of the inter-resource coverage (mapping WordNet literals to FrameNet frames) by weighing the candidate frames evoked by literals related to a given target literal through certain semantic relations (synonymy, hypernymy,

[1] http://adimen.si.ehu.es/web/WordFrameNet
[2] https://hlt-nlp.fbk.eu/technologies/mapnet
[3] https://verbs.colorado.edu/semlink/

antonymy). Another strategy is to exploit the relational structure of the resources – particularly that of WordNet – by mapping frames to synsets on the basis of the inheritance of conceptual features in hypernym trees, i.e. by assigning frames from hypernyms to hyponyms (Leseva et al., 2018).

We adopt this latter approach and using the initial automatic mapping of 4,306 synsets (cf. Section 2.1), we implemented a procedure of transferring the hypernyms' frames to their hyponyms in the cases where the hyponyms were not directly mapped to FrameNet frames (through the initial mapping). In such a way, we obtained an extended coverage of 13,226 synsets (synsets with an assigned FrameNet frame) out of the total of 14,103 verb synsets (Leseva and Stoyanova, 2020).

The main drawback of the inheritance approach is that especially for deeper level WordNet synsets the inherited frames may be underspecified. Thus, a natural follow-up was to look for ways to discover appropriately specific frames which have already been defined in FrameNet. This is where the FrameNet's relational structure comes into play as it may point to where to look first for probable candidate frames. Most of the procedures that are proposed below rely on the information and the overall relational structure of FrameNet and WordNet, as well as the semantic and lexical analysis of the components of the description in FrameNet and WordNet.

3. Consistency Checks and Procedures to Verify and Enhance the Conceptual Description

The main idea of these procedures is to explore: (i) the lexical mapping of the target synset's literals to lexical units in FrameNet frames that are related (through frame-to-frame relations) to the frame assigned to the target synset from its hypernym (i.e. exploring the vicinity of the frame inherited from the hypernym); (ii) the lexical mapping of the target synset's literals to lexical units in FrameNet frames that are assigned to synsets in the vicinity of the target synset (its hyponyms and sister synsets in particular); (iii) similarity, e.g. the similarity between keywords in WordNet glosses and the definitions of FrameNet lexical units, cf. Section 3.1. The methodology and implementation have been described in detail in Leseva and Stoyanova (2019) and Stoyanova and Leseva (2019).

3.1. Procedures based on Lexical and Semantic Analysis Involving Hierarchical FrameNet Relations

The procedures involve several steps, as described below:

(1) Check whether any of the literals of the target synset appears as a LU in: (a) the frame assigned from the synset's hypernym (to confirm its validity); (b) more specific frames the frame under discussion is linked to by means of any of the considered frame-to-frame inheritance relations (so as to try to find a suitable more specialised frame); (c) the sister frames of the assigned frame (the frames sharing a parent with the one assigned from the synset's hypernym).

Example 1. Synset: eng-30-01900255-v {flutter:3}
Gloss: flap the wings rapidly or fly with flapping movements
Assigned frame from hypernym: Body_movement
Suggested from (1a): Body_movement (LU: flutter)

The synset in Example 1 is assigned the frame Body_movement directly from its hypernym {beat:8, flap:3} 'move with a thrashing motion', which in its own right is assigned this frame through one or more of the automatic mappings described in Section 2.1. The appropriateness of the assignment through inheritance is confirmed by means of Procedure (1a) as the single literal in this synset, *flutter*, is found as a LU in the frame Body_movement.

Example 2 illustrates Procedure (1b). The synset is originally assigned the frame Cause_change from its hypernym {change:1; alter:1; modify:3} 'cause to change; make different; cause a transformation'. A more specific and better matching frame Cause_change_of_strength is suggested by the procedure on the basis of: (i) the fact that the three literals in the synset are found as LUs in this frame; (ii) there is an inheritance relation between the frame assigned from the hypernym and the newly suggested frame.

Example 2. Synset: eng-30-00220869-v {strengthen:1; beef up:1; fortify:1}
Gloss: make strong or stronger

Assigned frame from hypernym: Cause_change
Suggested from (1b): Cause_change_of_strength (LU: strengthen; beef up; fortify)

Example 3 illustrates Procedure (1c). The synset {educate:3, school:2, train:5, cultivate:3, civilize:1, civilise:1} is assigned the frame Cause_to_make_progress from its hypernym, while three of the literals are found as LUs in its sister frame Education_teaching (both Cause_to_make_progress and Education_teaching inherit from the frame Intentionally_affect).

Example 3. Synset: eng-30-02388403-v {educate:3, school:2, train:5, cultivate:3, civilize:1, civilise:1}
Gloss: teach or refine to be discriminative in taste or judgment
Assigned frame from hypernym: Cause_to_make_progress
Suggested from (1c): Education_teaching (LUs: educate, school, train)

To sum up, Procedure 1 aims at establishing whether a synset's literal or literals may be found as a LU/LUs in a substructure of frames related through inheritance or sisterhood to the frame assigned from the hypernym, thus expanding the frame window explored for literal-to-LU lexical match.

(2) Check whether any of the target synset' literals appears as a LU in: (a) any of the frames assigned to its hyponyms; (b) any of the frames assigned to its sister synsets; and (c) any of the frames related to the frames in (a) and/or (b).

Example 4. Synset: eng-30-00097621-v {regenerate:9; revitalize:1}
Gloss: restore strength
Assigned frame from hypernym: Cause_to_make_progress
Suggested frame from (2a): Rejuvenation (LU: revitalize)

In Example 4 the synset {regenerate:9; revitalize:1} is assigned the frame Cause_to_make_progress from its hypernym {better:2, improve:1, amend:2, ameliorate:1, meliorate:1} 'to make better'. A more specific and accurate frame is suggested through Procedure (2a) by virtue of the fact that the frame Rejuvenation, which is assigned to the single hyponym of {regenerate:9; revitalize:1} – {rejuvenate:3} 'make younger or more youthful', contains a LU matching one of the literals of {regenerate:9; revitalize:1}: *revitalize*.

Example 5. Synset: eng-30-00080705-v {nurse:1}
Gloss: try to cure by special care of treatment, of an illness or injury
Assigned frame from existing mapping: Medical_professionals
Suggested frame from (2b): Cure (LU: nurse)

The synset {nurse:1} in Example 5 was originally assigned the frame Medical_professionals (from the initial automatic mapping), which includes nouns denoting medical workers. Through the application of Procedure (2b), we found out that some of the sisters of the synset are assigned the frame Cure (e.g. {massage:2} 'give a massage to' and {insufflate:2} 'treat by blowing a powder or vapor into a bodily cavity') and that the literal *nurse* corresponds to a LU in the same frame. These two facts in conjunction motivate the suggestion of the frame Cure for the synset {nurse:1}.

