

FIFTH  
INTERNATIONAL  
CONFERENCE



**COMPUTATIONAL  
LINGUISTICS  
IN BULGARIA  
CLIB 2022**

**8 – 9 September 2022**

**Sofia, Bulgaria**

Organiser:



Department of Computational Linguistics  
Institute for Bulgarian Language  
Institute of Information and Communication Technologies  
**BULGARIAN ACADEMY OF SCIENCES**

**PROCEEDINGS**

The Fifth International Conference *Computational Linguistics in Bulgaria* (CLIB 2022) is organised with the support of the National Science Fund of the Republic of Bulgaria under Grant Agreement No. KP-06-MNF/7 of 20.07.2022.



The National Science Fund does not take responsibility for the contents of the papers presented at the Conference or for any of the Conference materials.

CLIB 2022 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 Fifth International Conference <i>Computational Linguistics in Bulgaria</i> (CLIB 2022)
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
Copyright:	Copyright of each paper stays with the respective authors. The works in the Proceedings are licensed under a Creative Commons Attribution 4.0 International Licence (CC BY 4.0).  License details: <a href="http://creativecommons.org/licenses/by/4.0">http://creativecommons.org/licenses/by/4.0</a> Copyright © 2022

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



8 – 9 September 2022  
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

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

**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, France

**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

**A. Seza Dođruöz** – Ghent University, Belgium

**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, Greece

**Stefan Gerdjikov** – Sofia University, Faculty of Mathematics and Informatics, Bulgaria

**Voula Giouli** – Institute for Language and Speech Processing, ATHENA Research Centre, Greece

**Ivan Koychev** – Sofia University, Faculty of Mathematics and Informatics, Bulgaria

**Cvetana Krstev** – University of Belgrade, Faculty of Philology, Serbia

**Eric Laporte** – University of Paris-Est Marne-la-Vallée, France

**Natalia Loukachevitch** – Research Computing Center of Moscow State University, Russia

**John P. McCrae** – National University of Ireland, Galway, Ireland

**Preslav Nakov** – Qatar Computing Research Institute, HBKU, Qatar

**Maciej Piasecki** – Wrocław University of Technology, Poland

**Vito Pirrelli** – Institute for Computational Linguistics, ILC-CNR, Italy

**Ewa Rudnicka** – Wrocław University of Technology, Poland

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

**Stan Szpakowicz** – University of Ottawa, Canada

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

**Hristo Tanev** – Joint Research Centre of the European Commission, Italy

**Irina Temnikova** – Big Data for Smart Society Institute (GATE), Bulgaria



**Tinko Tinchev** – Sofia University, Faculty of Mathematics and Informatics, Bulgaria

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

**Cristina Vertan** – University of Hamburg, Germany

**Katerina Zdravkova** – University St Cyril and Methodius in Skopje, North Macedonia

## **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

**Dimitar Hristov** – Cleversoft, Bulgaria

**Georgi Iliev** – Milestone Systems, Bulgaria

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

**Todor Lazarov** – New Bulgarian University

**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

<b>PLENARY TALKS</b> . . . . .	1
Prof. Shuly Wintner <i>The Hebrew Essay Corpus</i> . . . . .	2
Prof. Iryna Gurevych <i>Detect – Verify – Communicate: Combating Misinformation with More Realistic NLP</i> . . . . .	3
Prof. Bolette Sandford Pedersen <i>Lexical Conceptual Resources in the Era of Neural Language Models</i> . . . . .	4
Jose Manuel Gomez-Perez <i>Towards AI that Reasons with Scientific Text and Images</i>	5
<b>MAIN CONFERENCE</b> . . . . .	7
Hristo Tanev <i>OntoPopulis, a System for Learning Semantic Classes</i> . . . . .	8
Elena Callegari, Desara Khura <i>A corpus for Automatic Article Analysis</i> . . . . .	13
Timofey Atnashev, Veronika Ganeeva, Roman Kazakov, Daria Matyash, Michael Sonkin, Ekaterina Voloshina, Oleg Serikov, Ekaterina Artemova <i>Razmecheno: Named Entity Recognition from Digital Archive of Diaries “Prozhito”</i> . . . . .	22
Iglika Nikolova-Stoupak, Shuichiro Shimizu, Chenhui Chu, Sadao Kurohashi <i>Filtering of Noisy Web-Crawled Parallel Corpus: the Japanese-Bulgarian Language Pair</i> . . . . .	39
Radoslav Ralev, Jürgen Pfeffer <i>Hate Speech Classification in Bulgarian</i> . . . . .	49
Slavina Lozanova, Ivelina Stoyanova <i>WordNet-Based Bulgarian Sign Language Dictionary of Crisis Management Terminology</i> . . . . .	59
Petya Osenova <i>Raising and Control Constructions in a Bulgarian UD Parsebank of Parliament Sessions</i> . . . . .	68
Yovka Tisheva, Marina Dzhonova <i>Syntactic characteristics of emotive predicates in Bulgarian: A corpus-based study</i> . . . . .	75
Ekaterina Tarpomanova, Krasimira Aleksova <i>Evidential strategies and grammatical marking in clauses governed by verba dicendi in Bulgarian</i> . . . . .	81

Junya Morita <i>Corpus-Based Research into Verb-Forming Suffixes in English: Its Empirical and Theoretical Consequences</i> . . . . .	89
Ivan Derzhanski, Olena Siruk <i>Some Notes on p(e)re-Reduplication in Bulgarian and Ukrainian: A Corpus-based Study</i> . . . . .	98
Radu Ion, Andrei-Marius Avram, Vasile Păiș, Maria Mitrofan, Verginica Barbu Mititelu, Elena Irimia and Valentin Badea <i>An Open-Domain QA System for e-Governance</i> . . . . .	105
Daria Liakhovets, Sven Schlarb <i>Zero-shot Event Causality Identification with Question Answering</i> . . . . .	113
Svetla Koeva <i>Ontology of Visual Objects</i> . . . . .	120
Alexander Kirillovich, Natalia Loukachevitch, Maksim Kulaev, Angelina Bolshina, Dmitry Ilvovsky <i>Sense-Annotated Corpus for Russian</i> . . . . .	130
Verginica Barbu Mititelu, Mihaela Cristescu, Maria Mitrofan, Bianca-Mădălina Zgreabăn, Elena-Andreea Bărbulescu <i>A Romanian Treebank Annotated with Verbal Multiword Expressions</i> . . . . .	137
Aleksandar Petrovski <i>A Parallel English - Serbian - Bulgarian - Macedonian Lexicon of Named Entities</i> . . . . .	146
Silvia Gargova, Irina Temnikova, Ivo Dzhumerov, Hristiana Nikolaeva <i>Evaluation of Off-the-Shelf Language Identification Tools on Bulgarian Social Media Posts</i> .	152
Kamel Smaïli, David Langlois, Peter Pribil <i>Language rehabilitation of people with BROCA aphasia using deep neural machine translation</i> . . . . .	162
Travis Sorenson <i>Current Shortcomings of Machine Translation in Spanish and Bulgarian Vis-à-vis English</i> . . . . .	171
Cvetana Krstev, Duško Vitas <i>A Myriad of Ways to Say: "Wear a mask!"</i> . . . . .	181
Jordan Kralev, Svetla Koeva <i>Image Models for large-scale Object Detection and Classification</i> . . . . .	190
<b>SPECIAL SESSION ON WORDNETS, FRAMENETS AND ONTOLOGIES</b> . . . . .	202
Svetla Koeva, Emil Doychev <i>Ontology Supported Frame Classification</i> . . . . .	203
Svetlozara Leseva, Ivelina Stoyanova <i>Linked Resources towards Enhancing the Conceptual Description of General Lexis Verbs Using Syntactic Information</i> . . .	214
Matea Birtić, Ivana Brač, Siniša Runjaić <i>Croatian repository for the argument/adjunct distinction – SARGADA</i> . . . . .	225
Borislav Rizov, Tinko Tinchev <i>Towards Dynamic Wordnet: Time Flow Hydra</i> . . .	234



---

## PLENARY TALKS

---



## **The Hebrew Essay Corpus**

**Prof. Shuly Wintner (University of Haifa, Israel)**

---

The Hebrew Essay Corpus is an annotated corpus of Hebrew language argumentative essays authored by prospective higher-education students. The corpus includes both essays by native speakers, written as part of the psychometric exam that is used to assess their future success in academic studies; and essays authored by non-native speakers, with three different native languages, that were written as part of a language aptitude test. The corpus is uniformly encoded and stored. The non-native essays were annotated with target hypotheses whose main goal is to make the texts amenable to automatic processing (morphological and syntactic analysis).

I will describe the corpus and the error correction and annotation schemes used in its analysis. In addition, I will discuss some of the challenges involved in identifying and analyzing non-native language use in general, and propose various ways for dealing with these challenges. Then, I will present classifiers that can accurately distinguish between native and non-native authors; determine the mother tongue of the non-natives; and predict the proficiency level of non-native Hebrew learners. This is important for practical (mainly educational) applications, but the endeavor also sheds light on the features that support the classification, thereby improving our understanding of learner language in general, and transfer effects from Arabic, French, and Russian on nonnative Hebrew in particular.

## **Detect – Verify – Communicate: Combating Misinformation with More Realistic NLP**

**Prof. Iryna Gurevych (Technical University of Darmstadt, Germany)**

---

Dealing with misinformation is a grand challenge of the information society directed at equipping the computer users with effective tools for identifying and debunking misinformation. Current Natural Language Processing (NLP) including its fact-checking research fails to meet the expectations of real-life scenarios. In this talk, we show why the past work on fact-checking has not yet led to truly useful tools for managing misinformation, and discuss our ongoing work on more realistic solutions. NLP systems are expensive in terms of financial cost, computation, and manpower needed to create data for the learning process. With that in mind, we are pursuing research on detection of emerging misinformation topics to focus human attention on the most harmful, novel examples. Automatic methods for claim verification rely on large, high-quality datasets. To this end, we have constructed two corpora for fact checking, considering larger evidence documents and pushing the state of the art closer to the reality of combating misinformation. We further compare the capabilities of automatic, NLP-based approaches to what human fact checkers actually do, uncovering critical research directions for the future. To edify false beliefs, we are collaborating with cognitive scientists and psychologists to automatically detect and respond to attitudes of vaccine hesitancy, encouraging anti-vaxxers to change their minds with effective communication strategies.

## **Lexical Conceptual Resources in the Era of Neural Language Models**

**Prof. Bolette Sandford Pedersen (Copenhagen University, Denmark)**

---

Lexical conceptual resources in terms of e.g. wordnets, framenets, terminologies and ontologies have been compiled for many languages during the last decades in order to provide NLP systems with formally expressed information about the semantics of words and phrases, and about how they refer to the world. In most recent years, neural language models have become a game-changer in the NLP field – based, as they are, solely on text from large corpora. It is time we ask ourselves: What is the role of lexical conceptual resources in the era of neural language models? The claim of my talk is that they still play a crucial role since NLP systems based on textual distribution alone will always to some extent be insufficient and biased. Through my own work, which has over the years taken place in close collaboration with leading lexicographers in Denmark, I will illustrate how such conceptual resources can be compiled based on existing high-quality and continuously updated lexicographical resources and how they can be further curated by examining the distributional patterns captured in word embeddings.

## **Towards AI that Reasons with Scientific Text and Images**

**Jose Manuel Gomez-Perez (Expert.ai)**

---

Reading a textbook in a particular discipline and being able to answer the questions at the end of each chapter is one of the grand challenges of artificial intelligence, which requires advances in language, vision, problem-solving, and learning theory. Such challenges are best illustrated in the scientific domain, where complex information is presented over a variety of modalities involving not only language but also visual information, like diagrams and figures.

In this talk, we will analyze the specific challenges entailed in understanding scientific documents and share some of the recent advances in the area that enable the development of AI systems capable to answer scientific questions. In addition, we will reflect on what new developments will be required to address the next grand challenge: to create an AI system that can make major scientific discoveries by itself.





---

## MAIN CONFERENCE

---

# OntoPopulis, a System for Learning Semantic Classes

Hristo Tanev

Joint Research Centre, European Commission

via Enrico Fermi 2749

Ispira, Italy

hristo.tanev@ec.europa.eu

## Abstract

Ontopopulis is a multilingual weakly supervised terminology learning algorithm which takes on its input a set of seed terms for a semantic category and an unannotated text corpus. The algorithm learns additional terms, which belong to this category. For example, for the category “environmental disasters” the input seed set in English is *environmental disaster*, *water pollution*, *climate change*. Among the highest ranked new terms which the system learns for this semantic class are *deforestation*, *global warming* and so on.

Keywords: semantic classes, ontology learning, terminology extraction

## 1 Introduction

Ontologies are knowledge-representation models describing concepts of a domain, their properties and the semantic relations between them. These models are used in Natural Language Processing and other AI systems. Recently ontologies have been built and exploited predominantly in the area of biology and medicine. There are also many other domains for which they have been used: remote sensing, education, environment and security, etc.

The most fundamental building block of the ontology model are the concepts of the domain. Each concept has lexical representation, which shows how the concept is being referred at the level of a specific language. For example, the concept [TV-SET] in English can be referred to as: *TV*, *TV set*, *television receiver*, and *telly*. Words which describe ontology concepts form the lexical layer of the ontology.

Ontologies are created mainly manually by domain experts, however in populating their

lexical layer, terminology learning algorithms have successfully been exploited. In this paper we will describe such a multilingual algorithm which given a set of ontology concepts learns new set of related concepts.

For each concept and language under consideration, the language experts define a small seed set of terms, belonging to the concept and its sub-concepts. For example for the concept *disaster* the seed set for English can be: *disaster*, *flood*, *earthquake*, *forestfire*, *wildfire*.

The seed set is then expanded by the algorithm by learning new terms, referring to the same main concept (*disaster*) and its sub-concepts, for example it will learn words like *calamity*, *tsunami*, *landslide*. These terms belong to the category *disaster* and its sub-categories *tsunami* and *landslide*

The algorithm was named *OntoPopulis* (ONTOlogy learning and POPULation).

In this paper we describe the algorithm and its application in the domain of environment for English.

## 2 Related work

The first algorithms for ontology learning from text started to appear about 20 years ago. Currently, many approaches are described in the literature; they are designed to learn generic and domain-specific ontology resources.

One of the first comprehensive overview of these approaches is presented by (Cimiano, 2006).

More recent surveys of ontology learning from text are presented in (Lourdusamy and Abraham, 2019) and (Al-Aswadi et al., 2020).

The approach in this paper is inspired by an earlier work, described in (Tanev and Magnini,

2006).

### 3 OntoPopulis

OntoPopulis is a multilingual algorithm for learning semantic classes. It does not use any language-specific tools or annotations. The algorithm accepts as an input a set of seed terms for each ontology concept under consideration and an unannotated corpus. For example, for the concept *disaster* the seed set is *disaster*, *wildfire*, *earthquake* and for the concept *environmental disaster* it is *environmental disaster*, *water pollution*, *climate change*

The algorithm performs two processing steps to learn the new lexical items for the input concepts: (i) feature extraction and (ii) lexical learning.

#### 3.1 Feature Extraction and Weighting

For each category (e.g. environmental disaster), we consider left and right context features.

Each left context feature consists of uni-gram or bi-gram and can be followed by a preposition or another stop word. For example *primary cause of* is a left context feature; it occurs on the left side of words from the category *environmental disaster*. i.e. *primary cause of water pollution*

Similarly, right context feature appear on the right side of the seed terms, e.g. *and overfishing* is a right side context feature and it appears in phrases like *water pollution and overfishing*.

The left and right context features are weighted using a formula which considers the frequency of their co-occurrence with the seed terms.

Each context feature has to appear at least 3 times in the corpus with the seed set terms

For each such a context feature  $n$  and a semantic category  $C$  we calculate the score:

$$score(n, C) = \sum_{st \in C} PMI(n, st)$$

where  $seeds(C)$  are the seeds terms of the category  $C$  and  $PMI(n, st)$  is the point-wise mutual information which shows the co-occurrence between the feature  $n$  and the seed terms.

At the end of this learning phase there is a possibility for a linguist to perform manual feature selection from the list of the top ranked features. Manual cleaning is optional when high precision is the goal. In the reported experiments, however, we haven't used it.

Table 1 lists the top-ranked context feature for the semantic category environmental disaster

As one can observe, the context feature of OntoPopulis are very easy to evaluate semantically and linguistically. The table shows that these features come from semantic properties, predicates and related concepts of the considered semantic category.

#### 3.2 Term Extraction

The term extraction and learning stage takes the features, which were learned and manually selected for each category in the previous stage and extracts as candidate terms uni-grams and bi-grams, which frequently co-occur with these features and which do not contain stop words, numbers or capitalized letters. Weighting of the candidate terms was carried out with the view to optimize the efficiency of the calculations. For this reason, we avoid to obtain the frequency of each candidate term in the corpus and we rather calculate the term feature vector in a non-standard way. It would be statistically more correct to use as a feature weight the point-wise mutual information between the term and the feature. However, this would require to collect statistics about the term frequency, which will decrease the algorithm speed. We weighted the term candidates, using the following algorithm:

1. For each category  $C$  we define a feature space, whose dimensions are only the features selected for this category
2. For each category  $C$  we define a category feature vector

$$\vec{C} = (wf_1, wf_2, wf_3, \dots, wf_{nc})$$

where  $wf_i$  are the weights of the category features, calculated as  $wf_i = score(n_i, C)$ , where  $n_i$  is the  $n$ -gram used as  $i_{th}$  context feature in our model;  $score(n_i, C)$  is calculated with the point-wise-mutual-information based formula presented in the previous subsection.

Feature	Score
threatened by X	1.38
X and land degradation	0.87
impacts of X	0.79
pollution and X	0.75
impact of X	0.59
emissions and X	0.54
X and greenhouse	0.54
emissions and X	0.54
X and greenhouse	0.54
land use and X	0.52
contributor to X	0.51
X and global warming	0.48
primary cause of X	0.45
combating X	0.44
worst effects of X	0.42
degradation and X	0.38
X and other environmental	0.37
contributes to X	0.32
exacerbated by X	0.32
reducing X	0.31
global warming and X	0.29
X and overfishing	0.29
exposure to X	0.28
warming and X	0.27

Table 1: Top-ranked features for semantic category environmental disaster

3. We normalize each category feature vector  $\vec{C}$  by dividing its coordinates with its length and obtain its normalized form  $norm(\vec{C})$  (this is needed when several categories are considered at a time)
4. Then, for each candidate term  $t$  for the category  $C$  we define a term feature vector  $t_{\vec{C}} = (w_1, w_2, \dots, w_{nc})$  where

$$w_i = \frac{f_i}{f_i + 3}$$

, where  $f_i$  is the frequency with which term  $t$  appears with context feature  $i$ .