Example 6. Synset: eng-30-01013230-v {remonstrate:2; point out:3}
Gloss: present and urge reasons in opposition
Assigned frame from hypernym: Telling
Suggested frame from (2c): Judgment_communication (LU: remonstrate)

Example 6 illustrates the application of Procedure (2c) and the resulting suggestion of Judgment_communication as a possible replacement of the frame Telling originally assigned to {remonstrate:2; point out:3} from its hypernym. In this case, some of the sisters of the target synset, such as {announce:3; denote:3} 'make known; make an announcement', are assigned the frame Statement, which is a more general frame related through the relation *Inheritance* ('strong inheritance') to Telling, and through the relation *Using* ('weak inheritance') to Judgment_communication; in addition, the target synset's literal *remonstrate* is found as a LU in the latter frame. The frame assigned from the hypernym and the frame suggested through Procedure (2c) are 'step-sisters' as they are related to the same frame (Statement) through similar, but not identical relations.

The purpose of the procedures subsumed under Procedure 2 is to establish whether a target synset's literal(s) may be found as a LU/LUs in frames assigned to synsets in the WordNet substructure defined by the target synset's children (hyponyms) and/or sisters as well as in the vicinity of such frames.

(3) Check whether any of the synset literals appear as a LU in any other frame in FrameNet.

Example 7. Synset: eng-30-02217266-v {finance:1}
Gloss: obtain or provide money for
Assigned frame from hypernym: Commerce_pay
Suggested from (3): Funding (LU: finance)

The synset in Example 7 is assigned the frame Commerce_pay from its hypernym {pay:1} 'give money, usually in exchange for goods or services', which is close in meaning as financing involves paying. The more appropriate frame Funding is not related through any of the frame-to-frame relations to Commerce_pay and is suggested by virtue of the fact that the verb *finance* is a LU in this frame. As Procedure 3 is based primarily on the lexical correspondence between literals and LUs, it is more reliable for verbs with fewer senses or in cases where the suggested frame is additionally confirmed. The latter is true for the example under discussion: the proposed frame Funding receives support from the fact that two of the hyponyms of {finance:1} – {back:5} 'support financial backing for' and {fund:1} 'convert (short-term floating debt) into long-term debt that bears fixed interest and is represented by bonds' – are LUs in the same frame (hence their mapping to Funding is suggested by the same Procedure 3).

Procedure 3 may be the only available one when the assignment through inheritance has failed and hence no frame has been mapped to the synset from its hypernym (therefore procedures 1b and 1c are not applicable), consider 8 below.

Example 8. Synset: eng-30-02227741-v {abandon:4; give up:5}
Gloss: give up with the intent of never claiming again
Assigned frame from hypernym: none (ROOT synset)
Suggested from (3): Surrendering_possession (LU: give up); Abandonment (LU: abandon); Quitting_a_place (LU: abandon); Activity_stop (LU: abandon)

In this particular example four frames – Surrendering_possession, Abandonment, Quitting_a_place and Activity_stop – are suggested by Procedure 3 based on the literals (*give up, abandon*) found in them. Manual analysis is then performed to select the most appropriate frame by also considering frames related to the ones suggested.

(4) In this step we use keywords (words found in the name of a FrameNet frame, plus their derivatives collected from WordNet through the *eng_derivative* relation) and identify synsets with literals and/or glosses containing these keywords as candidates to be assigned the frame under discussion.

Example 9. Synset: eng-30-00768389-v {talk out of:1}
Gloss: persuade someone not to do something
Assigned frame: Suasion
Confirmed by (4): keyword:persuade (in gloss)

The synset {talk out of:1} is assigned the frame Suasion from its hypernym {dissuade:1, deter:2} 'turn away from by persuasion', which is the appropriate one. While Procedures (1-3) fail to suggest a frame, Procedure (4) confirms the assignment of Suasion through the keyword *persuade* in the gloss of {talk out of:1}, which is a LU in the frame Suasion.

3.2. Similarity-based Procedures

In addition to the procedures described in Section 3.1, we also introduce checks based on the similarity measures between synset glosses in WordNet and LU definitions in FrameNet. Similarity is measured as the degree of overlapping word roots (using stemming) where direct overlaps of words are given a higher score than the overlaps after stemming.

The procedures include:

(1) Direct similarity: In this step we identify candidate frames for a target synset by checking the similarity between its gloss and FrameNet LU definitions (even though there is no lexical correspondence between the synset's literals and the LUs).

Example 10. Synset: eng-30-01399821-v {beetle:3}
Gloss: beat with a beetle
Assigned frame from hypernym: Cause_ harm
Suggested from (1): Cause_ harm
Confirmed by: similarity of the WordNet gloss with LU definitions (bludgeon; cudgel; whip)

In this example {beetle:3} is assigned the frame Cause_ harm from its hypernym {beat:3} 'hit repeatedly'. This suggestion is confirmed by the similarity existing between the gloss of the synset and the definitions of the LUs *bludgeon*, *cudgel* and *whip* in the frame Cause_harm – 'beat with a bludgeon', 'beat with a cudgel', 'beat with a whip'.

(2) Indirect similarity: In this procedure we identify candidate frames for the target synset by checking the similarity between the glosses of the synsets that are derivationally related to it (as well as the glosses of their hypernyms which are their closest semantic generalisations) and FrameNet LU definitions.

In Example 11 the synset {solo:1} was initially mapped to the frame Shaped_part, an assignment originating from an error in an existing mapping between the synset {handle:4; palm:1} 'touch, lift, or hold with the hands' and this frame; the wrong assignment was then transferred in four steps down the tree from hypernyms to hyponyms to {solo:1}. This error was corrected after an appropriate frame, Operate_vehicle, was suggested using the indirect similarity procedure based on the lexical similarity between the gloss of the derivationally related noun {solo:3} and the definition of one of the LUs in the frame Operate_vehicle, *fly.v*. The similarity is calculated on the basis of matching words (in bold) excluding closed class lexemes and auxiliaries and taking into account the length of the glosses. Scores of over 1.0 are considered a strong indicator of similarity between the definitions.

Example 11. Synset: eng-30-01941987-v {solo:1}
Gloss: fly alone, without a co-pilot or passengers
Assigned frame from hypernym: Shaped_part
Derivationally related synset: eng-30-00304729-n {solo:3}
Gloss: a **flight** in which the **aircraft** pilot is unaccompanied
Suggested from (2): Operate_vehicle (1.11)
Confirmed by: similarity between the gloss of {solo:3} and the gloss of the LU fly.v
Gloss of LU fly.v: control the **flight** of (an **aircraft**)

3.3. Ranking the Suggestions

We consider separately the suggested frames that are related to the frame assigned from the hypernym as they are given higher priority over unrelated suggested frames. We give each lexical match a score based on the calculated similarity, then assign an overall cumulative score to each frame and rank the suggestions so that the more likely candidates are analysed first in order to optimise the manual work. In Example 12 the suggestion Motion_noise has been yielded by the LUs *crackle*, *squelch* and *hiss*, and this is why it is ranked higher than Fluidic_motion (yielded only by the LU *hiss*).

Example 12. Synset: eng-30-02069120-v {woosh:1, whoosh:1}
Gloss: move with a sibilant sound
Hypernym: eng-30-01850315-v {move:2, displace:4}
Assigned frame from hypernym: Cause_motion
Suggested frames from the procedures: Motion_noise:crackle (to move making soft sharp repeating sounds 1.17); Motion_noise:squelch (move with such a sound 1.2); Motion_noise:hiss (to move making a sibilant sound as of the letter s 1.12); Self_motion:wriggle (move with wiggling movements 1.2); Fluidic_motion:hiss ((for air) to move producing a sharp sibilant sound 1.29)
Assigned correct frame: Motion_noise

Future work will be directed to the development of a methodology for quantifying relevance and more precise ranking of the suggestions, so that manual work is minimised and an automatic procedure for filtering and selection of suggestions is implemented.

4. Manual Verification of Automatic Frame Suggestions

The output of all the applied procedures is produced as a list for the experts to analyse and possibly confirm the appropriate candidates. Consider Example 13 below. In this particular instance, each of the frames was suggested on the basis of a direct or an indirect derivational or semantic relation between the target synset and another synset – represented by the relevant literals – *terrifying*, *pleasing*, *loathing*, etc., which in their own right have been assigned the suggested frames on the basis of lexical mapping with LUs in the respective frames. A linguist needs to study the list of suggested frames and select an appropriate one if such is available.

Example 13. Synset: eng-30-01813668-v {exult:1, walk on air:1, be on cloud nine:1, jump for joy:1}
Gloss: feel extreme happiness or elation
Hypernym: eng-30-01813884-v {rejoice:1, joy:1}
Assigned frame from hypernym: Feeling
Suggested frames from the procedures: Stimulus_focus: terrifying (1.2); Stimulus_focus:pleasing (1.17); Stimulus_focus:exhilarating (1.17); Emotion_directed:agony (1.25); Emotion_directed:ecstatic (1.5); Emotion_directed:fury (1.25); Experiencer_focused_emotion:loathing (1.4)
Assigned correct frame: Experiencer_focused_emotion

The candidate frames produced via the assignment through inheritance and the devised procedures have been validated manually for approximately one third (4,522 out of 14,103) of the verb synsets. In the following Section 5 we present our findings on the performance of the various types of assignments.

5. Evaluation

We evaluate the development of the mappings between WordNet and FrameNet in terms of: (a) the improvement in the overall precision and coverage of the mapping; (b) the relevance of each procedure; and (c) the efficiency of each procedure. Precision is measured by counting, on the one hand, the number of mappings that we consider consistent, and on the other – the inconsistent mappings that have been identified and corrected, as well as any cases of missing elements (either in FrameNet or in WordNet, or both) that create imbalance and a skewed relation in any of the resources. Relevance takes into account the degree to which a frame suggested from the automatic procedures is directly related through a frame-to-frame relation to the manually validated frame (see details below). In addition, we pay attention to the efficiency of each procedure which we evaluate as the proportion of valid frame proposals out of all the suggestions obtained through a given procedure.

We perform a detailed analysis on each type of procedure and evaluate its results and efficiency with respect to the evaluation dataset containing 4,522 manually validated synsets (Section 4). We use the following data: (a) the initial mapping (baseline 1) – compiled from the evaluation dataset by applying only existing previous mappings before any extensions and consistency procedures are carried out (cf. Section 2.1); (b) the extended mapping (baseline 2) – compiled from the evaluation dataset baseline 1 by assigning a frame from a hypernym to its hyponyms in the cases where the hyponyms are not assigned a frame from the existing mappings (cf. Section 2.2); (c) the final mapping (manually validated) – after manual validation were carried out (cf. Section 4).

The baseline and the output mark the two extreme points on our evaluation scale. Our analysis takes into account the fact that the output is neither in its final state nor all the WordNet synsets are fully verified. The adopted level of detail in the classification of frames can vary for different resources and/or purposes, so we can always introduce more fine-grained frames, which will affect the evaluation of the procedures presented here with respect to the new output.

We introduce a detailed evaluation of the procedures involving not only precision and coverage but also relevance and efficiency of the result of each of the proposed procedures. Relevance measures the precision of the procedure itself and shows how close the output is to the desired result rather than whether it is precise since in the case of conceptual description assigning a more general frame to a synset is not considered wrong, although a more specific frame may be more informative.

The evaluation analysis tries to reflect the fact that the relevance of the assigned frames is not a

binary value (true/false). To this end, we introduce a scale from 0.00 to 1.00 to measure the relevance of a suggested assignment produced by the application of a given procedure with reference to the manually validated output: 1.00 is scored when a suggestion coincides with the final output; 0.50 – when a suggested frame is directly related to the final result via an inheritance relation; 0.25 – when a suggested frame is directly related to the final output via a different relation (e.g., Using, See also); 0.00 – when the two frames are not directly related.

The efficiency of each procedure is represented as the ratio between the number of changes undertaken and the total number of suggestions made using a particular procedure. The need to evaluate efficiency is related to the fact that manual verification is an expensive and time-consuming task, so the number of entities and suggestions to be manually checked needs to be optimised. Essentially, the efficiency measures the precision of the procedure itself. Procedures that require a lot of checks but identify very few relevant entries requiring changes are to be avoided unless essential for a particular task. The measure can also be used as a point of departure to optimising certain automatic procedures and consistency checks. One possible approach is the ranking of suggestions so that more likely ones appear first and thus, reduce the need to check lower ranking suggestions.

Table 5 shows the precision, coverage, relevance and efficiency when the discussed procedures are applied separately and in combination.

Procedure	**Precision**	**Coverage**	**Relevance**	**Efficiency**
BASELINE 1 (cf. sec. 2.1)	0.632	0.339	N/A	N/A
BASELINE 2 (cf. sec. 2.2)	0.782	0.774	N/A	N/A
Lexical & Semantic analysis (cf. sec. 3.1)	0.405	0.654	0.720	0.190
Similarity (cf. sec. 3.2)	0.334	0.590	0.691	0.117
All procedures (excl. BASELINES)	0.486	0.694	0.859	0.250
All procedures (with BASELINES)*	0.796	0.851	N/A	N/A

Table 1: Precision, coverage, relevance and efficiency of each of the procedures applied independently and in combination. *These results count the cases where either the baseline assignment was confirmed or the correct frame was suggested at least once by any of the procedures.

The reported results show that the semantic analysis and the similarity procedures as defined at present have limited contribution to improving the precision and coverage of the frame assignment. However, it is also evident that they complement the inheritance mapping and each other and that the improvement increases when all the procedures are applied in combination.