5. The weight for each candidate term  $t$  for a category  $C$  is defined as a scalar product in the vector space defined for the category  $C$ , multiplied by the square root of the number of the non-zero features of the term feature vector:

$$weight(t, C) = t_{\vec{C}} \cdot norm(\vec{C}) \cdot \sqrt{NNZF(t_{\vec{C}})}$$

, where  $NNZF$  returns the number of the non zero vector coordinates.

In plain words, this formula measures term suitability for a category by considering the co-occurrence of the term with the context features of this category and their weights.

6. Finally, the system orders the term candidates for each category by decreasing weight and filters out terms with a weight under a certain threshold.

## 4 Experiments

We run the Ontopopulis algorithm on a seed set of three words, modelling the concept *environmental disaster*: *environmental disaster*, *climate change*, *water pollution* The list of the highest ranked 27 newly learned terms is presented in table 2. The irrelevant ones are marked with asterisk. The relevant ones (77%)

shown in the table can be divided into two categories:

- hyponyms of the environmental disaster: global warming, deforestation, air pollution, acidification, rising sea, heat wave, ocean acidification, environmental degradation, desertification, warming temperatures, extreme weather, erosion, oil spills, habitat loss.
- factors, which cause environmental disasters: overfishing, illegal fishing, carbon emissions, greenhouse gases, noise.

The algorithm can be used in the process of building ontologies and semantic dictionaries for information extraction tasks, such as event detection, named entity recognition, sentiment analysis, etc. The algorithm can significantly speed up creation of language resources and it learns words which are rare and difficult to come up with by linguists.

In this paper we have evaluated this algorithm for English language, but it has no restriction on the language used, since it does not use any annotation or language-specific resources.

## References

- Fatima N Al-Aswadi, Huah Yong Chan, and Keng Hoon Gan. 2020. Automatic ontology construction from text: a review from shallow to deep learning trend. *Artificial Intelligence Review*, 53(6):3901–3928.
- Philipp Cimiano. 2006. *Ontology learning and population from text: algorithms, evaluation and applications*, volume 27. Springer Science & Business Media.
- Ravi Lourdusamy and Stanislaus Abraham. 2019. A survey on methods of ontology learning from text. In *International Conference on Information, Communication and Computing Technology*, pages 113–123. Springer.
- Hristo Tanev and Bernardo Magnini. 2006. [Weakly supervised approaches for ontology population](#). In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 17–24, Trento, Italy. Association for Computational Linguistics.



Term	Score
global warming	42.2
deforestation	21.4
overfishing	9.0
rising sea	8.8
<i>*naturaldisasters</i>	8.1
<i>*brexit</i>	5.4
<i>*covid</i>	4.4
greenhouse gases	4.2
air pollution	4.1
acidification	3.9
heat wave	3.8
environmental degradation	3.6
rising temperatures	3.4
ocean acidification	2.9
environmental damage	2.8
<i>*circumstances</i>	2.7
desertification	2.3
carbon emissions	2.2
illegal fishing	2.1
warming temperatures	2.1
extreme weather	2.1
erosion	2.1
<i>*globalization</i>	2.0
noise	2.0
<i>*fakenews</i>	2.0
oil spills	2.0
habitat loss	2.0

Table 2: Highest scored learned terms for semantic category environmental disaster

# A corpus for Automatic Article Analysis

**Elena Callegari**

Uni of Iceland / Árnagarður, Reykjavík  
SageWrite ehf. / Miðbær, Reykjavík  
ecallegari@hi.is

**Desara Xhura**

SageWrite ehf. / Miðbær, Reykjavík  
desara@sagewrite.com

## Abstract

We describe the structure and creation of the SageWrite corpus. This is a manually annotated corpus created to support automatic language generation and automatic quality assessment of academic articles. The corpus currently contains annotations for 100 excerpts taken from various scientific articles. For each of these excerpts, the corpus contains (i) a draft version of the excerpt (ii) annotations that reflect the stylistic and linguistics merits of the excerpt, such as whether or not the text is clearly structured. The SageWrite corpus is the first corpus for the fine-tuning of text-generation algorithms that specifically addresses academic writing.

**Keywords:** Natural Language Generation, Automatic quality assessment of text, Scientific articles, Academic writing

## 1 Introduction

The latest developments in Natural Language Processing (NLP) and Natural Language Generation (NLG) demonstrate a significant gain in performance on many domain-specific NLP tasks, by pre-training on a large corpus of text and fine-tuning using prompt engineering<sup>1</sup> in specific task (Liu et al., 2021)(Brown et al., 2020)(Han et al., 2021). The SageWrite corpus is a manually annotated corpus created as a training dataset for the development of automatic text-generation and quality-assessment tools for academic writing<sup>2</sup>.

When writing the different sections of an academic paper, authors often start by creating a rough draft or outline of what they want that section to say, which they then proceed to edit -and re-edit- until

<sup>1</sup>Prompt engineering is a way of fine-tuning, where the NLP algorithm gets fed with examples of input and expected results.

<sup>2</sup>In the future, the dataset could also be relevant for text summarization purposes similar to (Collins et al., 2017)

they are satisfied with it. An author writing the introduction of a linguistics paper may for example start by writing something along the lines of 1, which they will then proceed to edit until it looks something like 2:

1. *My intentions:*

*first: present core data on focus particles*  
*second, review different existing approaches*  
*3rd: say what I think about what works best*

2. My intentions in this article are threefold: first, to outline the key data that any successful account of focus particles should explain; second, to review existing approaches that attempt to account for these data; and third, to offer my own views about the direction any successful analysis should take.

Our primary goal is automate the process that leads from 1 to 2: we want to generate grammatical text starting from a rough draft of what the final text should look like. Put differently, what we aim to do is streamline the revision process that leads from 1 to 2. What is required to generate 2 out of 1 stands halfway between natural language generation out of a limited input (Qu, 2020) and advanced automatic paraphrasing(Palivela, 2021). Our secondary goal is to develop a classifier that can process scientific articles and automatically assess whether or not they exhibit certain qualities or flaws that we deem relevant to assess scientific publications, such as whether or not information is clearly presented and whether or not the text exhibits a good flow. Again, this is in an attempt to streamline the revision process: if the stylistics shortcomings of a paper are flagged automatically, the author(s) of said paper can more readily address them. The SageWrite corpus was created to assist in the training of both of these functionalities.

A first version of the corpus (version 0.1), consisting of 100 annotated excerpts, was published online in February 2022<sup>3</sup>. We plan on increasing the size of this dataset as more excerpts get annotated.

## 2 Text Selection

The 100 manually annotated excerpts were extracted from various types of academic articles. To obtain the excerpts, we first created a database containing scientific articles taken from Arxiv, PubMed, plus around 70 articles that we randomly selected from various disciplines in the Humanities. The articles taken from Arxiv were all dated March 2020 onwards.

To extract the excerpts, we wrote a Python program that automatically extracted excerpts of around 300 words from various points in an article. This was done to ensure that text belonging to various sections of a paper (e.g. introduction, abstract, conclusions) was included. Text was always selected from the beginning of a paragraph until the end of a paragraph. The average length of the excerpts was 193 words.

The excerpts were annotated by three annotators. As we wanted to work on academic texts, we hired annotators who had ties with academia and experience with academic writing. Accordingly, one of our annotators was a MA student, one was a university lecturer and one had a PhD degree. All annotators were also native speakers of (American) English.

Annotations were completed online on a dedicated platform where annotators could automatically log each part of the annotation for a given excerpt.

Annotators saw rotations consisting of one excerpt from a PubMed article, one from an Arxiv article and one from our Humanities articles. As we thought it would be interesting to see how different individuals would react to the same text, all annotators saw and hence annotated the same excerpts.

## 3 Structure of the Corpus

For each of the 100 excerpts, the corpus contains (A) three corresponding rough-draft versions of excerpt, each authored by a different annotator and (B) a list of tags that describe the stylistic and linguistic qualities of each excerpt.

<sup>3</sup><https://github.com/elenaSage/SageWrite0.1corpus>

### 3.1 The Drafts

When writing up a section of an academic paper, authors generally start out by writing a rough draft of what they want to say. Drafts are both lexically and syntactically different from the final version of a paper. (Bowen and Van Waes, 2020) and (Bowen and Thomas., 2020) used the key-logging software Inputlog ((Leijten and Van Waes, 2013)) to explore how seven MA students (four native speakers of English and three native speakers of Chinese, all enrolled at a British University) approach revisions when writing an academic paper. The authors discovered that drafts feature fewer subordinates, adverbials and nominal modifiers than the finished articles. For example, some sentence-initial adverbial clauses ((underlined in 3, ex. from Bowen & Van Waes: 348) and some sentence-initial adverbials ((underlined in 4, ex. Bowen & Van Waes: 349) do not appear in the initial draft but are only added during the revision stage. Based on our own experience with academic writing, we also expect drafts to contain various types of abbreviations (e.g. 5), to be more schematic in nature (e.g. articles, copulas, 1st person singular pronouns may be dropped (6a), or arrows and empty lines may be used in place of some types of adverbials (6b)), and to contain instances of colloquial language that do not appear in the final version of a paper ((8).

3. (a) **Draft**  
"Research in this area has also looked at the differences between collectivist and individualistic countries."
- (b) **Finished Paper**  
"As well as looking at the differences in class, research in this area has also looked at the differences between collectivist and individualistic countries."
4. (a) **Draft**  
"As previously mentioned, because of the close family bond (...)"
- (b) **Finished Paper**  
"However, as previously mentioned, because of the close family bond (...)"
5. (a) Contractions: "that's" vs. "that is"
- (b) Colloquialisms: "cause" for "because", "w" for "with"
- (c) Other types of abbreviations: "mvt" for "movement", "foc" for "(linguistic) Focus"

6. (a) "in sect 1, will be talking about foc marking patterns in Malayalam"
- (b) "in sect 1 -> foc patterns in Malayalam"
7. (a) **Draft**  
"In the essay Racisms, Kwame Anthony Appiah says what he thinks about the topic"
- (b) **Finished Paper**  
"In the essay Racisms, Kwame Anthony Appiah provides his thoughts on this issue."
8. (a) **Draft**  
"But are MCI patients actually aware of their cognitive deficits? That's debatable"
- (b) **Finished Paper**  
"However, whether patients with MCI are truly aware of the full extent of their cognitive deficits is a matter of debate"

We asked our three annotators to read each excerpt and try to reverse-engineer what the draft version of that excerpt might have looked like, and to write that down. We asked them to experiment with different drafting styles; for example, we explained that while some authors might use lots of abbreviations, others might prefer to spell out every or most words. While some authors might use extremely colloquial language, others might prefer to adhere to academic lexical standards already in earlier versions of a paper.

An example of an original text and the draft one of the annotators created can be seen in 9 (original excerpt from (Keay and Hind, 2020), page 5):

9. (a) **Original Text**  
"Participants (n=88) were recruited from clients attending a private physiotherapy clinic in Bath, United Kingdom. The physiotherapy clinic provides physiotherapy, strength and conditioning programmes and clinical input for a range of conditions, including those exercisers with suspected low energy availability. Invitation for participants was also disseminated through contacts in the vicinity such as university, sport clubs and healthcare providers referring to the physiotherapy practice. The inclusion criteria were males and females over the age of 20. The study was approved by by

the university research ethics committee and all participants provided informed consent prior to taking part."

- (b) **Draft Text**  
"Participants: n=88; recruited from client pool of private physio clinic, Bath, UK. Physio clinic offers physiotherapy, strength and conditioning, clinical input for many conditions, including exercisers with possible low energy availability. Also invited participants through local contacts at university, sports clubs, healthcare providers that refer to the physio clinic. Inclusion criteria: males/females >20 years Approved by uni research ethics committee; all subjects gave informed consent before participation."

As we are dealing with academic text, our goal is to develop NLG tools that do not generate too much beyond the original input: should the AI generate too much on top of the initial input provided by the user, one could question whether the resulting generated text is truly the work of the author or rather should be considered the work of the AI. Because of these concerns, we instructed our annotators not to leave out non-recoverable information from the drafts. For example, information occurring between parentheses in the original text was always included in the corresponding draft version (see 10).

10. (a) **Original Text**  
"It also presents methods that may be used for analyzing language interplays in general (**demonstrated using the PDT data**)"
- (b) **Draft Version** (as by Annotator 2)  
"Present methods to analyze language interplays in general (**see PDT**)"

Annotators first practiced annotation on a set containing 50 sample excerpts. During this practice run, annotators got direct feedback by the authors of this paper, who reviewed the annotations of the sample excerpts. These 50 practice excerpts are not included in the dataset we published online.

### 3.2 The Tags

We asked our annotators to evaluate the stylistic and linguistic merits of each excerpt by selecting

dedicated tags. We started out with a set of 13 tags that we came up with ourselves, based on our own personal perception of what common issues are found in scientific articles, as well as on the literature on the topic (Pinker, 2014),(Ventola and Anna Mauranen, 1996)(Badley, 2019)(Crompton, 1997). The 13 initial tags are listed below; we also provide a short explanation of those tags which may not be fully transparent.

- i. Colloquial Language: to be used whenever overly colloquial language is used;
- ii. Formal Language: whenever excessively formal language is used, e.g. when expressions like *et ceteris paribus* are used (too often);
- iii. Jumbled Vocabulary: to describe combinations of words that make little sense, e.g. “the council has a *strong objective*”(objectives cannot be *strong*);
- iv. Unnecessary jargon;
- v. Verbosity;
- vi. Opaque writing: for text that is obscure, hard to understand;
- vii. Overly long sentences;
- viii. Abuse of passive sentences: e.g. "It has been found that there had been many ...";
- ix. Excessively complex syntax: e.g. “It is expected that an exploration of the variables affecting the effectiveness of reading aloud will support us in designing lessons (...)”;
- x. Clear Structure: to mark text that is clear and well-structured, text that clearly communicates the writer’s intentions, data or results;
- xi. Pretentiousness;
- xii. Engaging Writing: text that is compelling, witty and makes one want to read more;
- xiii. Dull writing: text that is dry, boring and not engaging;

When selecting which tags to include in our inventory, we tried including tags that refer to different linguistic dimensions. For example, tags 1 to 5 relate to the **lexical** dimension, tags 7 to 10 capture **syntactic** properties, tag 10 relates to **pragmatics**

and tag 11 to 13 relate to the perceived **stylistic** merits or demerits of a text. We also tried to balance the number of positive and negative tags. We provided annotators with a document explaining each tag and where it should be used, which we went over together. We then let the annotators try out the tags over the 50 sample excerpts, providing them with personalized feedback and comments should they appear to be using some of the tags incorrectly. We also told annotators that they could suggest additional tags should they notice anything that was obviously missing. After this initial dry-run over the 50 sample excerpts, based on the suggestions from the annotators we added 6 additional tags:

- xiv. Redundant (content): to be used for words, phrases or clauses that are superfluous;
- xv. Repetition (style): for anaphoric repetitions, epiphoric repetitions and anytime sentence structure or vocabulary is not diverse enough; A problem which is encountered frequently in academic writing (Xiao and Carenini, 2020)
- xvi. Poor flow: if the logical flow of a text is whacky, or whenever there are no clear threads to follow;
- xvii. Non-sequitur: sentences that do not follow logically from anything that was said before;
- xviii. Unclear/vague: for unclear referents, ambiguous statements and anything that should have been explained in more detail;
- xix. Fragment: for sentences/ paragraphs that feel excessively telegraphic in style.

Also based on the suggestions from the annotators, we replaced the tag “jumbled vocabulary” with “word choice”:

- word choice: to be used for any questionable lexical choice, *whether at the sentence level or at the level of single words.*

The final tag inventory thus consisted of 19 tags. Annotators were given the option to select tags either globally or locally. Locally selected tags referred to specific sub-parts of an excerpt, e.g. to specific words, phrases, paragraphs. An example would be the tag “overly long sentences”, that could apply to a single sentence. A tag that was selected globally meant that the specific characteristic that the tag singled out applied to the entire excerpt;



an example would be the tag “poor flow”. In the corpus, each excerpt is associated with each of the 19 tags, and for each excerpt each of the 19 tags has a value ranging from 0 to 3: 3 if that tag was selected for that excerpt by all three annotators, 0 if it was selected by no annotator. To simplify the structure of the corpus, we eliminated the distinction between global and local tags (in version 0.1 at least): if a tag was selected by an annotator, it is associated with a “1” value, regardless of whether the tag was selected globally or locally. The same holds for cases in which the same tag was selected locally more than once within the same excerpt. In future versions of the corpus, we plan on making the distinction between global and local tags accessible.

## 4 Exploratory Data Analysis

### 4.1 Tags used

Table 1 below illustrates how often the tags were selected at least once for a given excerpt (whether locally or globally) by an annotator. We see that the most frequently selected tags were “opaque writing” (34 instances), “clear structure” (36 instances) and “word choice” (18 instances).

Some of the tags which were relatively underused are “formal language” (1 instance), “colloquial language” (2 instances), “repetition” (2 instances) and “abuse of passive sentences” (3 instances). There are different reasons that could explain why these tags were underused: the low frequency of “colloquial language” could be explained by assuming that academic papers displaying an overly colloquial style are fairly rare; if anything, academic papers tend to be *too* formal. The low frequency of the “formal language” tag could be explained by citing difficulties in determining when text is *too* formal in a field where the use of formal language is generally encouraged. The same explanation could be extended to account for the low frequency of “abuse of passive sentences”: passive sentences are a feature of academic writing. Annotators might have felt compelled to accept as good passive structures that they would have flagged otherwise precisely because they were aware they were dealing with academic text.

### 4.2 Length of Drafts

Figure 1 illustrates the length distribution of each of the 100 excerpts. The average length of the excerpts was 193 words.