The observations on the relevance of the procedures show that we need to evaluate the results more broadly, not only in terms of the direct contribution to improving the precision and coverage but also as a contribution to the manual verification by facilitating expert decisions or giving helpful clues. The efficiency of the procedures is very low, which points to the need of narrowing down the possible suggestions in order to optimise manual work.

6. Discussion

While they are not conclusive (as all the data are not yet manually validated), the proposed expansion procedures and consistency checks are promising with respect to the task of frame-to-synset alignment.

6.1. Inheritance Mapping, Semantic Analysis, Similarity

Inheritance-based assignment proves to be the most powerful procedure in terms of its impact on both precision and coverage. Given that the frames assigned to synsets (where such assignments are available) are correct, the transfer of a synset's frame to its hyponyms must also be correct as hyponyms inherit an essential part of their semantic and lexical properties from their hypernyms, although the parent frame may be too underspecified, especially for deeper level synsets (ones assigned a frame from a distant hypernym).

As noted above, the procedures based on semantic analysis and on similarity have a marginal effect in terms of their contribution as compared with the inheritance-based mapping. The suggestions coming from these additional procedures, however, merit attention as they do yield appropriately specific frame candidates that cannot be discovered if inheritance assignment is used alone. Moreover, they offer suggestions which significantly narrow down the scope (the vicinity of related frames) where a more suitable frame can be found in the FrameNet structure and thus contribute to optimising manual work. This aligns with the considerable relevance score of the semantic analysis and the similarity-based procedures.

6.2. Towards the Definition of New Frames

Due to the discrepancy between the lexical coverage of FrameNet and WordNet, even with the application of inheritance mapping, there may not be a suitable enough parent frame to be assigned. In such cases, new frames need to be created in order to be able to achieve better coverage.

Our current efforts are directed towards defining frames that elaborate on more general ones in a way that is consistent with the formulation of already existing frames. As noted in Leseva et al. (2019), while frames have been created that describe changes in various attributes, such as *temperature* (Cause_temperature_change), *consistency* (Cause_change_of_consistency), *phase* (Cause_change_of_phase), *strength* (Cause_change_of_strength), among others, corresponding frames are missing for equally specific properties, such as *colour*, *taste*, *chemical composition*, etc. The definition of such new frames as undertaken in our work is modelled on already formulated ones. For instance, in defining the new frames Cause_chemical_reaction, Cause_change_colour, Cause_change_taste, we follow the definitions of Cause_change_of_consistency and Cause_change_of_phase with which they most closely correspond.

In addition, in a number of cases certain frames are predictable from the FrameNet structure but have not been implemented. A notable example is the lack of frame correspondences between causative and inchoative parts of the lexicon where either of the members may be missing. We take as a model pairs of frames, such as Cause_change and Undergo_change or Cause_change_of_position_on_a_scale and Change_of_position_on_a_scale, among many others, where the causative frame is related to the inchoative frame by means of the *Causative of* relation; we then proceed to define a new causative or inchoative frame where one must exist and link it to its counterpart by means of this relation. For instance, the frame Cause_change_of_strength assigned to {strengthen:1, beef up:1, fortify:1} 'make strong or stronger' does not have an inchoative counterpart that should be assigned to {strengthen:2} 'gain strength'. In a like manner, the causative frame may be missing: Change_direction, Motion_directional, Self_motion do not have causative correspondences although this distinction is made for their parent frame Motion (with its counterpart Cause_motion). Thus, for instance, {march:3} 'walk fast, with regular or measured steps; walk with a stride' is assigned the frame Self_motion, but there is no corresponding frame to account for {march:2} 'force to march' and other verbs describing self propelled motion of a person, animal, vehicle, etc. brought about through the action of another participant. One should either resort to the more general frame Cause_motion or define a new one Cause_self_motion. The same procedure of defining both the causative and the inchoative correspondence is carried out when defining new frames such as the ones described in the previous paragraph.

Finally, we intend to explore new frame suggestions made by teams working on framenets for different languages within the Global FrameNet project[4] and incorporate suitable ones.

7. Conclusions and future work

The work envisaged in the near future is aimed at providing further validation of the frame assignment to verb synsets in WordNet. A challenging prospective research will be to devise new frames that provide description of parts of the verb lexicon that have not yet been tackled in FrameNet as well as of parts of the Bulgarian verb lexicon that have no English counterparts.

A further goal is to employ the obtained linked resource in tasks such as semantic role labelling, event detection, syntactic parsing, machine translation, among others. The mapping between WordNet and FrameNet as well as the newly devised frames will be made available to the research community.

[4] `https://www.globalframenet.org/`

Acknowledgements

This study has been carried out as part of the project *Towards a Semantic Network Enriched with a Variety of Semantic Relations* funded by the National Science Fund of the Republic of Bulgaria under the Fundamental Scientific Research Programme (Grant Agreement 10/3 of 14.12.2016).

References

Baker, C. F. and Ruppenhofer, J. (2002). FrameNet's Frames vs. Levin's Verb Classes. In Larson, J. and Paster, M., Eds., *Proceedings of 28th Annual Meeting of the Berkeley Linguistics Society*, pages 27–38.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *COLING-ACL '98: Proceedings of the Conference. Montreal, Canada*, pages 86–90.

Baker, C. F. (2008). FrameNet, present and future. In Webster, J., Ide, N., and Fang, A. C., Eds., *The First International Conference on Global Interoperability for Language Resources*, Hong Kong. City University, City University.

Burchardt, A., Erk, K., and Frank, A. (2005). A WordNet detour to FrameNet. In *Sprachtechnologie, mobile Kommunikation und linguistische Resourcen*, volume 8 of *Computer Studies in Language and Speech*. Lang, Frankfurt, Germany.

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

Ferrandez, O., Ellsworth, M., Munoz, R., and Baker, C. F. (2010). Aligning FrameNet and WordNet based on semantic neighborhoods. In *Proceedings of the 7th Conference on Language Resources and Evaluation (LREC 2010), May 17-23, Valletta, Malta*, pages 310 – 314.

Laparra, E. and Rigau, G. (2010). eXtended WordFrameNet. In *Proceedings of LREC 2010*, pages 1214–1219.

Leseva, S. and Stoyanova, I. (2019). Enhancing Conceptual Description through Resource Linking and Exploration of Semantic Relation. In *Proceedings of 10th Global WordNet Conference, 23 – 27 July 2019, Wroclaw, Poland*, pages 229–238.

Leseva, S. and Stoyanova, I. (2020). Beyond Lexical and Semantic Resources: Linking WordNet with FrameNet and Enhancing Synsets with Conceptual Frames. In *Towards a Semantic Network Enriched with a Variety of Semantic Relations*.

Leseva, S., Stoyanova, I., and Todorova, M. (2018). Classifying Verbs in WordNet by Harnessing Semantic Resources. In *Proceedings of CLIB 2018, Sofia, Bulgaria*, pages 115–125.

Leseva, S., Stoyanova, I., Todorova, M., and Kukova, H. (2019). Frame Specialisation Motivated by Inter-Frame Relations in FrameNet. In *Proceedings of the 14th International Conference on Linguistic Resources and Tools for Natural Language Processing, Cluj-Napoca, 18-20 November 2019, Editura Universităii Alexandru Ioan Cuza din Iasi*, pages 229–238.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41.

Palmer, M., Bonial, C., and McCarthy, D. (2014). SemLink+: FrameNet, VerbNet and Event Ontologies. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929–2014), Baltimore, Maryland USA, June 27, 2014*, pages 13–17. Association for Computational Linguistics.

Palmer, M. (2009). Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*. 9–15.

Petruck, M. R. (2015). *The Components of FrameNet*. `http://naacl.org/naacl-hlt-2015/tutorial-framenet-data/FNComponentsMRLP.pdf`.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., Baker, C. F., and Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California.

Shi, L. and Mihalcea, R. (2005). Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In Gelbukh, A., Ed., *Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science*, volume 3406. Springe, Berlin, Heidelbergr.

Stoyanova, I. and Leseva, S. (2019). Structural Approach to Enhancing WordNet with Conceptual Frame Semantics. In Angelova, G., Mitkov, R., Nikolova, I., and Temnikova, I., Eds., *Natural Language Processing in a Deep Learning World. Proceedings of RANLP 2019, Varna, Bulgaria*, pages 630 – 638.

Tonelli, S. and Pighin, D. (2009). New Features for Framenet – Wordnet Mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL'09), Boulder, USA*.

On WordNet Semantic Classes: Is the Sum Always Bigger?

Tsvetana Dimitrova
Institute for Bulgarian Language "Prof. Lyubomir Andreychin"
Bulgarian Academy of Sciences
cvetana@dcl.bas.bg

Abstract

The paper offers an approach to the validation of the data resulted from a previous effort on expansion of WordNet noun semantic classes by mapping them with the semantic types within the Corpus Pattern Analysis (CPA) ontology employed by the framework of the Pattern Dictionary of English Verbs (PDEV). A case study is presented along with a set of conditions to be checked when validating the combined data.

Keywords: WordNet, semantic classes, Corpus Pattern Analysis, semantic types

1. Introduction

The present work discusses the validity of the results of an effort ((Koeva et al., 2018)) on enriching WordNet through expansion of the WordNet noun semantic classes by mapping the WordNet data ((Miller, 1990)) with the data in another resource – the semantic types within the Corpus Pattern Analysis (CPA) ontology that is used by the Pattern Dictionary of English Verbs (PDEV) (Hanks, 2004); (Hanks and Pustejovsky, 2005); (Hanks, 2008). The discussion builds on results of work described in (Koeva et al., 2018), (Koeva et al., 2019b), (Koeva et al., 2019a), where the PDEV verb patterns were further automatically mapped to WordNet sentence frames thus adding information about the character of the arguments – the resulting patterns are considered conceptual frames whose arguments were specified for a set of lexical units – the semantic types assigned to the WordNet noun synsets. As far as we are aware, there were no previous attempts at mixing WordNet and CPA ontology, although there are proposals at mixing up information in WordNet and other resources (Longman Dictionary of Contemporary English (Dorr, 1997); (Korhonen, 2002); VerbNet and FrameNet (Shi and Mihalcea, 2005); VerbNet and PropBank ((Pazienza et al., 2006).

The main relation among words in the lexical-semantic network WordNet ((Miller et al., 1993), (Fellbaum, 1998)) is synonymy (or near-synonymy; synonyms are defined as words which denote the same concept and are interchangeable in many (but not all) contexts). The synonyms (called 'literals') are grouped into unordered sets (synsets) which are linked via the so-called 'conceptual relations'. Most relations between synsets connect words of the same part-of-speech (POS). Noun synsets are linked via hypernymy / hyponymy (superordinate) relation, and meronymy / holonymy (part-whole) relation. Verb synsets are arranged into hierarchies via hypernymy / hyponymy relation. Adjectives are organised in terms of antonymy and similarity, and relational adjectives ('pertainyms') are linked to the nouns they are derived from. Adverbs are linked to each other via similarity and antonymy relations.

In addition, nouns in WordNet are organised within the superordinate / subordinate (hypernymy / hyponymy) hierarchy which is limited in depth. Distinguishing features are added to create lexical inheritance system where each word inherits the distinguishing features (attributes (modification), parts (meronymy), functions (predication) from its superordinates (Miller, 1990). An example would be {chef:1}[1], which, as a hyponym of {cook:6}, could be an Agent of the verb synsets {cook:1}, {cook:3},

[1]Throughout the paper, the numbers of the literals follow those applied in the database used by the viewer Hydra available at: http://dcl.bas.bg/bulnet/. We do not give all literals and definitions due to space limitation. There may be changes to semantic

and {cook:4} just like its hypernym.
Noun synsets in WordNet are classified into 26 semantic classes (primitives or primes, (Miller, 1990)), namely nouns denoting humans (*noun.person*), animals (*noun.animal*), actions (*noun.act*), feelings and emotions (*noun.feeling*), etc.
The present paper discusses possible ways to check the validity of the data resulted from a previous effort to enrich the data in WordNet through expansion of the noun semantic classes by merging the WordNet data with the data in the CPA hierarchy used in the Pattern Dictionary of English Verbs (PDEV). The WordNet was enriched through merging the WordNet concepts and the Corpus Pattern Analysis (CPA) semantic types. The 253 CPA semantic types were mapped to the respective WordNet concepts. As a result of the mapping, the hyponyms of a synset to which a CPA semantic type was mapped, inherit not only the respective WordNet semantic class but also the CPA semantic type. The resources and the mapping are described in section (2) following (Koeva et al., 2018) – the resulting data on which the discussion in section (3) is based on, is publicly available at: http://dcl.bas.bg/PWN_CPA/ [2].

2. The mapping

2.1. WordNet noun hierarchy vs. CPA ontology

As already mentioned, noun synsets in WordNet are organized into the following 26 semantic classes (primitives or primes, (Miller, 1990)) – given in Table 1:

Nouns denoting	**semantic class**	**Nouns denoting**	**semantic class**
humans	*noun.person*	animals	*noun.animal*
plants	*noun.plant*	foods and drinks	*noun.food*
actions	*noun.act*	natural processes	*noun.process*
feelings and emotions	*noun.feeling*	cognitive processes	*noun.cognition*
spatial position	*noun.location*	time and temporal relations	*noun.time*
man-made objects	*noun.artifact*	natural objects	*noun.object*
body parts	*noun.body*	substances	*noun.substance*
possession and transfer of possession	*noun.possession*	quantities and units of measure	*noun.quantity*
relations between people or objects or ideas	*noun.relation*	natural phenomena	*noun.phenomenon*
groupings of people or objects	*noun.group*	two and three dimensional shapes	*noun.shape*
goals	*noun.motive*	communicative processes and contents	*noun.communication*
natural events	*noun.event*	attributes of people and objects	*noun.attribute*

Table 1: WordNet noun semantic classes.