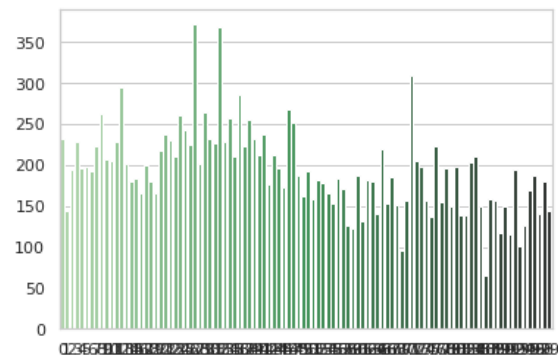


Figure 1: Length in words of each of the original 100 excerpts

Figure 2, 3 and 4 illustrate the length distribution of each of the drafts created by annotator 1, 2, and 3 respectively.

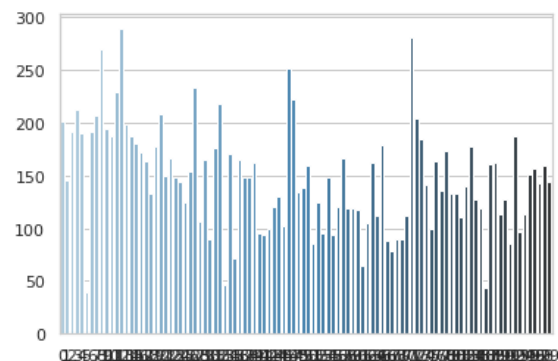


Figure 2: Length of each of the drafts created by annotator 1

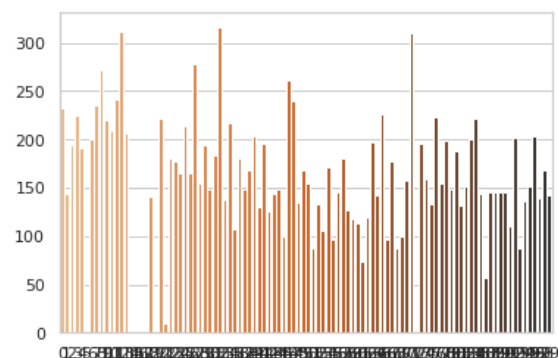


Figure 3: Length of each of the drafts created by annotator 2

Note that some of the data points are missing in figures 3 and 4 (annotators 2 and 3). This is because annotators were instructed to mark as not readable excerpts that would be too complex or time-consuming to annotate, e.g. excerpts containing lots of formulas or symbols. The missing data

colloquial language	2	abuse of passives	3	repetition	2
formal language	1	clear structure	36	fragment	9
jargon	4	pretentiousness	3	non-sequitur	4
verbosity	2	engaging	13	poor flow	8
opaque writing	34	dull	10	redundant	6
overly long sentences	4	unclear	12	complex syntax	4
word choice	18				

Table 1: Frequency of Tag Usage in corpus

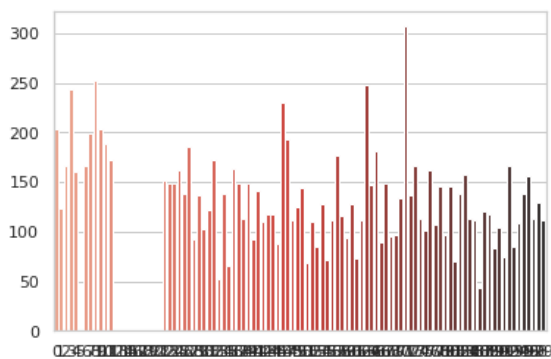


Figure 4: Length of each of the drafts created by annotator 3

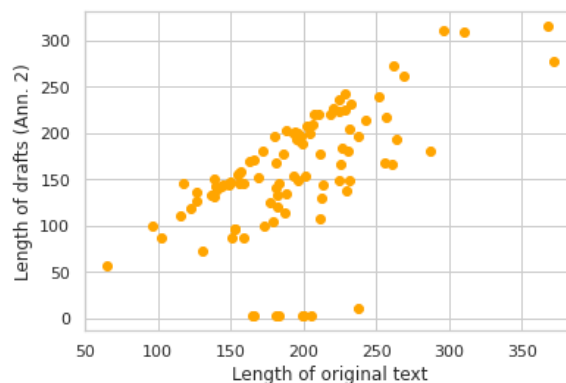


Figure 6: Ratio between length of excerpts and corresponding draft for annotator 2.

points in Fig 3-4 then represent excerpts that the annotators decided to mark as not readable.

Figures 5, 6 and 7 illustrate the ratio between length of the original excerpt and the corresponding draft for each of the 3 annotators. For annotator 1, the average ratio corresponds to 0.774; for annotator 2, to 0.813; for annotator 3, to 0.633. We see that the length of a draft increases more or less incrementally with the length of the original text for annotators 1 and 2. In the case of annotator 3, on the other hand, the length of the initial outline is less reliable of an indicator of the length of the corresponding draft.

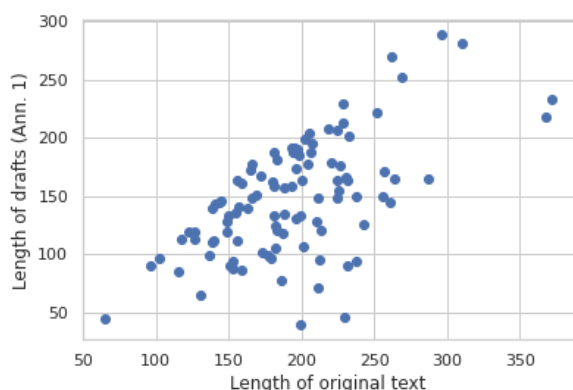


Figure 5: Ratio between length of excerpts and corresponding draft for annotator 1.

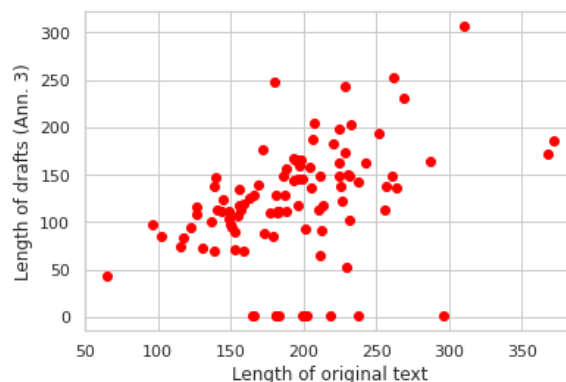


Figure 7: Ratio between length of excerpts and corresponding draft for annotator 3.

The average amount of words per draft was 146.5 for annotator 1, 155 for annotator 2 and 119 for annotator 3. Figures 8, 9 and 10 (teal scatter plots) help us further qualify these numbers by showing us how draft length compares among annotators. Figure 8 illustrates how the length of the drafts created by annotator 1 compares to those created by annotator 2. Figure 9 compares the drafts written by annotator 1 to those written by annotator 3. Finally, figure 8 compares annotator 1 with annotator 2. We see that there is indeed a difference in style between annotator 1 and 2 (some of the

drafts created by annotator 2 are longer than those created by annotator 1), and in between annotator 2 and 3 (annotator 3 writes shorter drafts). The difference between annotator 1 and annotator 3, on the other hand, seems to also be an artefact the missing data points (i.e. the original texts that the annotator decided not to annotate) for annotator 3.

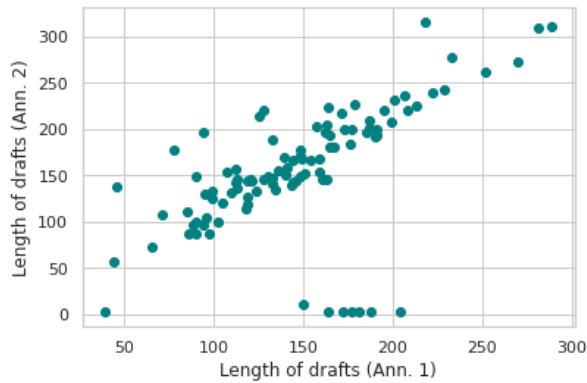


Figure 8: Ratio between length of drafts by annotator 1 and drafts by annotator 2.

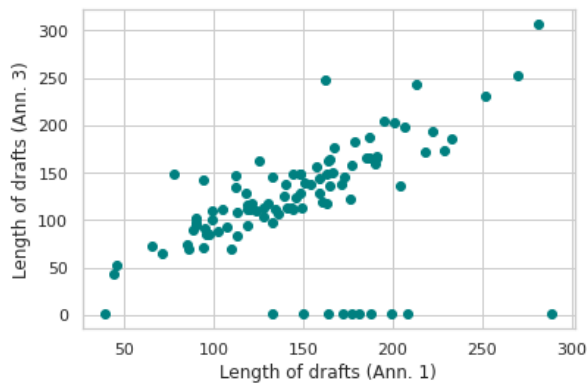


Figure 9: Ratio between length of drafts by annotator 1 and drafts by annotator 3.

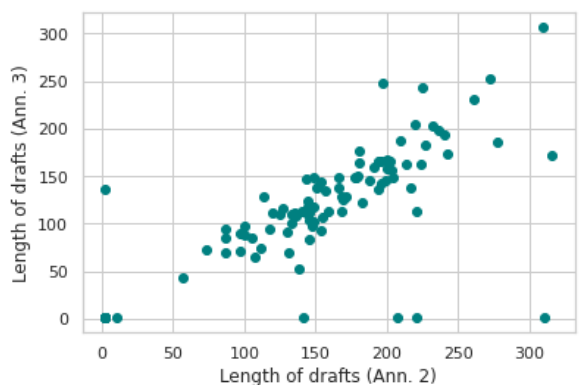


Figure 10: Ratio between length of drafts by annotator 2 and drafts by annotator 3.

We also computed the total number of words versus the number of unique words for the original excerpts, the drafts by annotator 1, those by annotator 2 and those by annotator 3 (table 2). Note that we only extracted expressions containing letter characters (symbols and digits were excluded) and with a length between 2 to 20 characters; this was mainly done to exclude formulas and mathematical symbols from the analysis. We see that the percentage of unique words over the total word count is remarkably similar overall: it is identical in both the original texts, the drafts created by annotator 1 and the drafts created by annotator 2. Annotator 3, on the other hand, appears to make a higher use of unique words. This is likely connected to the fact that annotator 3 is both a university lecturer and a writer.

## 5 Conclusions & Limitations

In this paper, we have illustrated the structure and creation process behind the SageWrite corpus, a manually annotated corpus created to support automatic language generation and automatic quality assessment of academic articles. Version 0.1 of the corpus contains annotations for 100 excerpts taken from various academic articles; each excerpt was annotated by three different annotators, all of whom were native English speakers. For each of these excerpts, the corpus contains (i) a draft version of the excerpt (ii) a selection of tags that reflect the stylistic and linguistics merits of the excerpt. Regarding drafts, on average drafts were around 26% shorter than the corresponding original text, although there was definitely variation among different annotators. More specifically, we saw that the ratio between length of the original excerpt and the corresponding draft was 0.77 for annotator 1, 0.81 for annotator 2 and 0.63 for annotator 3. This suggests that one should aim for drafts to be around  $26 \pm 9.6\%$  shorter than the original text; this value is particularly interesting in the context of automatically generating draft-like text from finished papers by selectively removing specific words and phrases, which is something we are also currently working on. Our data also suggests that 23.6% is a good value to aim for when it comes to lexical diversity in the drafts: this is the value obtained by computing the mean value of the lexical diversity indexes for annotators 1, 2 and 3 (22%, 22% and 27% respectively). Of interest is the fact that 22% was also the lexical diversity index of the original texts. An



	Total Words	Unique Words	% of unique words
Original texts	14187	3150	22%
Drafts by Annotator 1	12069	2753	22%
Drafts by Annotator 2	13320	2998	22%
Drafts by Annotator 3	9986	2774	27%

Table 2: Total words vs. Unique words

issue that reduces the power of our analysis is the missing data points for annotators 2 and 3: these are excerpts that the annotators decided not to annotate because they were deemed sub-optimal examples of text. This generally happened when the original excerpts contained a lot of mathematical formulas or other types of symbols. Annotators clearly disagreed on what was deemed "annotation-worthy": annotator 1 annotated all examples, while annotators 2 and 3 did not. Future rounds of annotations could be made more efficient by analyzing more in detail what kind of excerpts did not get annotated by the most stringent annotator (annotator 3 in our case), and then adjusting our extraction code to automatically exclude text that contains whatever features are common to those sub-optimal excerpts (e.g. a ratio of symbols and formulas higher than a certain value). Regarding tag usage, we saw that the most frequently selected tags were "opaque writing" (34 instances), "clear structure" (36 instances) and "word choice" (18 instances). The tags that were selected the least, on the other hand, were "formal language" (1 instance), "colloquial language" (2 instances), "repetition" (2 instances) and "abuse of passive sentences" (3 instances). To optimize future round of annotations, a possibility might be that of dropping these four tags from the list of tags annotators can choose from; this would reduce the total number of selectable tags to 15, something that would likely also simplify the annotation process.

## References

Graham Francis Badley. 2019. Post-academic writing: Human writing for human readers. *Qualitative Inquiry*, 25(5):180–191.

Neil Evan Jon Anthony Bowen and Nathan Thomas. 2020. Manipulating texture and cohesion in academic writing: A keystroke logging study. *Journal of Second Language Writing*, 50:100773.

Neil Evan Jon Anthony Bowen and Luuk Van Waes. 2020. Exploring revisions in academic text: Closing the gap between process and product approaches in

digital writing. *Written Communication*, 37(3):322–364.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).

Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017. [A supervised approach to extractive summarization of scientific papers](#).

Peter Crompton. 1997. Hedging in academic writing: Some theoretical problems. *English for specific purposes*, 16(4):271–287.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. [Pre-trained models: Past, present and future](#).

Gavin Francis Keay, Nicola and Karen Hind. 2020. Bone health risk assessment in a clinical setting: an evaluation of a new screening tool for active populations. *medRxiv*.

Mariëlle Leijten and Luuk Van Waes. 2013. [Keystroke logging in writing research using inputlog to analyze and visualize writing processes](#). *Written Communication*, 30(3):358–392.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. [Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing](#).

Hemant Palivela. 2021. Optimization of paraphrase generation and identification using language models in natural language processing. *International Journal of Information Management Data Insights*, 1(1):100025.

Steven Pinker. 2014. Why academics stink at writing. *The chronicle of higher education*, 61(5).

et al. Qu, Yuanbin. 2020. A text generation and prediction system: pre-training on new corpora using bert and gpt-2. 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC).

Eija Ventola and eds Anna Mauranen. 1996. Academic writing: Intercultural and textual issues. 41(2).

Wen Xiao and Giuseppe Carenini. 2020. [Systematically exploring redundancy reduction in summarizing long documents.](#)

# Razmecheno: Named Entity Recognition from Digital Archive of Diaries “Prozhito”

Timofey Atnashev<sup>♡\*</sup>, Veronika Ganeeva<sup>♡\*</sup>, Roman Kazakov<sup>♡\*</sup>  
Daria Matyash<sup>♡\$\*</sup>, Michael Sonkin<sup>♡\*</sup>, Ekaterina Voloshina<sup>♡‡\*</sup>  
Oleg Serikov<sup>♡◇‡‡</sup>, Ekaterina Artemova<sup>♡♣</sup>

<sup>♡</sup> HSE University   <sup>◇</sup> DeepPavlov lab, MIPT   <sup>‡</sup> AIRI

<sup>‡‡</sup> The Institute of Linguistics RAS   <sup>♣</sup> Lomonosov Moscow State University   <sup>\$</sup> Sber AI Centre

oserikov@hse.ru

## Abstract

The vast majority of existing datasets for Named Entity Recognition (NER) are built primarily on news, research papers and Wikipedia with a few exceptions, created from historical and literary texts. What is more, English is the main source for data for further labelling. This paper aims to fill in multiple gaps by creating a novel dataset “Razmecheno”, gathered from the diary texts of the project “Prozhito” in Russian. Our dataset is of interest for multiple research lines: literary studies of diary texts, transfer learning from other domains, low-resource or cross-lingual named entity recognition.

Razmecheno comprises 1331 sentences and 14119 tokens, sampled from diaries, written during the Perestroika. The annotation schema consists of five commonly used entity tags: person, characteristics, location, organisation, and facility. The labelling is carried out on the crowdsourcing platform Yandex.Toloka in two stages. First, workers selected sentences, which contain an entity of particular type. Second, they marked up entity spans. As a result 1113 entities were obtained. Empirical evaluation of Razmecheno is carried out with off-the-shelf NER tools and by fine-tuning pre-trained contextualized encoders. We release the annotated dataset for open access.

**Keywords:** named entity recognition, text annotation, datasets

## 1 Introduction

Modern Named Entity Recognition (NER) systems are typically evaluated on datasets such as ACE, OntoNotes and CoNLL 2003, collected from news or Wikipedia. Other common setups to test NER systems include cross-lingual evaluation (Liang et al., 2020) and evaluation in domains, other than

general, such as biomedical domain (Weber et al., 2020; Wang et al., 2019).

Additionally, the vast majority of NER datasets are in English. A few large-scale datasets for other languages are NoSta-D (Benikova et al., 2014) (German), NorNE (Jørgensen et al., 2020) (Norwegian), AQMAR (Mohit et al., 2012) (Arabic), OntoNotes (Hovy et al., 2006) (Arabic, Chinese), FactRuEval (Starostin et al., 2016) (Russian).

We present in this work a new annotated dataset for named entity recognition from diaries, written in Russian, – “Razmecheno”<sup>1</sup>. The texts are provided by the project “Prozhito”<sup>2</sup> which digitizes and publishes personal diaries. Diaries exhibit different surface and style features, such as complex narrative structure, and author-centricity, mostly expressed in simple sentences with predominance of verbs and noun phrases. NER annotation is the first step for summarisation and coreference resolution tasks.

Design choices, made for the corpus construction, are the following. We follow the standard guidelines of named entity annotation and adopt four commonly-used types Person (PER), Location (LOC), Organization (ORG), Facility (FAC). We add one more type, CHAR, which is used for personal characteristic (e.g., nationality, social group, occupation). Texts, used in the corpus, are sampled from the diaries, written in the late 1980s, the time period addressed as Perestroika. We utilized crowdsourcing to label texts.