The synsets labeled *noun.Tops* are the top-level synsets in the hierarchy – these are the so-called unique beginners which divide the nouns into (sub-)hierarchies as illustrated in Table 2:

classes between the PWN and the version on http://dcl.bas.bg/bulnet/, for detail see (Leseva et al., 2015).

[2] To every noun <SYNSET> element there are <CPA> elements assigned.

noun.Tops	**hyponyms**	**semantic class of the hyponym(s)**
{entity:1}	{physical entity:1}	*noun.Tops*
	{abstraction:1; abstract entity:1}	*noun.Tops*
	{thing:4}	*noun.artifact*
{physical entity:1}	{thing:1}	*noun.Tops*
	{object:1; physical object:1}	*noun.Tops*
	{causal agent:1; cause:1; causal agency:1}	*noun.Tops*
	{matter:1}	*noun.substance*
	{substance:7}	*noun.substance*
	{process:1; physical process:1}	*noun.process*
{thing:1}	*varying...*	*noun.object*
{object:1; physical object:1}	*varying...*	*noun.object*, *noun.artifact*
{causal agent:1; cause:1; causal agency:1}	*varying...*	*noun.person*, *noun.phenomenon*, *noun.state*, *noun.object*, *noun.substance*
{matter:1}	*varying...*	*noun.substance*, *noun.object*
{abstraction:1; abstract entity:1}	{psychological feature:1}	*noun.attribute*
	{attribute:1}	*noun.attribute*
	{group:1; grouping:1}	*noun.group*
	{relation:1}	*noun.relation*
	{communication:1}	*noun.communication*
	{measure:1; quantity:1; amount:1}	*noun.quantity*
	{otherworld:1}	*noun.cognition*
	{set:41}	*noun.group*

Table 2: (Sub-)hierarchies under *noun.Tops*.

Different entities may inherit information for their features from different sub-hierarchies as they may have more than one hypernym, as in (1) where the two hypernyms are members of two sub-hierarchies which can be followed down to two main opposite concepts – {physical entity:1} and {abstract entity:1}:

(1)
{substance:1} *noun.substance*
hypernym: {matter:1}
...hypernym: {physical entity:1}
hypernym: {part:18; portion:7; component part:1; component:3; constituent:6} *noun.relation*
...hypernym: {abstract entity:1}

The second resource that was mapped, is the PDEV framework which employs the so-called semantic types which are members of the Corpus Pattern Analysis (CPA) ontology. The CPA semantic types refer to properties shared by a number of nouns that are found in the argument positions of verb patterns

and are formulated if repeatedly observed in verb patterns in a corpus – so these are corpora-based. The CPA semantic types are organized into a relatively shallow ontology (up to 10 sublevels for some types), where the top-level type is [Anything] which has six (sub)types: [Entity], [Eventuality], [Group], [Part], [Property], and [Not_Connected]. These are further subcategorised as follows:

Type	**Subtypes**
[Entity]:	[Abstract_Entity], [Energy], [Physical_Object], [Particle], [Self]
[Eventuality]:	[Event], [State_of_Affairs]
[Group]:	[Human_Group], [Vehicle_Group], [Animal_Group], [Physical_Object_Group]
[Part]:	[Language_Part], [Music_Part], [Physical_Object_Part], [Speech_Act_Part], [Document_Part], [Movie_Part], [Recording_Part]
[Property]:	[Cognitive_State], [Role], [Visible_Feature], [Character_Trait], [Injury], [Institution_Role], [Pace], [Use], [Weight]
[Not_Connected]	

Each semantic type inherits the formal property of the type above it in the hierarchy ((Cinkova and Hanks, 2010)). The CPA semantic types represent cognitive concepts in the context of their use but they are not linked to sets of concrete concepts and their lexical representations – this is achieved through mapping the CPA with WordNet.

2.2. Mapping CPA ontology and WordNet noun hierarchy

The WordNet noun synset hierarchy was mapped onto the semantic type hierarchy in the CPA ontology by matching the CPA semantic types with the WordNet synsets. The matching was done manually – the most probable (according to the definition) and populated (i.e., having the most hyponyms) synset was matched to a CPA semantic type by two independent annotators with a third annotator validating the cases of disagreement; the resulting assignments are inherited along the whole hypernym / hyponym 'tree'. The semantic types borrowed from the CPA ontology were added – as complementary semantic primitives – to the WordNet semantic classes (the process is described in (Koeva et al., 2018)).
As a result, the hyponyms of a synset to which a CPA semantic type is mapped, are labeled by the respective WordNet semantic class and the CPA semantic type as in (2) – the assigned CPA semantic type [Artifact] encodes the information that the New York State Barge Canal is an artificial system:

(2)
{New York State Barge Canal:1} 'a system of canals crossing New York State and connecting the Great Lakes with the Hudson River and Lake Champlain' *noun.location*
[Artifact], [Watercourse], [Waterway]

The 253 CPA semantic types are mapped to the respective WordNet concept (synset) with: (a) 199 semantic types mapped directly to one concept; (b) 39 semantic types mapped to two WordNet concepts ([Route] is mapped to {road:2; route:4} 'an open way (generally public) for travel or transportation', *noun.artifact*, and {path:3; route:5; itinerary:3} 'an established line of travel or access', *noun.location*); (c) 12 semantic types are mapped to three concepts; 2 semantic types are mapped to four concepts; and 1 semantic type is mapped to five concepts. Not each CPA semantic type can be mapped to one synset but the hyponyms of the respective nodes in the WordNet hierarchy inherit the semantic specifications of the specific class.

It is assumed that the concepts in WordNet are divided into {abstract entity:1} and {physical entity:1} [3], thus the CPA types are marked as follows (matching the CPA subtypes in the respective sub-hierarchies with probable noun synset(s), which are linked to either of the two *noun.Tops*; some types – [Energy], [Self], [Event] – involve subtypes that are matched to WordNet concepts that can be traced back to both {abstract entity:1} and {physical entity:1}):

CPA	WordNet
[Entity]	entity:1
[Abstract_Entity]	abstract entity:1
[Energy]	abstract entity:1, physical entity:1
[Physical_Object]	physical entity:1
[Particle]	physical entity:1
[Self]	abstract entity:1, physical entity:1
[Eventuality]	abstract entity:1
[Event]	abstract entity:1, physical entity:1
[State_of_Affairs]	abstract entity:1
[Group]	abstract entity:1
[Part]	abstract entity:1
[Language_Part]	abstract entity:1
[Physical_Object_Part]	physical entity:1
[Property]	abstract entity:1

As a result, the new semantic types borrowed from the CPA ontology were added to the WordNet structure as complementary semantic classes. The semantic types of the hypernym were inherited by its hyponyms if the hyponyms was not assigned other semantic types – for example, {wine:4; vino:1} was assigned the types [Part], [Drug], [Abstract_Entity], [Wine], [Food] because there is the CPA type [Wine] while its co-hyponym {home brew:1; homebrew:1} was assigned [Part], [Drug], [Abstract_Entity], [Alcoholic_Drink], [Food] types that are inherited from the hypernym {alcohol:1; alcoholic drink:1; alcoholic beverage:1}. However, certain errors and mismatches were found in the hypernym / hyponym structure under the top-level concepts as not every of their hyponyms instantiates another hypernym / hyponym tree.