Our dataset enables assessing performance of the NER models in a new domain or in a cross-domain transferring. We make the following contributions:

<sup>1</sup>“Got annotated”. The short form of the past participle neuter singular of the verb *размечать* (“to annotate”). <https://github.com/hse-cl-masterskaya-prozhito/main>

<sup>2</sup>“Got lived”. The short form of the past participle neuter singular of the verb *прожить* (“to live”). <https://prozhito.org/>

These authors contributed equally to this work.

1. We present a new dataset for Named Entity Recognition of 14119 tokens from 124 diaries from Prozhito. Entity types, used in the dataset, follow standard guidelines. The dataset will be freely available for download under a Creative Commons Share-Alike 4.0 license at <https://github.com/hse-cl-masterskaya-prozhito/main>;
2. We assess the performance of the off-the-shelf NER taggers and fine-tuned BERT-based model on this data.

## 2 Related work

Most of the standard datasets for named entity recognition, as ACE (Walker et al., 2005) and CoNLL (Sang and De Meulder, 2003), consist of general domain news texts in English. For our study, there are two related research lines: NER for the Russian language and NER in Digital Humanities domain.

### 2.1 NER for Russian language

The largest dataset for Russian was introduced by (Loukachevitch et al., 2021). In NEREL, entities of types PER, ORG, LOC, FAC, GPE (Geopolitical entity), and FAMILY were annotated, and the total number of entities accounts to 56K.

(Starostin et al., 2016) presented FactRuEval for NER competition. The dataset included news and analytical texts, and the annotation was made manually for the following types: PER, ORG and LOC. As of now, it is one of the largest datasets for NER in Russian as it includes 4907 sentences and 7630 entities.

Several other datasets for Russian NER, such as *Named Entities 5*, WikiNER, are included into project Corus<sup>3</sup>. Its annotation schema consists of 4 types: PER, LOC, GEOLIT (geopolitical entity), and MEDIA (source of information). Another golden dataset for Russian was collected by (Gareev et al., 2013). The dataset of 250 sentences was annotated for PER and ORG. For the BSNLP-2019 shared task, a manually annotated dataset of 450 sentences was introduced (Piskorski et al., 2019). The annotation includes PER, ORG, LOC, PRO (products), and EVT (events). RuREBus (Ivanin et al., 2020) is an example of NER dataset for a specific domain: it was introduced for a shared task in relation extraction

for business. Business-related documents were annotated manually with the help of active learning algorithm.

Several silver datasets exist for Russian NER. WikiNEuRal (Tedeschi et al., 2021) uses multilingual knowledge base and transformer-based models to create an automatic annotation for PER, LOC, PRG, and MISC. It includes 123,000 sentences and 2,39 million tokens. In Natasha project, a silver annotation corpus for Russian Nerus<sup>4</sup> was introduced. The corpus contains news articles and is annotated with three tags: PER, LOC, and ORG. For Corus project, an automatic corpus WikiNER was created, based on Russian Wikipedia and methodology of WiNER (Ghaddar and Langlais, 2017).

### 2.2 NER applications to Digital Humanities

Bamman et al. (2019) introduced LitBank, a dataset built on literary texts. The annotation was based on ACE types of named entities, and it includes the following types: PER, ORG, FAC, LOC, GPE (geo-political entity) and VEH (Vehicle). The annotation was made by two of the authors for 100 texts. The experiments with models trained on ACE and on LitBank showed that NER models trained on the news-based datasets decrease significantly in the quality on literary texts. Brooke et al. (2016) trained unsupervised system for named entity recognition on literary texts, which bootstraps a model from term clusters. For evaluation, they annotated 1000 examples from the corpus. Compared to NER systems, the model shows better results on the literary corpus data.

Apart from English LitBank, a dataset for Chinese literary texts was created and described by Xu et al. (2017). The dataset for Chinese literature texts had both rule-based annotation and machine auxiliary tagging, hence, only examples where gold labels and predicted labels differ were annotated manually. The corpus of 726 articles were annotated by five people. Besides standard tags, as PER, LOC, and ORG, the authors used tags THING, TIME, METRIC, and ABSTRACT.

Another approach to annotation was presented by Wohlgenannt et al. (2016). The authors' purpose was to extract social networks of book characters from literary texts. To prepare an evaluation dataset, the authors used paid micro-task crowdsourcing. The crowdsourcing showed high quality results and appeared to be a suitable method for

<sup>3</sup><https://github.com/natasha/corus>

<sup>4</sup><https://github.com/natasha/nerus>

digital humanities tasks.

### 3 Dataset collection

#### 3.1 Annotation schema

Our tag set consists of five types of entities. This tag set was designed empirically for texts of diaries from common tags used in related works (Walker et al., 2005; Bamman et al., 2019).

- **PER**: names/surnames of people, famous people and characters (see Example 1);
- **CHAR**: characteristics of people, such as titles, ranks, professions, nationalities, belonging to the social group (see Example 4);
- **LOC**: locations/places, this tag includes geographical and geopolitical objects such as countries, cities, states, districts, rivers, seas, mountains, islands, roads etc. (see Example 2);
- **ORG**: official organizations, companies, associations, etc. (see Example 3);
- **FAC**: facilities that were built by people, such as schools, museums, airports, etc. (see Example 4);
- **MISC**: other miscellaneous named entities.

We introduce a novel tag CHAR for the following reasons. In diaries, people are often referred with their social status or specialty. Annotation of such mentions allows for further exploration of a social spectrum. See Appendix G.4 for the exact definition of the tag as it has been presented to the assessors. Among the annotated characteristics, plenty of emotional coloured judgements (such as “rebel”, “alcoholic”, “liar”) can be found. While this highlights the subjective nature of this class of entities, it also provides a way to consider the perception of the epoch by various social groups, which we find promising for further studies.

Unlike datasets based on news, when working with diaries, it is important to know not only a person’s name (which is sufficient for news because famous people usually get into them), but also one’s social status. The reason for this is that it gives an opportunity to make assumptions about lifestyle of this person.

These five entity types can be clearly divided into two groups: the first one, PER-CHAR, is related

to people and the second one, ORG-LOC-FAC, is related to places and institutions.

We annotated flat entities, so that the overlap between two entities is not possible. The main principle of the annotation is to mark up the longest possible span for each entity, not to divide them when not required, because our schema does not assume multi-level annotation, when one entity can include another ones. For example, a name and a surname coalesce in single PER entity, rather than being two different ones (see Example 1).

- (1) А ведь Леон просил меня  
 And really Leon asked me  
PER  
 отозваться лишь о Жаке Ланге  
 to.talk only about Jack Lang  
PER

‘And Leon asked me to talk only about Jack Lang’.

- (2) Орёл самый литературный город в  
Orel the.most literary city in  
LOC  
 России  
Russia  
LOC

‘Orel is the most literary city in Russia’.

- (3) Позвонил в “Урал”: надо все-таки  
 called in “Ural” need after.all  
ORG  
 дать им знать о моем прилете.  
 give them know about my arrival

‘I called the “Ural”: after all, I have to let them know about my arrival’.

- (4) Солдаты живут в вагоне на этой  
soldiers live in car on this  
CHAR  
 станции.  
station  
FAC

‘Soldiers live in a car at this station’.

In ambiguous cases entity tags were identified based on the context, so the same entity in different



sentences could be tagged as two different types, for instance, *university* could be annotated as ORG or FAC. If an entity was used in a metaphorical sense, it would not be annotated with any tag.

- (5) Будет и на нашей улице праздник  
will and on our street a.festival  
‘Every dog has its own day’.

### 3.2 Preliminary markup

We performed preliminary analysis of the random subsets of the “Prozhito” corpus. The analysis revealed that most of the sentences contain no entities at all. To avoid costly looping over all sentences, we developed a two-stage annotation pipeline. The first stage aims at selecting sentence candidates, which may include entities of interest. This helps to reduce the amount of sentences sent to assessors and exclude sentences with no entities at all. During the second stage, entity spans are labeled in the pre-selected candidates from the first stage.

Two classifiers were trained on a small manually annotated training set — for PER-CHAR and ORG-LOC-FAC groups, respectively. The task of these classifiers is to predict, whether an entity from a group is present in a sentence, or not. These classifiers do not aim at entity recognition, but rather at binary entity detection.

We leverage upon four possible base models as classifiers: ruBERT-tiny<sup>5</sup>, ruBERT<sup>6</sup> (Kuratov and Arkhipov, 2019), ruRoBERTa<sup>7</sup>, XLM-RoBERTa<sup>8</sup>. Table 1 presents with the classification scores. A few marked up sentences (198) were taken as test sample.

Models	Precision	Recall	Micro f1-score
ruBERT-tiny	0.81	0.88	0.84
ruBERT	0.89	0.91	<b>0.90</b>
ruRoBERTa	<b>0.90</b>	0.88	0.89
XLM-RoBERTa	0.80	<b>0.99</b>	0.89

Table 1: Transformer-based binary classifiers scores

As a result, ruRoBERTa was chosen as the base model. In this task, the precision is more impor-

<sup>5</sup><https://huggingface.co/cointegrated/rubert-tiny>

<sup>6</sup><https://huggingface.co/DeepPavlov/rubert-base-cased>

<sup>7</sup><https://huggingface.co/sberbank-ai/ruRoberta-large>

<sup>8</sup><https://huggingface.co/xlm-roberta-base>

tant than the recall, since we mark up only part of the corpus and, therefore, we still miss some information, but at the same time we want to have any entities in the selected sentences with a high probability.

To train both classifiers, a random sample of size 1500 was taken from diaries belonging to the Perestroika period. Texts were independently marked up by assessors for the presence of ORG-LOC-FAC and PER-CHAR. Due to the fact that it was important to achieve a balance of classes in the training sample, and there were more texts with PER-CHAR than ORG-LOC-FAC, the training samples for ORG-LOC-FAC and PER-CHAR turned out to be different – 829 and 1465 records accordingly (see Table 2 for the validation set scores).

All available sentences were marked up by binary classifier and after that were chosen sentences with following conditions:

1. In the sentence there are entities from PER-CHAR and ORG-LOC-FAC groups, respectively;
2. Classifier was the most confident on these sentences.

Entity Type	Precision	Recall	F1-score
ORG-LOC-FAC	0.94	0.92	0.94
PER-CHAR	0.89	0.81	0.82

Table 2: ruRoBERTa scores in the binary classification task

Most confidence here means the average probabilities of each entity groups. Finally, the sentences selected this way were given to the assessors for further marking.

### 3.3 Crowdsourcing annotation

**Annotation setup** For annotation, we used Russian crowdsourcing platform Yandex.Toloka<sup>9</sup>. We prepared two tasks for assessors: determination of PER-CHAR and of ORG-LOC-FAC in “Prozhito” texts. The task was made available only to Russian native speakers. Before annotation, it is necessary to get through the learning pool with hints (20 sentences) and an exam (10 sentences) that show whether assessors understand the meaning of the

<sup>9</sup><https://toloka.yandex.ru/>

given NE tags. The sentences were tokenized with Razdel tokenizer<sup>10</sup>.

The tasks for learning, exam and control were initially annotated by the co-authors with help of annotation tool BRAT<sup>11</sup>.

Each assessor, who succeeded in the learning and exam phases, (mark  $\geq 50\%$  for learning and  $\geq 80\%$  for exam), got access to assessment of sentences in the main pool. Our main pools in both tasks consist of approximately 1500 tasks and 400 control sentences. Tasks were given to assessors on pages, Figure 3 depicts the task interface. Each page consisted of 4 normal tasks and 1 control task. A fee for one page was 0.05\$. The average time of completion of a page was about one minute. Overall, the fee per hour exceeded minimum wage in Russia. The overlap for each sentence given in Toloka is 3 in order to choose the most popular variant of markup as a correct one. Control tasks are necessary for monitoring of an annotation quality. We banned users if they skipped more than 7 task suites in a row or if they had less than 30% correct control responses.

**Assessors agreement analysis** While in most of the cases assessors had no dispute, voting mechanism has been involved in nearly one third of cases provided in the corpus (38% in the ORG-LOC-FAC task, 36% in PER-CHAR tasks, respectively).

In both tasks, the typical assessors' disagreement pattern was two competing annotation hypotheses. In the ORG-LOC-FAC task, that was mostly caused by different labels plausible for certain rare events. The ability to correctly disambiguate such terms relied on rather rare factual knowledge, thus provoking annotation errors (as in *Сижу в гостинице "Одесса"*. ('Staying in the hotel "Odessa"'), the challenging choice is 'hotel "Odessa"' is a FAC or an ORG entity). While the same group of assessors disagreements was found in the PER-CHAR task, there also emerged two more disagreements patterns: (i) identifying the proper span for the characteristics (annotating the whole *полковник в отставке* ('the retired colonel') or only *полковник* ('colonel')) and (ii) inaccurate boundaries' detection for persons initials, which mostly emerged when the assessors missed to highlight the dot in the name shortenings (as with *М. С.* in *М. С. его очень ценил поначалу*. ('M.S. valued him a lot in the beginning')).

<sup>10</sup><https://github.com/natasha/razdel>

<sup>11</sup><https://brat.nlplab.org/>

Rare cases with more than two competing annotations were mostly of random nature (as with birds being annotated as PER), or caused by the appearance of rare words (as with calzones being annotated as Person).

### 3.4 Dataset statistics

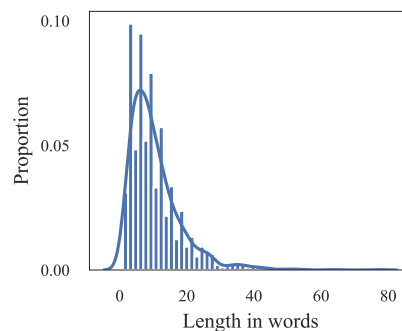


Figure 1: Distribution of sentence lengths

The total number of sentences in the dataset is 1331 and the total number of tokens is 14119. The average sentence length is 10.61 tokens (see Figure 1). 1113 entities were identified at all (1474 mentions). The average length of entity in tokens is 1.32 token.

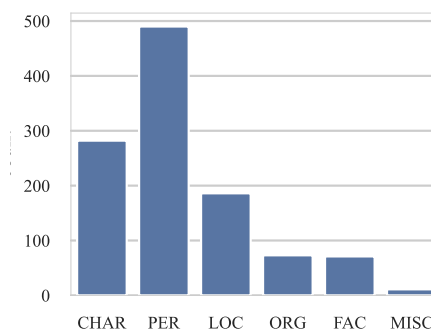


Figure 2: Distribution of entity types

Table 3 and Figure 2 describe dataset statistics.

Type	# Entities	% Entities	# Mentions	% Mentions
CHAR	282	25.0%	290	19.7%
FAC	71	6.4%	106	7.2%
LOC	186	16.7%	221	15.0%
ORG	73	6.6%	137	9.3%
PER	490	44.0%	708	48.0%
MISC	11	1.0%	12	0.8%
<b>Total</b>	<b>1113</b>	<b>100.0%</b>	<b>1474</b>	<b>100.0%</b>

Table 3: Dataset entities statistics

Entity Type	Top-10 mentions
CHAR	<i>ребёнок</i> ('child'), <i>женщина</i> ('woman'), <i>президент</i> ('president'), <i>друг</i> ('friend'), <i>поэт</i> ('poet'), <i>папа</i> ('dad'), <i>писатель</i> ('writer'), <i>жена</i> ('wife'), <i>отец</i> ('father'), <i>военный</i> ('military')
FAC	<i>театр</i> ('theatre'), <i>аэропорт</i> ('airport'), <i>дом</i> ('house'), <i>школа</i> ('school'), <i>музей</i> ('museum'), <i>кафе</i> ('cafe'), <i>станция</i> ('station'), <i>библиотека</i> ('library'), <i>посольство</i> ('embassy'), <i>тюрьма</i> ('prison')
LOC	<i>город</i> ('city'), <i>Москва</i> ('Moscow'), <i>Россия</i> ('Russia'), <i>улица</i> ('street'), <i>Ленинград</i> ('Leningrad'), <i>проспект</i> ('avenue'), <i>Кандагар</i> ('Kandagar'), <i>озеро</i> ('lake'), <i>страна</i> ('country'), <i>запад</i> ('west')
ORG	<i>ЦК</i> ('Central Committee'), <i>совет</i> ('council'), <i>парламент</i> ('parliament'), <i>Политбюро</i> ('Politburo'), <i>Правда</i> ('Pravda'), <i>КПСС</i> ('the Communist Party of the Soviet Union'), <i>издательство</i> ('publishing house'), <i>верховный</i> ('supreme'), <i>Мосфильм</i> ('Mosfilm'), <i>союз</i> ('union')
PER	<i>Горбачев</i> ('Gorbachev'), <i>Борис</i> ('Boris'), <i>Ельцин</i> ('Yeltsin'), <i>Володя</i> ('Volodya'), <i>Таня</i> ('Tanya'), <i>Витя</i> ('Vitya'), <i>Рыжков</i> ('Ryzhkov'), <i>Яковлев</i> ('Yakovlev'), <i>Сергей</i> ('Sergey'), <i>Иван</i> ('Ivan')

Table 4: Top-10 mentions for each entity type

PER is the most frequent tag, a little less than a half of all entities are of this type. Persons are often provided via a few tokens. The rest of types does not represent the same variance between mentions and entities. MISC entities are only 1% of all entities.

As expected, popular mentions of entities actually represent concepts and personalities of the Perestroika period (see Table 4). As we can see, there are main politic figures in the list (e.g., Boris Yeltsin, Mikhail Gorbachev, Nikolai Ryzhkov) as well as old soviet political authorities (e.g., Central Committee, the Communist Party of the Soviet Union, Politburo). Some words that were new at that time, such as 'a president' (since Gorbachev became the first president of USSR in 1990) or 'parliament' (the Parliament of USSR was founded in 1989) are among the most frequent words. The mixture of old Soviet terms and new words illustrates this period as a time of transition. Another important trend is the discussion of the Soviet-Afghan war, as Kardagan was one of the centres of soviet troops' dislocation.

Top-10 entities of each type in all diaries for Perestroika period can be found in Appendix H. Texts were marked up by the ruBERT model, trained on texts annotated by assessors.

## 4 Evaluation

We've benchmarked two groups of models on the presented dataset. Off-the-shelf tools were evaluated without any modifications, while transformer-based models were evaluated after a fine-tuning.

### 4.1 Off-the-shelf tools

We use a selection of publicly available, NER systems: DeepPavlov-NER, Natasha-SlovNet, Stanza, and SpaCy.

DeepPavlov-NER is a BERT-based model for NER<sup>12</sup> implemented in DeepPavlov library (Burtsev et al., 2018). Its markup includes 18 tags, including PERSON, ORGANIZATION, FACILITY, and LOCATION.

SlovNet is a neural network based tool for NLP tasks, including NER annotation. SlovNet is a part of Natasha project.<sup>13</sup> SlovNet's annotation includes PER, LOC and ORG.

Stanza is a Stanford state-of-art model<sup>14</sup>. Stanza is based on Bi-LSTM model and CRF-decoder. Stanza for Russian is a 4-entity system, which includes PER, LOC, ORG and MISC.

NER system developed by SpaCy is a transition-based named entity recognition component. We use Natasha-SpaCy<sup>15</sup> model trained on two resources - Nerus<sup>16</sup> and Navec<sup>17</sup>. Natasha-SpaCy model can detect PER, LOC and ORG entities in our dataset.

We have compared results of these models on our dataset.

<sup>12</sup><http://docs.deeppavlov.ai/en/master/features/models/ner.html>

<sup>13</sup><https://github.com/natasha/slovnet#ner>

<sup>14</sup><https://stanfordnlp.github.io/stanza/>

<sup>15</sup><https://github.com/natasha/natasha-spacy>

<sup>16</sup><https://github.com/natasha/nerus>

<sup>17</sup><https://github.com/natasha/navec>



Models	PER	LOC	ORG	Overall
DeepPavlov	0.55	0.0	<b>0.33</b>	0.93
SpaCy	0.64	<b>0.54</b>	0.16	0.95
Stanza	0.69	0.4	0.11	0.94
Natasha	<b>0.77</b>	<b>0.54</b>	0.14	<b>0.96</b>

Table 5: The performance of off-the-shelf tools (accuracy)

As seen from the table 5, *Natasha-SlovNet* showed the best performance on our dataset for PER and LOC, while *SpaCy* was the best on LOC and *DeepPavlov* showed the best results on ORG detection. However, the results of all models are significantly worse than the results on other datasets (Appendix A). Such results prove our hypothesis that off-the-shelf tools do not recognise entities on a diary-based dataset, for they were trained on news data.

Model performance analysis (Figure 5) reveals main entity recognition issues. Most of the models often detect false LOC and PER entities. In this case, *SpaCy* shows the best results. *Natasha-SlovNet* has the greatest recall, especially on LOC and PER. All models often annotated ORG as a non-entity. As our texts come from diaries written in the 1990s, some organisations could not exist anymore, and models do not recognise them.

FAC and CHAR were not on the entity lists of the models, therefore, the models did not recognise these tags. However, we would expect the models to mark CHAR as PER and FAC as LOC or ORG because those tags are related. Indeed, this happens for FAC but not for CHAR. This happens as most of the named entities are proper nouns and start with capitalized letters, unlike CHAR. All models annotated FAC more often as ORG than as non entity.

Another problem is caused by false detection of named entities’ span boundaries. To account for this, we introduced the following approach. We counted all cases when models did not find entities at all, detected false entities or used a wrong tag (combined as ‘false detected’) or models included more or less words from one or both sides. *Natasha* showed the best results, for it detects right boundaries for the most of the spans. The most common error though for all models was not finding an entity. Other mistakes include a shift of boundaries to the left and including more or less

words on the left side, especially for PER recognition. It could be possibly explained that CHAR entity proceeds PER entity (for instance, *профессор Иванов* (‘professor Ivanov’) where ‘professor’ is CHAR). Off-the-shelf models do not include CHAR entity and could annotate them as PER. Problems of narrower boundaries could be caused by excluding quoting markers in automatic annotation.

## 4.2 Fine-tuned models

We fine-tuned multiple Transformer models for NER: *ruBERT*, *ruBERT-tiny*, *ruRoBERTa*, *XLM-RoBERTa*. The performance was evaluated according to F1-scores per named entity and overall micro F1-score.

We used weighted cross-entropy as a loss function. An inverse tag frequency was taken as weights for cross-entropy, which helped us gain better results on unbalanced data. We also sorted the dataset by the length of tokens and then split it in batches, which slightly improved models’ performance. Models were trained in an unfrozen manner. The detailed hyperparameters values used to train the models are provided in the Appendix B. The performance was evaluated according to per-class and overall micro-averaged F1-score.

## 4.3 Results

*Natasha* had the best F1-score among all off-the-shelf tools. Nevertheless, results achieved for our corpus are below *Natasha*’s results on news-based datasets.

Fine-tuned transformers showed better results than off-the-shelf tools. Predictions made by *ruBERT* had the highest overall F1-score, the model’s performance had the best F1-scores for most tags (FAC, LOC, ORG) and top-3 best results for CHAR and PER tags. According to Table 6, we can consider *ruBERT* the best model for our datasets, as it successfully predicts major and minor classes.

The number of epochs was chosen according to the following criteria: the model does not overfit on the train data and shows high results on the development data. To this end, we used early-stopping. For *ruBERT-tiny* even 50 epochs were not sufficient for reaching results comparable to other models’ performances.

According to Figure 5, CHAR and PER entities were mostly wrongly detected as O by *Natasha*, *SpaCy* and *Stanza* assessors. ORG tags were also erroneously detected by these parsers, which

Models	CHAR	FAC	LOC	ORG	PER	Overall
ruBERT-tiny	0.712	0.8	0.748	<b>0.4</b>	0.738	0.731
ruBERT	0.757	<b>1.0</b>	<b>0.793</b>	<b>0.4</b>	<b>0.854</b>	<b>0.813</b>
ruRoBERTa	0.703	0.333	0.729	0.166	0.795	0.739
XLM-RoBERTa	<b>0.817</b>	0.363	0.742	0.333	0.825	0.8

Table 6: Transformer architectures F1-scores

was quite similar to the results of transformer models’ results. LOC tags almost in all cases were detected correctly both by pre-trained parsers’ transformer models, while FAC tags were significantly better found by the former ones.

According to Figure 6, XLM-RoBERTa’s performance could be considered quite successful: CHAR tags, as well as PER and LOC, were almost infallibly predicted. More exactly, PER entity was never predicted as another entity on test data. FAC entity was mixed with ORG tag in XLM-RoBERTa’s predictions, while ORG tag itself is nearly in all cases is considered as O tag by the model.

Figure 6 also presented ruBERT-tiny’s performance: CHAR and ORG entities were erroneously predicted as O more often, if compared to XLM-RoBERTa. Nevertheless, in most cases the model predicts correctly. ruBERT-tiny extracted all FAC and almost all PER tags without major errors.

As for ruBERT’s results, O tags were rarely misclassified as CHAR, while all other tags were predicted entirely correctly or with inconsequential mistakes.

ruRoBERTa’s performance was far from being perfect, as O-entities were heavily confused with other tags, but most predictions of other entities were correct.

As for major tendencies in models’ predictions, we can notice that ORG entity in most cases was detected as O tag which although was not desired, but still can encourage us to reanalyse ORG entities and collect substantially more examples of ORG tag occurrence. FAC entities were either (in most cases) correctly predicted, or mispredicted as ORG. O tags were sometimes detected as PER entity.

Given the evaluation results, one can conclude that while off-the-shelf NER tools sometimes lack desired tags, fine-tuning popular language models allows to support the chosen subset with somewhat reasonable yet far from perfect performance. This highlights the need for better few- and zero-shot sequence tagging tools capable of quickly generalizing onto novel tag-sets.

## 5 Conclusion

This paper introduces Razmecheno, a novel dataset for Named Entity Recognition. The texts in the dataset are sampled from the project “Prozhito”, which comprises personal diaries, written in Russian, from the 17th century up to the end of the 20th century. In particular, texts, marked up in Razmecheno belong to the mid-1980 years, the period in Russia, commonly known as Perestroika. Razmecheno is a middle-scale dataset so that it contains enough data to carry out literal and historical studies.

The annotation schema, used in Razmecheno, is simplistic. It consists of five named entity types, of which four are commonly used in NER datasets, namely, **persons**, **locations**, **organization**, and **facilities**. An only named entity type, introduced in this project, is **characteristics** of the different groups of people. The annotations are flat; overlapped, or nested entities are not allowed at the moment.

As our annotation schema matches a commonly used inventory of named entity types, it is possible to leverage upon pre-trained models and transfer learning techniques. The experimental evaluation of Razmecheno is two-fold. First, we carry out an extensive analysis of how available off-the-shelf NER tools cope with the task. The results reveal, that Natasha outperforms other tools under consideration by a small margin. However, of five named entity types, the off-the-shelf tools used to support only three. Next, we experiment with four state-of-the-art pre-trained Transformers. A monolingual model, ruBERT significantly outperforms other Transformers, followed by a multilingual model XLM-RoBERTa.

There are a few directions for Razmecheno development. We plan to annotate the collected sentences for other information extraction tasks, including co-reference resolution, relation extraction, and entity linking. Providing NER is the first step to present the diary’s plot in a concise form. This can be beneficial for studying the narratives and events present in diaries. This way, Razmecheno could serve as a test-bed for end-to-end information extraction models. Experiments in domain adaptation and cross-lingual transfer from other languages are another research line. Finally, we have set up the whole environment to annotate texts from “Prozhito”, so that diaries from other periods can be marked up with a little effort.

## Acknowledgments

The project is supported by the Russian Science Foundation, grant # 20-11-20166.

## 6 Bibliographical References

### References

- David Bamman, Sejal Papat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144.
- Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. Nosta-d named entity annotation for german: Guidelines and dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531.
- Julian Brooke, Adam Hammond, and Timothy Baldwin. 2016. Bootstrapped text-level named entity recognition for literature. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 344–350.
- Mikhail S Burtsev, Alexander V Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, et al. 2018. Deepavlov: Open-source library for dialogue systems. In *ACL (4)*, pages 122–127.
- Rinat Gareev, Maksim Tkachenko, Valery Solovyev, Andrey Simanovsky, and Vladimir Ivanov. 2013. Introducing baselines for russian named entity recognition. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 329–342. Springer.
- Abbas Ghaddar and Philippe Langlais. 2017. Winer: A wikipedia annotated corpus for named entity recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 413–422.
- Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.
- Vitaly Ivanin, Ekaterina Artemova, Tatiana Batura, Vladimir Ivanov, Veronika Sarkisyan, Elena Tutubalina, and Ivan Smurov. 2020. Rurebus-2020 shared task: Russian relation extraction for business. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog” [Komp’uternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog”]*, Moscow, Russia.
- Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. 2020. Norne: Annotating named entities for norwegian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4547–4556.
- Usama Khalid, Mirza Omer Beg, and Muhammad Umair Arshad. 2021. Rubert: A bilingual roman urdu bert using cross lingual transfer learning. *arXiv preprint arXiv:2102.11278*.
- Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.
- Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fengei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.
- Natalia Loukachevitch, Ekaterina Artemova, Tatiana Batura, Pavel Braslavski, Iliia Denisov, Vladimir Ivanov, Suresh Manandhar, Alexander Pugachev, and Elena Tutubalina. 2021. NEREL: A Russian dataset with nested named entities, relations and events. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 876–885, Held Online. INCOMA Ltd.
- McEnery A. et. al. 2004. The emille/ciil corpus.
- Behrang Mohit, Nathan Schneider, Rishav Bhowmick, Kemal Oflazer, and Noah A Smith. 2012. Recall-oriented learning of named entities in arabic wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 162–173.
- Jakub Piskorski, Laska Laskova, Michał Marcińczuk, Lidia Pivovarova, Pavel Přibáň, Josef Steinberger, Roman Yangarber, et al. 2019. The second cross-lingual challenge on recognition, normalization, classification, and linking of named entities across slavic languages. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. ACL.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.
- Speecon Consortium. 2011. Catalan speecon database.

- AS Starostin, VV Bocharov, SV Alexeeva, AA Bordova, AS Chuchunkov, SS Dzhumaev, IV Efimenko, DV Granovsky, VF Khoroshevsky, IV Krylova, et al. 2016. Factrueval 2016: Evaluation of named entity recognition and fact extraction systems for russian. In *Proc Dialogue, Russian International Conference on Computational Linguistics*.
- Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Ceconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2005. Ace 2005 multilingual training corpus-linguistic data consortium. URL: <https://catalog.ldc.upenn.edu/LDC2006T06>.
- Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2019. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 35(10):1745–1752.
- Leon Weber, Mario Sanger, Jannes Munchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. 2020. Hunflair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *arXiv preprint arXiv:2008.07347*.
- Gerhard Wohlgenannt, Ekaterina Chernyak, and Dmitry Ilvovsky. 2016. Extracting social networks from literary text with word embedding tools. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 18–25.
- Jingjing Xu, Ji Wen, Xu Sun, and Qi Su. 2017. A discourse-level named entity recognition and relation extraction dataset for chinese literature text. *arXiv preprint arXiv:1711.07010*.

## Appendix A Models performance on different datasets

Models	factru			ne5			bsnlp			razmecheno		
	PER	LOC	ORG	PER	LOC	ORG	PER	LOC	ORG	PER	LOC	ORG
DeepPavlov	0.91	0.886	0.742	0.942	0.919	0.881	0.866	0.767	0.624	0.55	0	0.33
SpaCy	0.901	0.886	0.765	0.967	0.928	0.918	0.919	0.823	0.693	0.64	0.54	0.16
Stanza	0.943	0.865	0.687	0.923	0.753	0.734	0.938	0.838	0.724	0.69	0.4	0.11
Natasha	<b>0.959</b>	<b>0.915</b>	<b>0.825</b>	<b>0.984</b>	<b>0.973</b>	<b>0.951</b>	<b>0.944</b>	0.834	0.718	0.77	0.54	0.14
ruBERT-tiny	0.619	0.395	0.558	0.619	0.414	0.564	0.318	0.333	0.180	0.738	0.748	<b>0.4</b>
ruBERT	0.548	0.358	0.461	0.883	0.777	0.856	0.483	0.451	0.423	<b>0.854</b>	<b>0.793</b>	<b>0.4</b>
ruRoBERTa	0.468	0.261	0.406	0.768	0.593	0.687	0.192	0	0	0.795	0.729	0.166
XLM-RoBERTa	0.879	0.763	0.78	0.963	0.936	0.944	0.762	<b>0.899</b>	<b>0.726</b>	0.825	0.742	0.333

Table 7: See Section 2.1 for the review of these corpora in the Nerus suite. The data on the performance for off-the-shelf were taken from Natasha project<sup>18</sup>

## Appendix B Transformers hyper-parameters

Models	Number of epochs	Learning rate	Weight decay
ruBERT-tiny	50	1e-5	3e-5
ruBERT	10	1e-4	2e-5
ruRoBERTa	5	1e-5	2e-5
XLM-RoBERTa	10	3e-5	1e-4

Table 8: Transformer architectures' hyperparameters

## Appendix C Crowd-sourcing task interface

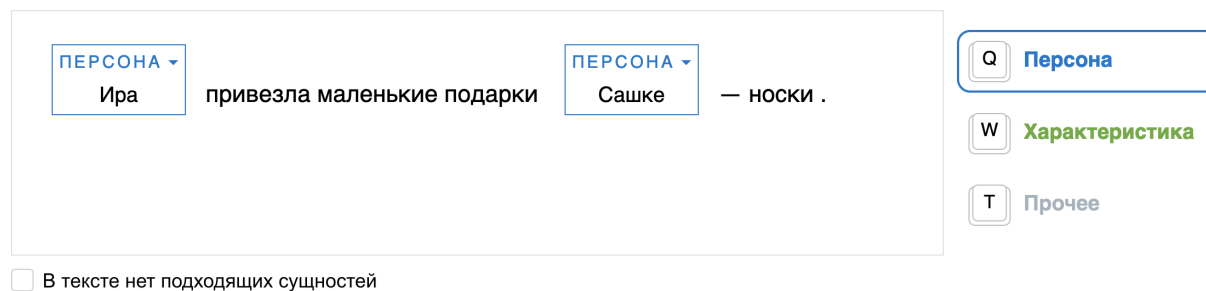


Figure 3: Annotation of a phrase given in Yandex.Toloka: *Ира привезла маленькие подарки Сашке — носки.* ('Ira brought socks as small presents for Sasha.')

Available annotations (hotkeys to annotate the selection are depicted on the right) are: *Персона* ('Person', PER, blue), *Характеристика* ('Characteristics', CHAR, green), *Прочее* ('Misc', MISC, grey), *В тексте нет подходящих сущностей* ('No entities present', checkbox).

<sup>18</sup><https://github.com/natasha/slovnet#ner>



## Appendix D Off-the-Shelf models' span recognition

To evaluate how precise off-the-shelf models are in span recognition, we divide all cases of recognition in 11 groups:

- **left more:** the right border of a span was detected correctly but on the left border a model included more words than in our annotation;
- **right more:** more words were included into a span on the right side;
- **left less:** the right border was correctly detected but on the left side one or more words were missing;
- **right less:** the left border was detected but on the right side less words were included;
- **more:** on both sides a model annotated more words than in the data;
- **less:** on the both sides a model detected a smaller span;
- **equal:** a model detected a span correctly;
- **left right:** the borders of a span were shifted from left to right, i.e., on the left side less words were included and on the right side a model detected some extra words;
- **right left:** the borders of a span were shifted from right to left;
- **not found:** models did not find a span or annotated it with a wrong tag;
- **false detected:** models found spans that were not in the manual annotation.

Figure 4 shows the absolute number of cases of each type described above.