3. Validation

Following the mapping of the CPA semantic types to the Wordnet noun hierarchy with the hyponyms inheriting the CPA type, we checked whether the assigned CPA semantic type was the correct one according to a number of conditions drawing upon already available data in WordNet. Since a synset may be assigned one or more CPA semantic types, an error may arise at each assignment.

3.1. Noun.foods

To illustrate, we will focus our attention on the noun synsets that refer to foods and drinks and are (mostly) labeled *noun.food* with the assigned CPA semantic type of [Food]. Table 1 gives the numbers of synsets assigned the type [Food] in a combination with other CPA semantic types.

[3] The third synset under the hypernym {entity:1} – {thing:4} which is classified as *noun.artifact* – comprises 8 synsets only.

CPA semantic type	**No**	**Examples**
[Stuff], [Food]	978	fast food, meal, wiener roast (*noun.food*)
[Natural_Landscape_Feature], [Stuff], [Food]	2	multivitamin, vitamin pill (*noun.food*)
[Part], [Abstract_Entity], [Stuff], [Food]	3	milk, mother's milk, colostrum (*noun.body*)
[Human_Group], [Abstract_Entity], [Stuff], [Food]	1	power breakfast (*noun.group*)
[Drug], [Stuff], [Food]	1	powdered mustard (*noun.substance*)
[Solid], [Food]	726	takeout, sugarloaf, quiche, cherry tomato (*noun.food*)
[Part], [Abstract_Entity], [Material], [Food]	2	fish meal, pantothenic acid (*noun.substance*)
[Part], [Abstract_Entity], [Beverage], [Food]	97	coffee substitute, chicory, cow's milk (*noun.food*)
[Part], [Drug], [Abstract_Entity], [Beverage], [Food]	2	elixir, elixir of life (*noun.food*)
[Part], [Drug], [Abstract_Entity], [Alcoholic_Drink], [Food]	164	malt liquor, kvass (*noun.food*)
[Part], [Abstract_Entity], [Water], [Food]	12	bottled water, sparkling water (*noun.food*)
[Solid], [Natural_Landscape_Feature], [Tree_Part], [Food]	197	edible fruit, strawberry, apple (*noun.food*)

Table 3: Foods and drinks.

The least populous combinations are the ones whose members are synsets of another semantic class, different from the expected *noun.food*.

In six combinations, there is an [Abstract_Entity] semantic type and in two – [Natural_Landscape_Feature] semantic type. In some cases, the [Abstract_Entity] is admissible as with 'power breakfast' and 'elixir' and 'elixir of life' while the [Natural_Landscape_Feature] can be applied to natural objects which is true in the case of fruit but is not applicable to the vitamil pill.

WordNet is heavily anthropocentric, thus {milk:4} 'produced by mammary glands of female mammals for feeding their young' is *noun.body* and [Part], [Abstract_Entity], [Stuff], [Food] but {milk:5} 'a white nutritious liquid secreted by mammals and used as food by human beings' is *noun.food* and [Part], [Abstract_Entity], [Beverage], [Food] – the same difference is kept between {mother's milk:1} and {cow's milk:1}. Here, the [Abstract_Entity] type is not appropriate, though.

Further, the vegetables are classified as solid foods, while fruit are solid foods which are parts of a tree – [Tree_Part] (including berries). For example, {edible fruit:1}, which is *noun.food* and [Solid], [Natural_Landscape_Feature], [Tree_Part], [Food], inherits its features (semantic types) from its two hypernyms – {produce:8; green goods:1; green groceries:1; garden truck:1} 'fresh fruits and vegetable grown for the market' which is *noun.food* and [Solid], [Food], and {fruit:5} 'the ripened reproductive body of a seed plant' which is *noun.plant* and [Natural_Landscape_Feature], [Tree_Part].

Drinks can be classified as abstract entities (with the semantic type [Abstract_Entity]) as many of them are man-made products. For example, {sparkling water:1} is *noun.food* and [Part], [Abstract_Entity], [Water], [Food] while {tap water:1} is also *noun.food* but [Part], [Abstract_Entity], [Material], [Water].

Further, there is {power breakfast:1} 'a meeting of influential people to conduct business while eating breakfast' which is *noun.group* but has two hypernyms whose semantic types it inherits: {breakfast:3} which is *noun.food* and {meeting:5; get together:5} which is *noun.group*. On the other hand, there is {dinner:2; dinner party:1} 'a party of people assembled to have dinner together' which has only one hypernym {party:3}, therefore the semantic types it inherits are only [Human_Group], [Abstract_Entity]. Such single instances should be taken into consideration.

3.2. Steps for validation

In order to check the correctness of the assigned CPA semantic types, we have to observe several conditions in relation to the data encoding information that is already available in WordNet, in the following order:
1. Check whether the WordNet semantic class is compatible with the assigned CPA semantic type, as in (3) where *noun.food* projects to [Food] but not in (4) where the semantic class is *noun.group*.
(3)
{dish:6} *noun.food* [Stuff], [**Food**] (True)
(4)
{power breakfast:1} *noun.group* [Human_Group], [Abstract_Entity], [**Stuff**] [**Food**] (False)

2. Check whether there are literals in the synset that are compatible with the assigned CPA semantic type, as in (5) where the literal is the same.
(5)
{alcohol:1; alcoholic drink:1; alcoholic beverage:1} *noun.food* [**Alcoholic_Drink**] [Food] (True)

3. Check whether the hypernym synset is assigned a CPA semantic type that is compatible, as these are inherited as in (6) where the *noun.food* {liqueur:1; cordial:1} inherits the type [Alcoholic_Drink] from its hypernym and transfers it to its hyponym {absinth:1; absinthe:1}.
(6)
{absinth:1; absinthe:1} *noun.food* [**Alcoholic_Drink**] [Food]
hypernym: {liqueur:1; cordial:1} *noun.food* [**Alcoholic_Drink**] [Food]
hypernym: {alcohol:1; alcoholic drink:1; alcoholic beverage:1} *noun.food* [**Alcoholic_Drink**] [Food] (True)