	DeepPavlov						Natasha						
left_more	0	0	0	0	0	0	left_more	0	0.75	0	0	0	0
right_more	0	0	0	0	0	0	right_more	0	0	0	0	0	0
left_less	0	13.56	0	0	0	0	left_less	0	0	0	0	0	0
right_less	0	0.85	0	0	0	0	right_less	0	0.75	0	0	0	0
more	0	0	0	0	0	0	more	0	0	0	0	0	0
less	0	0.85	0	0	0	0	less	0	0	0	0.75	0	0
equal	0	45.76	0	0.85	0	0	equal	0	48.87	13.53	0.75	0	0
left_right	0	0	0	0	0	0	left_right	0	0	0	0.75	0	0
right_left	0	0	0	0	0	0	right_left	0	0	0	0	0	0
not_found	32.2	1.69	1.69	0.85	0	1.69	not_found	26.32	5.26	1.5	0	0	0.75
false_detected	0	0	0	0	0	0	false_detected	0	0	0	0	0	0
	CHAR	PER	LOC	ORG	LOC	FAC	CHAR	PER	LOC	ORG	LOC	FAC	

	SpaCy						Stanza						
left_more	0	1.5	0	0	0	0	left_more	0	2.33	0	0	0	0
right_more	0	0	0	0	0	0	right_more	0	0	0	0	0	0
left_less	0	4.51	0	0	0	0	left_less	0	0.78	0	0	0	0
right_less	0	1.5	0	0	0	0	right_less	0	1.55	0	0	0	0
more	0	0	0	0	0	0	more	0	0	0	0	0	0
less	0	0	0	0.75	0	0	less	0	0	0	0	0	0
equal	0	39.85	13.53	0.75	0	0	equal	0	43.41	12.4	0.78	0	0
left_right	0	0	0	0.75	0	0	left_right	0	0	0	0	0	0
right_left	0	0	0	0	0	0	right_left	0	0	0	0	0	0
not_found	27.07	7.52	1.5	0	0	0.75	not_found	27.13	6.98	3.1	0.78	0	0.78
false_detected	0	0	0	0	0	0	false_detected	0	0	0	0	0	0
	CHAR	PER	LOC	ORG	LOC	FAC	CHAR	PER	LOC	ORG	LOC	FAC	

Figure 4: Off-the-shelf tools' mistakes in span recognition for each entity

## Appendix E Off-the-shelf tools confusion matrix

		DeepPavlov				Natasha				SpaCy				Stanza						
CHAR	0	1.2	0	0	CHAR	0	1.15	0	0.05	CHAR	0	1.18	0	0.03	CHAR	0	0.03	1.13	0	0.05
FAC	0	0.08	0	0	FAC	0	0.03	0.05	0	FAC	0	0.03	0.05	0	FAC	0.05	0	0.03	0	0
LOC	0	0.44	0	0.1	LOC	0.49	0.05	0	0	LOC	0.49	0.05	0	0	LOC	0.44	0	0.1	0	0
O	0.1	92.4	0	2.61	O	0.79	92.5	0.49	1.33	O	0.79	92.4	0.44	1.46	O	1.08	1.13	91.3	0.33	1.28
ORG	0	0.1	0.03	0	ORG	0	0.08	0.05	0	ORG	0	0.08	0.05	0	ORG	0	0	0.1	0.03	0
PER	0	0.77	0	2.1	PER	0	0.2	0	2.66	PER	0.03	0.97	0	1.87	PER	0.08	0.15	0.44	0	2.2
		LOC	O	ORG	PER			LOC	O	ORG	PER			LOC	O	ORG	PER			

Figure 5: Confusion matrix for off-the-shelf tools per token in relative weights

## Appendix F Transformers confusion matrix

		ruBERT						ruBERT-tiny							
CHAR		1.1	0	0	0.13	0	0	CHAR	1.2	0	0	0.37	0	0.01	
FAC		0	0.06	0	0	0	0	FAC	0	0.07	0	0	0	0	
LOC		0	0	0.49	0.02	0	0	LOC	0	0	0.69	0.04	0	0.03	
O		0.52	0	0.24	93	0.04	0.99	O	0.52	0.01	0.39	91	0.03	1.9	
ORG		0	0	0	0.15	0.02	0	ORG	0	0	0	0.15	0.01	0	
PER		0	0	0	0.02	0	3.4	PER	0	0	0	0.09	0	3.8	
		CHAR	FAC	LOC	O	ORG	PER			CHAR	FAC	LOC	O	ORG	PER

		ruRoBERTa						XLM-R							
CHAR		1.2	0	0	0.08	0	0	CHAR	1.3	0	0	0.11	0	0	
FAC		0	0.1	0	0	0	0	FAC	0	0.048	0	0	0.048	0	
LOC		0	0	0.62	0.06	0	0	LOC	0	0	0.58	0.048	0	0	
O		0.97	0.21	0.43	90	0.21	2.2	O	0.34	0.13	0.29	91	0.032	1.5	
ORG		0	0	0	0.14	0.02	0	ORG	0	0	0	0.11	0.016	0	
PER		0	0	0	0.02	0	4	PER	0	0	0	0	0	4.1	
		CHAR	FAC	LOC	O	ORG	PER			CHAR	FAC	LOC	O	ORG	PER

Figure 6: Confusion matrix of ruBERT, ruBERT-tiny, ruRoBERTa and XLM-RoBERTa models' results on the test dataset

## Appendix G Crowd-sourcing tasks guidelines

### G.1 Binary annotation for LOC, ORG, and FAC

Please note that this task is only for Russian native speakers.

Notice if the sentence contains references to places or organizations.

Here are examples of sentences that mention places or organizations:

1. *Whatever you say, **Orel** is the most literary city in **Russia**.*
2. *A dark dream: we are going to some agricultural work along an embankment highway in a low place, a flood meadow (like the intersection of the **Kyiv** highway with **the Ugra River**)*
3. *I called "**Ural**": I had to let them know about my arrival*
4. *At eight in the morning they called us to **the headquarters** and put on the bus*
5. *A ferry on **the Danube** and **Czechoslovakia** are seen from the parapet*
6. *From the very beginning I did not like the name, but I remembered a twenty-five-year-old meeting in our **House of Culture** with a group of poets.*
7. *Soldiers live in a carriage at **this station**.*

Here are examples of sentences where there is no mention of entities:

1. *Which of the **Muscovites** is a great writer?*
2. *Unpleasant letters caught my eye in the morning.*
3. *Everything should be harmonious and beautiful.*

### G.2 Binary annotation for PER and CHAR

Please note that this task is only for Russian native speakers.

Note whether the sentence mentions people or not.

Here are examples of sentences that include mentions of people.

1. *Which of **the Muscovites** is a great writer? Well, **Pushkin**, of course.*
2. *What time did **the parents** call **the boys**?*
3. ***Asya** laughed like crazy.*
4. ***Father Alexander** came to our house from a neighboring church.*
5. ***Comrade J. V. Stalin** never trusted **that Englishman**.*
6. ***We** entered the Viennese shrine - the church of **St. Stephan** - with the flow of city guests.*

Here are examples of sentences where there is no mention of entities:

1. *In Chernobyl, we stood in line for two hours for dinner for two hours.*
2. *Unpleasant letters caught my eye in the morning.*
3. *Everything should be harmonious and beautiful.*



### G.3 Span annotation for LOC, ORG, and FAC

Please note that this task is only for Russian native speakers.

Find mentions of entities in the text and highlight them in different colors: highlight a place in **blue**, an organisation in **green** and a facility in **red**. If you can't decide on a color to mark an entity, highlight them in **gray**.

Annotation schema

- **Place** includes the names of countries, cities, states, etc. (when they designate a place), as well as natural features: mountains, bodies of water, etc.
- **Organization** is an official association, such as names of firms, companies, etc.
- **Facility** is an institution built by humans: schools, museums, offices, airports, railway stations, etc.
- **Other** is used if there is some named entity in the text (Place or Organization), but you cannot determine which one.

**Advice.** Select all the entities that you found in the text (see Example 1, there are two entities in it).

**Advice.** If several consecutive words form one entity, extend the selection to all these words (see Example 6, where the House of Culture is one entity).

Entity examples

Location: **Orel**, **Russia**, **Kyiv highway**, **Ugra river**

Organization: **“Ural”**, **headquarters**

Institution: **Lyceum 1535**, **Tretyakov Gallery**, **Kyiv Railway Station**

Markup Examples

1. Whatever you say, **Orel** is the most literary city in Russia.
2. A dark dream: we are going to some agricultural work along an embankment highway in a low place, a flood meadow (like the intersection of the **Kyiv highway** with **the Ugra River**).
3. I called **“Ural”**: I had to let them know about my arrival.
4. At eight in the morning they called us to **the headquarters** and put on the bus.
5. A ferry on **the Danube** and **Czechoslovakia** are seen from the parapet.
6. From the very beginning I did not like the name, but I remembered a twenty-five-year-old meeting in our **House of Culture** with a group of poets.
7. Soldiers live in a carriage at **this station**.

### G.4 Span annotation for PER and CHAR

Please note that this task is only for Russian native speakers.

Mark references to people in the text and highlight it in different colors: highlight a person in **blue** and a characteristic in **green**. If you can't decide on a color to tag a person, highlight them in **gray**.

Annotation schema

- **Person** is a name (as well as a surname, pseudonym, etc.) of a person or group of people, including fake and famous ones.
- **Characteristic** is a characteristic of a person (**rank, profession, nationality, belonging to a social group**)

- Other is used if there is some named entity (Person or Characteristic) in the text, but you cannot determine which one.

**Advice.** Select all the entities that you found in the text (see Example 4, there are two entities in it).

**Advice.** If several consecutive words form one entity, extend the selection to all these words (see Example 5, where J. V. Stalin is one entity).

#### Entity examples

Persons: [Asya](#), [Pushkin](#), [J. V. Stalin](#) (J.V. Stalin is one person, so you should extend one selection to all three words.)

Characteristics: [schoolchildren](#), [girls](#), [women](#), [priests](#), [Americans](#)

#### Markup Examples

1. [Asya](#) laughed like crazy. (Asya is a person's name)
2. Which of [the Muscovites](#) is a great writer? Well, [Pushkin](#), of course. (Pushkin is the name of a person, Muscovite is a characteristic)
3. What time did [the parents](#) call [the boys](#)? (the parents is a characteristic, the boys is a social group)
4. [Father Alexander](#) came to our house from a neighboring church (the word father here is a profession (his characteristic), Alexander is the name of a person)
5. Comrade [J. V. Stalin](#) never trusted [that Englishman](#). (Comrade is definitely something like Characteristics, but it seems that it does not fall under the description of Characteristics; J.V. Stalin is the name of a person; Englishman is a nationality))

## Appendix H Top-10 entities of each type in the Prozhito diaries

Entity Type	Top-10 mentions
CHAR	<i>ребёнок</i> ('child'), <i>жена</i> ('wife'), <i>секретарь</i> ('secretary'), <i>женщина</i> ('women'), <i>мама</i> ('mom'), <i>отец</i> ('father'), <i>командир</i> ('commander'), <i>писатель</i> ('writer'), <i>президент</i> ('president'), <i>начальник</i> ('chief')
FAC	<i>театр</i> ('theatre'), <i>музей</i> ('museum'), <i>школа</i> ('school'), <i>институт</i> ('institute'), <i>церковь</i> ('church'), <i>университет</i> ('university'), <i>училище</i> ('college'), <i>госпиталь</i> ('hospital'), <i>кафе</i> ('cafe'), <i>монастырь</i> ('monastery')
LOC	<i>Москва</i> ('Moscow'), <i>Россия</i> ('Russia'), <i>Ленинград</i> ('Leningrad'), <i>Кандагар</i> ('Kandagar'), <i>город</i> ('city'), <i>Кабул</i> ('Kabul'), <i>Афганистан</i> ('Afghanistan'), <i>советский</i> ('soviet'), <i>страна</i> ('a country'), <i>СССР</i> ('USSR')
ORG	<i>ЦК</i> ('Central Committee'), <i>Политбюро</i> ('Politburo'), <i>партия</i> ('party'), <i>КПСС</i> ('the Communist Party of the Soviet Union'), <i>МИД</i> ('Foreign Ministry'), <i>КГБ</i> ('Committee for State Security'), <i>член</i> ('member'), <i>союз</i> ('union'), <i>СП</i> ('Union of writers'), <i>правительство</i> ('government')
PER	<i>Горбачев</i> ('Gorbachev'), <i>М. С.</i> ('M. S., Gorbachev's initials'), <i>Ельцин</i> ('Yeltsin'), <i>Веничек</i> ('Venichek'), <i>Любимов</i> ('Lubimov'), <i>Ерофеев</i> ('Yerofeyev'), <i>Яковлев</i> ('Yakovlev'), <i>Сталин</i> ('Stalin'), <i>Галя</i> ('Galya'), <i>Володя</i> ('Volodya')

Table 9: Top-10 mentions for each entity type on the whole Prozhito diaries during the Perestroika period

# Filtering of Noisy Web-Crawled Parallel Corpus: the Japanese-Bulgarian Language Pair

Iglika Nikolova-Stoupak   Shuichiro Shimizu   Chenhui Chu   Sadao Kurohashi  
Kyoto University

{iglika, sshimizu, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp

## Abstract

One of the main challenges within the rapidly developing field of neural machine translation is its application to low-resource languages. Recent attempts to provide large parallel corpora in rare language pairs include the generation of web-crawled corpora, which may be vast but are, unfortunately, excessively noisy. The corpus utilised to train machine translation models in the study is CCMatrix, provided by OPUS. Firstly, the corpus is cleaned based on a number of heuristic rules. Then, parts of it are selected in three discrete ways: at random, based on the “margin distance” metric that is native to the CCMatrix dataset, and based on scores derived through the application of a state-of-the-art classifier model (Acarcicek et al., 2020) utilised in a thematic WMT shared task. The performance of the issuing models is evaluated and compared. The classifier-based model does not reach high performance as compared with its margin-based counterpart, opening a discussion of ways for further improvement. Still, BLEU scores surpass those of Acarcicek et al.’s (2020) paper by over 15 points.

**Keywords:** neural machine translation, low-resource language pairs, Bulgarian language, Japanese language, corpus filtering, web-crawled corpora

## 1 Introduction

In recent years, web-crawled corpora have come as an attempt to tackle the problem of limited parallel corpora, notably when it comes to machine translation involving low-resource language pairs. They are the product of unsupervised covering of portions of the web based on a widely used metric, such as the cosine distance between sentence embeddings, and they tend to be produced in excess, leading to problems like redundancy and data of low quality (Schafer

et al., 2014). Large web-crawled corpora are often associated with a lack of documentation and require further work before they can be used within the field of machine translation (Dodge et al., 2021).

In their study, Khayrallah and Koehn (2018) discuss the types of noise that tend to occur in web-crawled corpora, as well as their effect on potential machine translation systems. Notably, neural machine translation is affected by such noise to a considerably greater extent as compared with its statistical counterpart, derived BLEU scores decreasing dramatically at its experimental introduction (Khayrallah and Koehn, 2018).

Motivated by a desire to mitigate the described problems, associated with similarly derived parallel corpora, WMT has organised three shared tasks in 2018-2020, addressing their cleaning, the last two of which have specifically centred on low-resource language scenarios. Several excellent state-of-the-art models have been produced to handle the task. In this paper, a representative model (Acarcicek et al., 2020) is selected and applied to a particular, extremely under-resourced language pair: Japanese-Bulgarian. Acarcicek et al.’s model uses a classifier on top of RoBERTa in order to score sentence pairs according to the level of certainty that they are mutual translations.

The corpus discussed in this study is CCMatrix, the largest parallel dataset that is currently available in the addressed language pair. It is provided by the OPUS collection (Tiedemann, 2012) and contains over four million multi-domain web-crawled sentences, derived based on “margin distance.” The last is an improved implementation of cosine distance that considers the ratio of the cosine distance between two candidate sentences’ embeddings as compared with the average cosine distance that a sentence has with its nearest neighbours (Schwenk et al., 2019). Following preprocessing based on heuristic

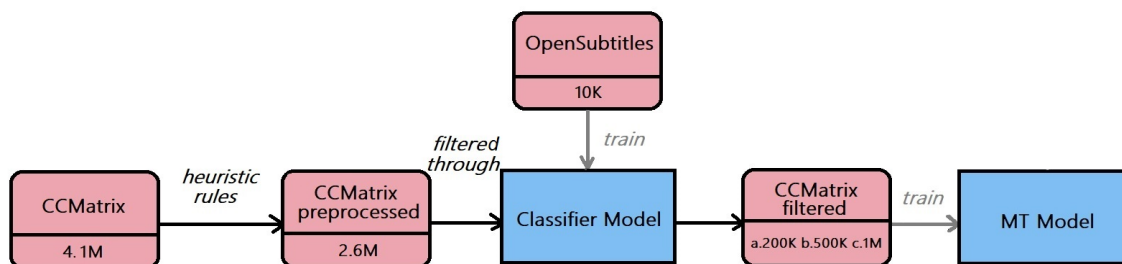


Figure 1: pipeline of the current study

rules that keep in mind the characteristics of the two languages in question, discrete subcorpora of three sizes (200K, 500K, and 1M) are selected based on margin distance and on the classifier-derived scores. They are compared to randomly selected subcorpora of the same size (see Figure 1). The margin-distance-based models show significantly improved performance. Conversely, the performance of the classifier models is largely non-optimal, showing the need for improvement of the selection techniques, such as through higher focus on the morphological and semantic specificities of the two languages. Importantly, the best derived model outperforms the one offered by Acarcicek et al. (2020) by over 15 BLEU points.

## 2 Related Work

### 2.1 WMT Corpus Filtering Shared Tasks

The particular languages addressed in this paper have not been involved in substantial research regarding the cleaning of noisy parallel corpora. This being said, the current study is highly inspired by the WMT corpus filtering shared tasks conducted in 2019 and 2020, which specifically targeted low-resource languages as an entity. Participants were prompted to provide a method of scoring the quality of each sentence within a provided noisy parallel corpus in order to then use the best scored pairs to train a translation model. In the process, they were allowed to use available clean parallel or monolingual data. The winning papers apply several distinct filtering techniques, including various uses of monolingual data, sentence embeddings, transfer learning, back translation, as well as the tool discussed in this paper, classifiers. In their highly successful model

(which was consequently taken as a baseline within the shared task), Chaudhary et al. (2019) use only parallel data as they apply LASER sentence embeddings and calculate the cosine distance between sentences in order to obtain similarity scores. Lo and Joanis (2020) in turn utilise the semantic metric Yisi-2 in their scoring method, underlining the importance of vocabulary coverage. In their SMT system, Sen et al. (2019) come up with a fuzzy matching method akin to the one to be used in this paper, via which they calculate the Levenshtein distance between the corpus’s English sentences and English translations of the additional language’s sentences.

### 2.2 Use of Classifiers

A number of successful submissions to WMT’s 2018-2020 shared tasks opt for a classifier model that differentiates between positive and negative examples of parallel sentences. In the 2018 edition of the shared task, Junczys-Dowmunt et al. (2018) assign cross-entropy scores to a noisy corpus’s sentence pairs after first generating an inverse translation model trained on clean parallel data in the languages in question. Sánchez-Cartagena (2018) makes use of a classifier composed using the free open-source tool Bicleaner and enhanced with randomised trees and heuristic rules.

In fact, the use of classifier models in machine translation far predates the mentioned shared tasks as well as current state-of-the-art tools and recently assembled corpora. Munteanu and Marcu (2005) use a classifier to improve translation memory. Tyagi et al. (2015) apply support-vector machines and a Naive Bayes classifier in the ranking of translated sentences

into several categories ranging from “excellent” to “bad.” Yogi et al. (2015) in turn rate the quality of produced machine translation with a Kneser-Ney smoothing language model that assigns probability scores to translated output. A year prior to the launching of WMT’s shared tasks, Xu and Koehn (2017) come up with the data cleaning system Zipporah, which classifies the quality of translated sentences using bag-of-words.

### 3 Noise in the CCMatrix Corpus (Japanese-Bulgarian)

Akin to an experiment in Khayrallah and Koehn’s (2018) study, a random 200-sentence sample from the described corpus is examined in an attempt to identify the nature of the different types of noise present.<sup>1</sup> The examined sentences demonstrate a large variety of domains and registers and feature a wide range of vocabulary, notably including a number of proper nouns. The main types of noise discovered include: non-corresponding numbers and dates, inappropriate punctuation, wrong use of abbreviations, presence of foreign languages, and machine-translated text.

As numbers and dates widely mismatch between the two languages within a sentence pair, they are regarded as noise. The next largest source of noise in the Bulgarian sentences comes in the face of problems with punctuation (for instance, a frequent use of “...”) and capitalisation. What follows are instances of “non-standard language,” including a large number of sentence fragments (for example, “Ако по някаква причина се преместят в друго училище,”). However, if one disregards the lack of final punctuation within these fragments, they read smoothly and match unproblematically between the two languages. In fact, the mandate for a sentence to contain a main verb, largely influenced by English grammar, is not intrinsic to either the Bulgarian or the Japanese language. While the Cambridge dictionary states that a sentence is “a group of words, *usually containing a verb*, that expresses a thought” (“Sentence”, 2022; emphasis added), Bulgarian (“Изречение”, 2022) and Japanese (“文”, 2022) counterparts do not make a reference to the concept of “verb” in their definitions of a sentence.

<sup>1</sup> See Appendix A for a detailed description of the

One Bulgarian sentence contains the word “сопи,” a slang transliteration of the English “sorry” (the respective Japanese sentence does not demonstrate any parallelism). Seemingly machine-translated sentences come at as much as closely five per cent and are therefore placed in a separate category.

An example is the sentence “Как мога да защитавам моята PC?”, which contains a gender mismatch and an unnatural English abbreviation. Other types of noise include abbreviations in both the Cyrillic and Latin alphabets, excessively large sentences and sentences written in (or partly in) a foreign language, predominantly English. Foreign language within a sentence ranges between a single word or phrase that can safely be regarded as a proper noun (e.g. “Google Assistant”) and a full sentence written in a foreign language with a few seemingly mistakenly inserted Bulgarian words.

Similar patterns are observable when it comes to the noise in Japanese sentences: the use of numbers and dates, followed by abbreviations in the Latin alphabet and wrong punctuation. An additional problem related to punctuation is the fact that it differs significantly between the two languages; as a result, for instance, a Bulgarian “...” may be rendered as either “...” or “--” in the parallel Japanese sentence. Other examples of “non-standard language” come in the face of language attributable to “texting” (e.g. a “laughter” kanji in the end of a sentence) and supplementary hiragana renditions of kanji and katakana scripts, placed in brackets.

Some of the observed types of noise can be addressed directly during the preprocessing step (see Section 4.1). Such an issue as machine-translated language, however, is difficult to tackle using heuristic rules.

## 4 Methodology

### 4.1 Preprocessing

Like the majority of submissions for WMT’s corpus filtering shared tasks, this study starts off with a preprocessing step that applies a series of heuristic rules to the noisy corpus. In concordance with observations described in Section 3, the

types of noise found.

following preprocessing pipeline is applied: N/A entries and duplicates are removed; sentences in different languages are removed; Japanese sentences are tokenised; Japanese sentences with more than two pairs of brackets are removed (as they may indicate the use of multiple scripts); punctuation is removed; capitalisation is removed from Bulgarian sentences; sentences that show a large mismatch in size are removed; dates are replaced with the tag “DATE”; and numbers are replaced with the tag “NUM.” The library *datefinder*<sup>2</sup> is utilised to locate dates written in a variety of formats. The tool used to identify sentences in languages other than the expected ones is *langdetect*<sup>3</sup>. Conveniently, in the case of short amounts of text in a foreign language, language is labelled in accordance with the large portion of text, thus allowing for sentences with words and phrases in English that take the role of proper nouns to remain in the corpus. Several patterns of wrong labelling are established and taken into consideration (e.g. Bulgarian text is occasionally mistakenly guessed to be in Russian or Macedonian).

Where applicable, the mentioned cleaning rules bear in mind the morphological and syntactic specificities of the two languages in question. For instance, the thresholds that are assumed to indicate unlikely proportions in sentence lengths are determined following observations of translation examples. Also, even though the later utilised neural models do not mandate prior tokenisation, a decision is made for Japanese text to be tokenised as part of preprocessing due to the language’s notorious lack of space delimiters between words. The tool used for tokenisation is Juman++, developed in Kyoto University (Tolmachev et al., 2018).

## 4.2 “Proxy Filter” Classifier

This study sought to apply a winning state-of-the-art model from the WMT corpus filtering shared tasks to the selected Japanese-Bulgarian corpus. Several criteria were considered within the choice of a model. Firstly, the focus was on 2019 and 2020 tasks, as they explicitly target low-resource language pairs (albeit in an English-centred setting). Simplicity, availability and

reproducibility of research were also sought, thus dismissing for instance ensemble methods. Due to a strong recent shift toward NMT, SMT models were also disregarded, and so were models that involved not only corpus cleaning but also their own alignment of candidate parallel sentences (an option introduced in 2020’s shared task). In the case of high similarity, newer models were preferred over older ones (for instance, Acarcicek et al. of 2020 was regarded as a better choice than Bernier-Colborne et al. of 2019). Finally, in the case of several experiments utilised within the same submission, only authors’ best attempts were to be made use of.

Consequently, Acarcicek et al.’s 2020 model was selected. The authors enhance a multilingual RoBERTa-Large model (Liu et al., 2019) with a “proxy filter” i.e. a classifier that is trained to differentiate between positive and negative examples of parallel sentences. Specific attention is placed on the generation of challenging negative examples. The utilised technique is “fuzzy string matching,” also known as “approximate string matching,” which applies Levenshtein distance in the calculation of levels of similarity between texts.

A notable difference between Acarcicek’s work and the one presented in this paper is the fact that the CCMatrix corpus already contains a metric pertaining to the level of parallelism of sentence pairs, the “margin distance.” As a result, the study benefits from a comparison between a use of this native unsupervised metric and the newly derived classifier scores in the later translation model.

## 5 Experiments

### 5.1 Data

The parallel corpus whose cleaning is undertaken in this study is CCMatrix by OPUS (Japanese-Bulgarian). In its original form, the corpus contains 4.1M web crawled parallel Japanese-Bulgarian sentences. Following preprocessing based on heuristic rules, the corpus contains a little over 2.5M sentences.

The test and validation sets of the translation model comprise of 1,000 clean parallel sentence pairs each. The sentences are randomly taken from the top scoring 20K sentences following the

---

<sup>2</sup> [datefinder.readthedocs.io](https://datefinder.readthedocs.io)

<sup>3</sup> <https://pypi.org/project/langdetect/>

classification task and are then removed from the training set. In order to guarantee quality and remove a bias toward sentences selected by the classification task, thorough manual editing and translation are applied.

The “proxy filter” classifier is trained on 10K parallel sentences from the OpenSubtitles (Japanese-Bulgarian) corpus. This corpus is significantly cleaner than CCMatrix, and it has notably been used by Koeva et al. (2012) in the construction of the “Bulgarian X-Language Parallel Corpus,” the largest systematized Bulgarian bilingual corpus to date. Importantly, however, the OpenSubtitles corpus is more domain-specific as compared with CCMatrix, thus encouraging the extraction of a specific subtype of sentences from the latter.

All data is preprocessed following the same general pipeline as described in 4.1.

## 5.2 Classifier Model

The hyperparameters of the classifier model to be utilised were selected via grid searching: training epochs (0, 5), learning rate (2e-6, 2e-4, 2e-2, 0.2), negative random sampling<sup>4</sup> (2, 5, 8, 10), fuzzy ratio<sup>5</sup> (2, 1, 5), fuzzy max score<sup>6</sup> (30, 60, 100) and positive oversampling<sup>7</sup> (1, 2, 10). The models were trained on a single TITAN RTX GPU.

## 5.3 Translation Models

After the CCMatrix corpus was preprocessed, subcorpora were obtained through the application of three techniques: at random, based on margin distance and based on the classifier scores. Japanese-Bulgarian Transformer neural machine translation models<sup>8</sup> were trained as per the FAIRSEQ toolkit (Ott et al., 2019). In addition, three sizes of training data were introduced in an attempt to determine the optimal level of compromise between data size and data quality: 200K, 500K and 1M parallel sentences. The transformer models were trained on 8 TITAN X

GPUS at a learning rate of 5e-4, using square root scheduler and a dropout of 0.3; early stopping was applied. The models were evaluated using BLEU scores.

## 6 Results

### 6.1 Classifier Models

Over two thirds of the derived classifier models received an F1 score of 0 while at the same time showing high accuracy scores. An F1 of 0 implies that the value of either precision or recall is 0. A plausible reason is that such a model falsely identifies all examples as negative. While overall trends are difficult to pinpoint in relation to the models with highest F1 scores, all of them are trained for two epochs at a learning rate of 2e-6. Fuzzy matching scores and fuzzy ratios vary. When it comes to negative random sampling and positive oversampling, a general tendency is discernable for high values of the latter and slightly lower ones for the former (see Table 1).

Experiments with the application of several random seeds and a different amount of parallel data showed that, whilst a different random seed does not lead to significantly lower F1 scores for the best models, a different amount and organisation of parallel sentences often does reduce the score to 0. Training loss decreases smoothly with all models, the lowest score being associated with a model whose F1 score is 0.58.

	Fuzzy Ratio	Fuzzy Max Score	Positive Over-sampling	Negative Random Sampling	F1 Score
#1	1	100	2	10	0.72
#2	5	60	10	8	0.7
#3	5	30	10	8	0.7
#4	2	30	10	8	0.7
#5	1	30	10	8	0.7

Table 1: Varying hyperparameters among the top five classifier models according to F1 score

Due to the fact that the best scoring model (F1

<sup>4</sup> the ratio of negative examples in the classifier

<sup>5</sup> the number of similar sentences taken based on a sentence’s fuzzy matching score

<sup>6</sup> a threshold (in percent) for the fuzzy matching similarity a sentence is allowed to exhibit; used in order to avoid the inclusion of duplicates or extremely similar sentences

<sup>7</sup> oversampling of the classifier’s positive examples in order to maintain a given ratio with negative examples

<sup>8</sup> 6 layers, learning rate 5e-4, dropout 0.3, early stopping, vocabulary size 8,000



score of 0.72) demonstrates a slightly irregular pattern, such as the only negative ratio of 10 among the top five models and a fuzzy max score of 100 (a value that in fact negates the parameter's influence), the second best model (F1 score of 0.7) was selected as baseline.

## 6.2 Derived Scores

Following application of the classifier to the preprocessed CCMatrix corpus, each sentence pair received a score between 0 and 1, denoting its level of parallelism. The derived scores exhibit the following characteristics: their values range between 0.028 and 0.977, and their mean comes at 0.926.

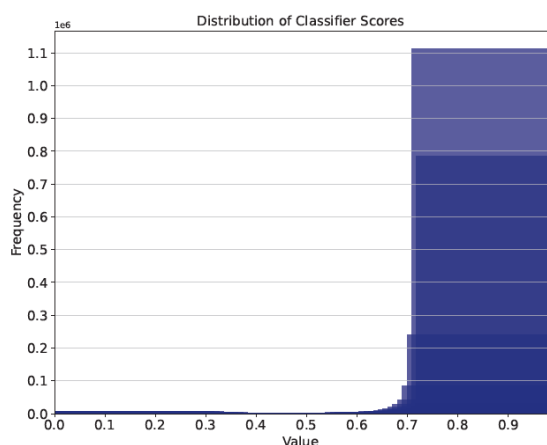


Figure 2a: Distribution of classifier scores.

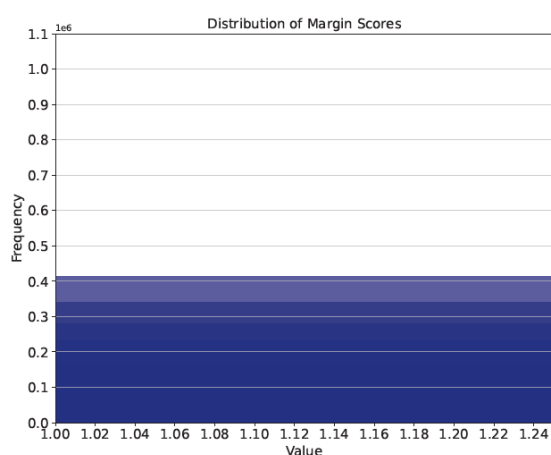


Figure 2b: Distribution of margin distance.

Figure 2 shows the distribution of classifier scores (a) as compared with the distribution of the native to CCMatrix margin scores (b). Whilst the latter demonstrates full uniformity at the given scale, the former exhibits high concentration as scores approach their maximum value.

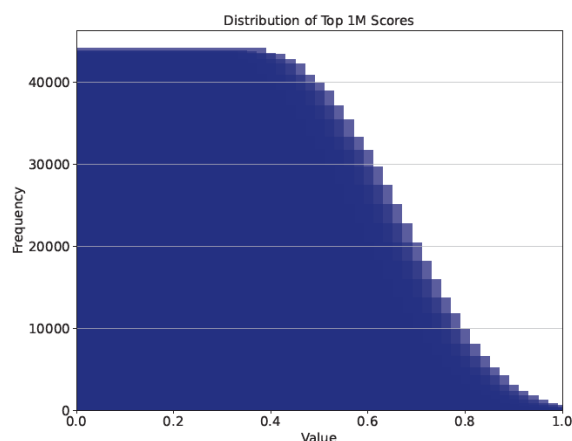


Figure 3: Distribution of the top 1M classifier scores.

In addition, Figure 3 provides a close-up overview of the distribution of the top 1M scores (that is to say, the scores corresponding to the sentences used in the study's translation model). These scores range between 0.967 and 0.977.

Manual evaluation of 20 scored sentence pairs (five with a score of over 0.9 and five with a score of under 0.9 from both the beginning and end of the corpus<sup>9</sup>) shows that classifier scores provide a discernibly better evaluation of sentence parallelism.

## 6.3 Translation Models

With a BLEU score of 28.49, the highest scoring model is the one that is trained on 1M parallel sentences and uses the CCMatrix margin distance metric (Table 2). Its classifier-based counterparts score even lower than the randomly selected sample with a BLEU score of 22.28 vs 25.25. A possible reason for better performance of margin-based and randomly selected models as compared with classifier-based ones is the variety of domains and registers that is retained from the

margin distance scores.

<sup>9</sup> The CCMatrix corpus is ordered in descending order of

original web-crawled corpus. In contrast, classifier scores, which are derived following training on a corpus of a narrower domain, encourage a focus on a specific type of sentences in addition to a higher level of cleanliness and are likely to have favored sentences “crawled” from the same or related sources.

Translation Model	Size	BLEU Score
Preprocessing + Random	200K	18.24
Preprocessing + Margin-Based	200K	19.85
Preprocessing + Classifier-Based	200K	17.02
Preprocessing + Random	500K	21.10
Preprocessing + Margin-Based	500K	23.91
Preprocessing + Classifier-Based	500K	20.52
Preprocessing + Random	1M	25.25
Preprocessing + Margin-Based	1M	<b>28.49</b>
Preprocessing + Classifier-Based	1M	22.28

Table 2: BLEU scores of the NMT models

In contrast, Acarcicek et al.’s (2020) best scoring classifier model increase the shared task’s LASER-based baseline by 1.1 and 1.3 points for the two considered language pairs. It is worth noting, however, that overall BLEU scores are significantly lower, the highest results coming at 13.3 (Acarcicek, 2020). It is possible that this difference is partly explainable through the examined languages’ characteristics combined with appropriate preprocessing.

## 7 Conclusion and Future Work

Although the exposed study exhibits high similarity to WMT’s corpus filtering shared tasks, several crucial elements that distinguish it should be made note of. Firstly, the English language is

not featured in either translation direction, and the examined language pair is not selected merely quantitatively based on its associated resources but is closely associated with the study and its goals. As a result, preprocessing is key within the filtering process. Part of the corpus’s preprocessing is language-specific, and a suggested direction for future improvement of the utilised classifier model would involve further application of the two languages’ morphological features (such as the use of an alternative, more morphologically-aware fuzzy search algorithm and the inclusion of Universal Dependencies annotations and relations).

Additionally, in this study a customised filtering model benefits from a comparison with one that uses margin scores, thus allowing for specific conclusions to be made, such as the effect of domain-specific data on the machine translation models. In order for this narrowing of the corpus to be avoided, clean multi-domain data could be attained if a manually cleaned portion of CCMatrix is used in training the classifier model.

Also, as performance increases steadily with subcorpora sizes, even larger models should be experimented with.

Importantly, the current work does not claim to propose a high quality translation system in the low-resource Japanese-Bulgarian language pair. Rather, it provides methods for improving the quality of noisy parallel sentences and for the selection of specific portions of higher-quality data. The study may be used as the starting point for further work toward an improved translation model in the described language pair as well as a general frame of reference in terms of a filtering pipeline that can be adapted to other corpora and language pairs.

## References

- Acarcicek, H. Colakoglu, T., Aktan, P. E., Huang, C., Peng, W. (2020). Filtering Noisy Parallel Corpus Using Transformers with Proxy Task Learning, *Proceedings of the Fifth Conference on Machine Translation (WMT)*.
- Bernier-Colborne, G. and Lo, C. (2019). NRC Parallel Corpus Filtering System for WMT 2019, *Proceedings of the Fourth Conference on Machine Translation (WMT)*, Florence, Italy, pp.252-260.
- “文” (2022) *Goo*. Available at:

<https://dictionary.goo.ne.jp/thsrs/10547/meaning/m0u/文/> (Accessed 14 June 2022).