4. Check the inheritance along the hypernym / hyponym tree as in (7) where {mother's milk:1} inherits [Abstract_Entity] semantic type from its hypernym {milk:4} which, on its turn, has inherited it from one of its own two hypernyms. However, this semantic type is found much further down the tree and is probably not applicable here – such cases should be studied and eventually corrected.
(7)
{mother's milk:1} noun.food [**Abstract_Entity**] (False), [Stuff], [**Food**] (True)
hypernym: {milk:4} *noun.body* (with two hypernyms)
hypernym: {liquid body substance:1; bodily fluid:1; body fluid:1; humor:4; humour:4} *noun.substance* [**Abstract_Entity**] (False), [**Stuff**] (True)
hypernym: {body substance:1} *noun.body*
hypernym: {substance:1} *noun.substance* (with two hypernyms)
hypernym: {matter:1} *noun.substance*
hypernym: {physical entity:1}
hypernym: {part:18; portion:7; component part:1; component:3} *noun.relation*
hypernym: {relation:1} *noun.relation*
hypernym: {abstract entity:1} [**Abstract_Entity**]
hypernym: {nutriment:1; nourishment:2; nutrition:2; sustenance:2; aliment:2; alimentation:2; victuals:2} *noun.food* [Stuff], [Food]
. . . hypernym: {physical entity:1}

In all the data, there is persistent assignment of the semantic type of [Natural_Landscape_Feature], as in (8):
(8)
{paring:1} *noun.food* [Natural_Landscape_Feature]
{multivitamin:1; multivitamin pill:1} *noun.food* [Natural_Landscape_Feature], [Stuff], [Food]

In addition, there are other types of foods that are not classified with [Food] semantic type but with types referring to their source ([Meat]) or the fact that they are part of some other entity ([Quantity]) – see Table 4.

CPA semantic type	**No**	**Examples**
[Solid], [Meat]	198	sirloin steak, roast, confit (*noun.food*)
[Quantity]	193	wing, turkey wing, oyster, cutlet (*noun.food*)

Table 4: Foods not assigned [Food] type.

Information about the classification of these concepts as foods is already available from the WordNet semantic class. This means that somewhere down the hypernym / hyponym tree there is a synset which contains a literal that repeats the semantic class at hand (e.g., {substance:1} and {substance:2} have a semantic class of *noun.substance*, {person:1} has a semantic class of *noun.person*, {artifact:1; artefact:1} has a semantic class of *noun.artifact*, etc.).
The mapping can be additionally applied to the following pairs of WordNet semantic classes and CPA semantic types: *noun.substance* = [Stuff]; *noun.person* = [Human]; *noun.artifact* = [Artifact]; *noun.plant* = [Plant]; *noun.animal* = [Animal]; *noun.location* = [Location]; *noun.group* = [X_Group] ([Human_Group], [Physical_Object_Group], [Animal_Group], etc.), *noun.time* = [Time_Period], etc.

4. Conclusion

The paper discussed the the results of an effort on enriching the WordNet through expansion of the noun semantic classes by mapping the WordNet data with the semantic types within another corpus-based ontology within the Corpus Pattern Analysis (CPA). The validity of the results was checked on the basis of synsets about food and drinks, a couple of erroneous assignments was discussed along with the conditions behind these assignments based on the data in WordNet.
Although the CPA semantic types may add explicit information to WordNet semantics, this information is already available on different levels in the WordNet structure.

Acknowledgements

The work is funded under the project "Towards a Semantic Network Enriched with a Variety of Relations" (DN 10-3 / 14.12.2016), financed by the Bulgarian National Science Fund (BNSF).

References

Cinkova, S. and Hanks, P. (2010). *Validation of Corpus Pattern Analysis – Assigning Pattern Numbers to Random Verb Samples*. `http://ufal.mff.cuni.cz/project/spr/data/publications/annotation_manual.pdf`.

Dorr, B. J. (1997). Large-scale Dictionary Construction for Foreign Language Tutoring and Interlingual Machine Translation. *Machine Translation*, 12(4):271–322.

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

Hanks, P. and Pustejovsky, J. (2005). A Pattern Dictionary for Natural Language Processing. *Revue française de linguistique appliquée*, 10 (3):63–82.

Hanks, P. (2004). Corpus Pattern Analysis. In *Proceedings of Euralex*, pages 87–97.

Hanks, P. (2008). Mapping Meaning onto Use: A Pattern Dictionary of English Verbs. In *Proceedings of the AACL 2008, Utah*.

Koeva, S., Dimitrova, T., Stefanova, V., and Hristov, D. (2018). Mapping WordNet concepts with CPA ontology. In *Proceedings of the 9th Global WordNet Conference (GWC'2018)*, pages 70–77.

Koeva, S., Dimitrova, T., Stefanova, V., and Hristov, D. (2019a). Towards Conceptual Frame. *Chuzhdoezikovo obuchenie*, 46(6):551–564.

Koeva, S., Hristov, D., Dimitrova, T., and Stefanova, V. (2019b). Enriching Wordnet with Frame Semantics. In *Dokladi ot Mezhdunarodnata godishna konferentsiya na Instituta za balgarski ezik Prof. Lyubomir Andreychin" (Sofiya, 14 – 15 may 2019 godina)*, pages 300–308.

Korhonen, A. (2002). Assigning Verbs to Semantic Classes via Wordnet. In *Proceedings of the 2002 Workshop on Building and Using Semantic Networks*, volume 11.

Leseva, S., Stoyanova, I., Todorova, M., Dimitrova, Tsvetana, R., and Koeva, S. (2015). Automatic classification of WordNet morphosemantic relations. In *The 5th Workshop on Balto-Slavic Natural Language Processing*, pages 59–64.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K., Eds. (1993). *Introduction to WordNet: an On-line Lexical Database. Five Papers on WordNet*. Princeton, NJ: Princeton University.

Miller, G. (1990). Nouns in WordNet: a lexical inheritance system. *International journal of Lexicography*, 3.4:254–264.

Pazienza, M. T., Pennacchiotti, M., and Zanzotto, F. M. (2006). Mixing wordnet, verbnet and propbank for studying verb relations. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006)*.

Shi, L. and Mihalcea, R. (2005). Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In *Computational linguistics and Intelligent Text Processing*, page 100–111.

Organised by:

Department of Computational Linguistics
Institute for Bulgarian Language

Institute for Information and Communication Technologies

Bulgarian Academy of Sciences

The **Special Session on Wordnets and Ontologies** at the Fourth International Conference Computational Linguistics in Bulgaria (CLIB 2020) is organised with the support of the National Science Fund of the Republic of Bulgaria under the project *Towards a Semantic Network Enriched with a Variety of Relations*, Grant Agreement DN10/3/2016.

NATIONAL SCIENCE FUND
Ministry of Education and Science

ISSN: 2367-5675