Chaudhary, V. Tang, Y., Guzmán, F., Schwenk, H., Koehn, P. (2019). Low-Resource Corpus Filtering using Multilingual Sentence Embeddings, *Proceedings of the Fourth Conference on Machine Translation*.

Dodge, J. Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groenvelde, D., Mitchell, M., Gardner, M. (2021). Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. arXiv:1907.11692.

“Изречение” (2022) *OnlineRechnik.com*. Available at: [m.onlinerechnik.com/duma/Изречение](http://m.onlinerechnik.com/duma/Изречение) (Accessed 14 June 2022).

Junczys-Dowmunt, M. Grundkiewicz, R., Guha, S., Heafield, K. (2018). Approaching Neural Grammatical Error Correction as a Low-Resource Machine Translation Task, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, pp.595-606.

Khayrallah, H. and Koehn, P. (2018). On the Impact of Various Types of Noise on Neural Machine Translation, *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, Melbourne, Australia, pp.74-83.

Koeva, S. Stoyanova, I., Dekova, R., Rizov, B., Genov, A. (2012). Bulgarian X-language Parallel Corpus, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, pp.2480–2486.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). Roberta: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Lo, C. and Joanis, E. (2020). Improving Parallel Data Identification using Iteratively Refined Sentence Alignments and Bilingual Mappings of Pre-Trained Language Models, *Proceedings of the Fifth Conference on Machine Translation*, pp.972–978

Munteanu, D. S. and Marcu, D. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel Corpora, *Computational Linguistics*, 31(4):477–504.

Ott, M. Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., Auli, M. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, Minnesota, pp. 48–53.

Sánchez-Cartagena, V. M. (2018). Prompsit’s Submission to WMT 2018 Parallel Corpus Filtering Shared Task, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Belgium, Brussels, pp.955-962.

Schafer, R. et al. (2014). Focused Web Corpus Crawling, *Proceedings of the 9th Web as Corpus Workshop (WAC-9)*.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., Guzmán, F. (2021). WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp.1351-1361.

Sen, S. Ekbal, A., Bhattacharyya, P. (2019). Parallel Corpus Filtering Based on Fuzzy String Matching. *Proceedings of the Fourth Conference on Machine Translation*.

“Sentence” (2022) *Cambridge Dictionary*. Available at: <https://dictionary.cambridge.org/dictionary/english/sentence> (Accessed 14 June 2022).

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS, *Proceedings of the Eighth International Conference on Language Resources*.

Tyagi, S. Chopra, D., Mathur, I., Joshi, N. (2015). Classifier-Based Text Simplification for Improved Machine Translation, *Proceedings of International Conference on Advances in Computer Engineering and Applications*.

Xu, H. and Koehn, P. (2017). Zipporah: a Fast and Scalable Data Cleaning System for Noisy Web-Crawled Parallel Corpora, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp.2945–2950.

Yogi, K. Jha, C. K., Dixit, S. (2015). Classification of Machine Translation Outputs Using NB Classifier and SVM for Post-Editing. doi:10.5121/mlaij.2015.2403

**Appendix A Detailed Presentation of Noise in the CCMatrix Coprus (based on a 200-sentence sample)**

Type of Noise	% of sentences
Punctuation and capitalisation	11.5
"..."	5.5
Capitalisation	1
Symbols	4.5
Misplaced Punctuation	0.5
Numbers/Dates	15
Numbers	11.5
Dates	2
Years	1.5
URLs	1
Long sentences	6
Abbreviations	7
In EN	2.5
In BG	4.5
Foreign language	5
EN	4.5
Other	0.5
Machine-translated	4.5
Non-standard language	8.5
Sentence fragments	7.5
Typoes	0.5
Slang	0.5

Table 3: Noise in Bulgarian sentences

Type of Noise	% of sentences
Punctuation	6.5
"..."	2
Symbols	4.5
Numbers/Dates	16.5
Numbers	12.5
Dates	2.5
Years	1.5
URLs	0.5
Long Sentences	1
Abbreviations (EN)	7.5
Foreign Language	5
EN	4.5
Other	0.5
Non-standard language	6
Hiragana + kanji/katakana	2
Sentence fragments	3.5
"Texting" language	0.5

Table 3: Noise in Japanese sentences

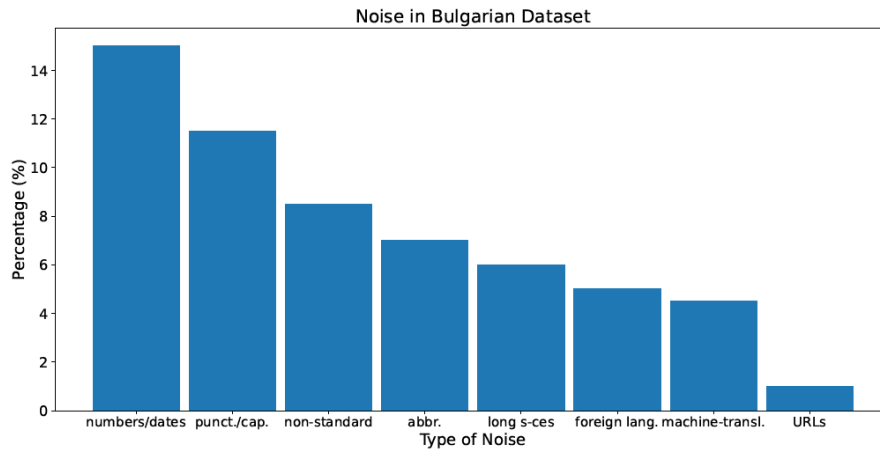


Figure 3: Noise in CCMatrix’s Bulgarian sentences by type.

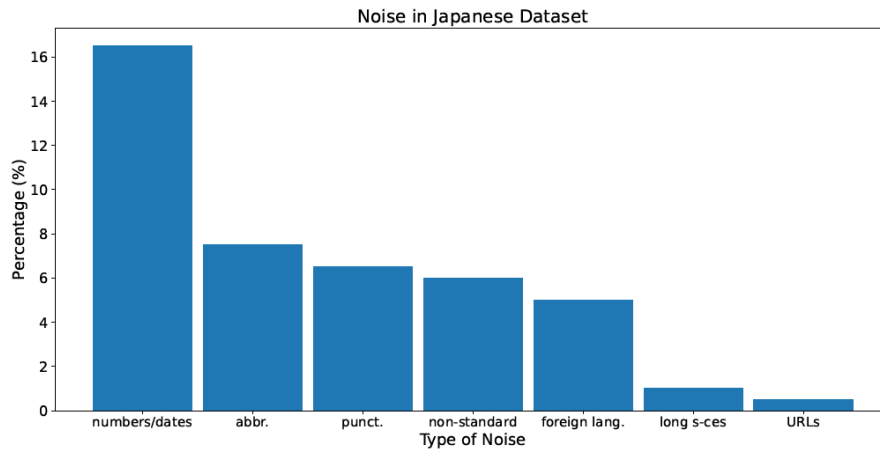


Figure 4: Noise in CCMatrix’s Japanese sentences by type.

# Hate Speech Classification in Bulgarian

**Radoslav Ralev**

Technical University of Munich  
Department of Informatics  
80333 Munich, Germany  
radoslav.ralev@tum.de

**Jürgen Pfeffer**

Technical University of Munich  
School of Social Sciences and Technology  
80333 Munich, Germany  
juergen.pfeffer@tum.de

## Abstract

In recent years, we have seen a surge in the propagation of online hate speech on social media platforms. According to a multitude of sources such as the European Council, hate speech can lead to acts of violence and conflict on a broader scale. That has led to increased awareness by governments, companies, and the scientific community, and although the field is relatively new, there have been considerable advancements in the field as a result of the collective effort. Despite the increasingly better results, most of the research focuses on the more popular languages (i.e., English, German, or Arabic), whereas less popular languages such as Bulgarian and other Balkan languages have been neglected. We have aggregated a real-world dataset from Bulgarian online forums and manually annotated 108,142 sentences. About 1.74% of which can be described with the categories racism, sexism, rudeness, and profanity. We then developed and evaluated various classifiers on the dataset and found that a support vector machine with a linear kernel trained on character-level TF-IDF features is the best model. Our work can be seen as another piece in the puzzle to building a strong foundation for future work on hate speech classification in Bulgarian.

**Keywords:** hate speech, natural language processing, classification, Bulgarian

## 1 Introduction

The term "hate speech" means public speech that expresses hate or encourages violence toward a person or group based on race, religion, sex, or sexual orientation<sup>1</sup>. Hate speech is not something new. We can find evidence of it throughout history ranging from Ancient Greece, through Rome and the middle ages up to modern times. It is no

<sup>1</sup>[www.dictionary.cambridge.org/us/dictionary/english/hate-speech](http://www.dictionary.cambridge.org/us/dictionary/english/hate-speech)

surprise that during times when the most prominent thinkers were freely expressing their hateful opinions and discrimination against minorities was part of both the law and religion, hate speech was omnipresent. However, identifying hate speech is a complex problem. Who decides what hate speech is? Aristotle would probably not consider his writings hateful, but two thousand years later, we might.

Today, social media platforms can enable people with discriminatory views to express their opinions more openly and under anonymity. Furthermore, there have been multiple occasions in which there is a connection between online hate speech and increased violent hate-based activities. Two very prominent examples of increased hate speech online following real-world events are a) hate speech towards immigrants and Muslims following the Manchester and London attacks after the UK left the EU. (Travis, 2017); b) an uptick in racist and xenophobic harassment incidents following the Presidential election in the US. (Okeowo, 2017). By the year 2020 hate crime had already achieved global recognition. In total, 118 countries and international organizations have laws on hate speech<sup>2</sup>.

The connection between hate speech and hate crime has also already been studied more thoroughly in academia (Müller and Schwarz, 2021). In general, studying human behavior at scale by utilizing social media data has been the focus of researchers' attention for 15 years (Lazer et al., 2009). While much research has been devoted to big platforms like Facebook and Twitter and focuses on a small number of languages, more recently, research on smaller and more specialized communities (Mooseder et al., 2022) and less popular languages (Nurce et al., 2021; Shekhar et al., 2020; Ljubešić et al., 2018) has become increas-

<sup>2</sup>[www.futurefreespeech.com/global-handbook-on-hate-speech-laws](http://www.futurefreespeech.com/global-handbook-on-hate-speech-laws)

ingly visible. We follow this branch of research and focus our attention on content in a language underrepresented in scientific research, namely Bulgarian.

Bulgarian is a language spoken by approximately 8 million people around the globe, however, it plays an important historical role as the first Slavic language to have an official alphabet. It was created and developed in the 9th century AD by the Saints Cyril and Methodius and their disciples. It was the first Slavic language into which the Bible was translated.

Although not as damaging as the examples mentioned above, Bulgaria suffers from an extremely high incidence of hate speech towards representatives of ethnic, religious, or sexual minorities (Lozanova et al., 2017; Ivanova, 2018). Figure 1—showing the sentiment distribution of 1,475 comments in Bulgarian following the Syrian refugee wave—illustrates the gravity of the problem. Article 162, paragraph 1 of the Bulgarian penal code penalizes the more extreme forms of hate speech, hence one can conclude that Bulgaria currently suffers from two problems in terms of hate speech prevention. First, the detection of hate speech and encouragement towards violent acts. Second, is the enforcement of the law. This paper aims to address the first of these two points by collecting, filtering, and manually annotating real-world data, and by implementing, evaluating, and comparing various supervised learning models.

The contributions of this article are:

- We have manually annotated 108,142 Bulgarian sentences and made this dataset publicly available.<sup>3</sup>
- 1,878 of these sentences can be described as being hate speech, namely in the categories racism, sexism, rudeness, and profanity.
- We have tested multiple classifiers and approaches on the dataset and compared their performances.
- The best model in terms of F1 score is a support vector machine with a linear kernel trained on character-level TF-IDF features. The model achieved a macro F1 score of 0.73.

<sup>3</sup><http://www.pfeffer.at/data/bulgarian/>

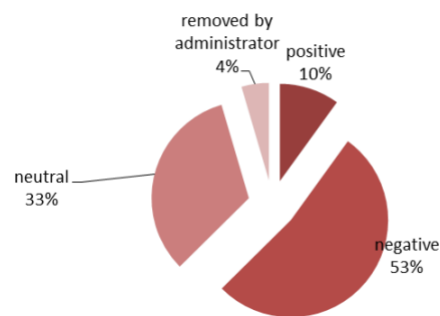


Figure 1: Sentiment distribution of 1,475 comments in Bulgarian, following the Syrian refugee wave (Lozanova et al., 2017).

## 2 Related work

The rising visibility of hate speech on the online social platform has resulted in a continuously growing rate of published research into different areas of hate speech (Tontodimamma et al., 2021). Due to the enormous volume of data that needs to be checked, more focus has been put on automatic detection algorithms.

Detecting hate speech has become an essential topic in the natural language processing community (Mohiyaddeen and Siddiqi, 2021). As a result, a wide range of approaches to text classification was applied, and new datasets were created (Waseem and Hovy, 2016). The issue is that some more minor, less represented languages go under the radar. There have been efforts for language-agnostic text classification (Feng et al., 2020), however, these languages remain mainly ignored by the scientific community. Bulgarian, for example, is one such language.

Some efforts (Dinkov et al., 2019) have been made toward detecting toxicity in news articles in Bulgarian, but the datasets tend to be too small. In recent years there have been numerous advances in natural language processing conducted by Bulgarian researchers on various topics. In (Koeva et al., 2020) the authors present a new corpus of national legislative documents. In (Zhikov et al., 2012) a multi-class multi-label classifier for social news is presented. In (Marinova, 2019) the author compares the performance of classifiers trained on features generated by a variety of state-of-the-art pre-trained embeddings models for tasks such as Named Entity Recognition and Classification (NERC) and Part-of-Speech (POS) Tagging. In (Kapukaranov and Nakov, 2015) a movie review dataset in Bulgarian, sentiment lexicon, and a first-



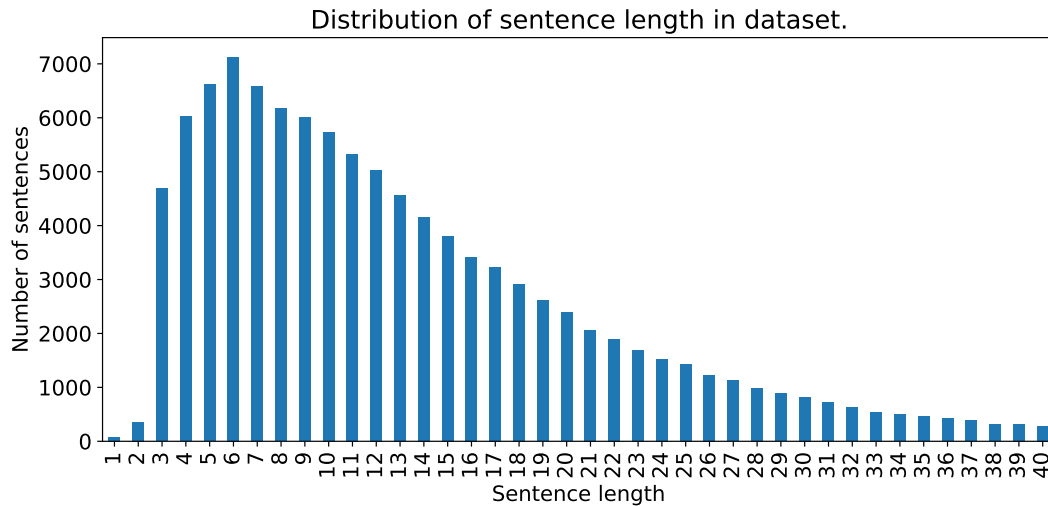


Figure 2: The sentence lengths in our dataset follow a long tail distribution. (Tail has been cut at 40 for better readability).

of-its-kind fine-grained sentiment classifier are presented. Word normalization methods such as stemming (Nakov, 1998) and lemmatization (Iliev et al., 2015) have also been explored, enabling more advanced natural language processing pipelines, sentiment analysis, and others.

To address the problem of hate speech, an automatic detection algorithm has to be created. Usually, this is done by training a machine learning model in a supervised manner for which huge amounts of annotated data are required. Some authors (Waseem and Hovy, 2016) also incorporate social network data features in the model training, however, we have abstained from this and focused explicitly on natural language.

### 3 Data

The data required for our purpose was natural language written informally in Bulgarian and, if possible, written as part of a dialogue or a comment on a subject.

The biggest portion of data was directly provided by the Bulgarian forum BG-Mamma<sup>4</sup> which is the biggest Bulgarian forum and its main user base is mostly comprised of current or future parents. Except the data provided by them we also scraped other forums such as BG-Jargon<sup>5</sup> and BG-Nacionalisti<sup>6</sup> (BG-Nationalists). BG-Jargon is a website that collects Bulgarian slang words and

<sup>4</sup>[www.bg-mamma.com](http://www.bg-mamma.com)

<sup>5</sup><https://www.bgjargon.com/>

<sup>6</sup><https://bg-nacionalisti.org/>

phrases and includes example sentences of how each word is used in everyday life. We have scraped exactly those sentences. BG-Nationalists is an extremist right-wing political forum. About 80% of the data is from BG-Mamma.

The sentences consist on average of 14.3 words (median 11, standard deviation 12.2). All of the websites above contain mostly informal communication. This is further confirmed by the distribution of the sentence length as seen in Figure 2 with most sentences being short but also having a very long tail. While we can find one "sentence" with 685 words, 75% of sentences are 18 words or less. Due to the nature of the main source of the data (BG-Mamma) we were expecting predominantly non-hateful sentences.

#### 3.1 Labeling

Text classification is almost always performed in a supervised way. For this reason, a labeled dataset is required.

At first, we approached this by manually labeling entire "comments" or "opinions" which are multi-sentence posts, however, after a few thousand samples we noticed that within multi-sentence hateful posts, hate usually occurs within only one sentence, hence we decided to do sentence classification instead. We split the initial "comments" dataset into a sentence dataset which greatly increased the sample count. The final result was a total of 108,142 manually annotated sentences, Unfortunately, even before the split, the data was severely imbalanced.

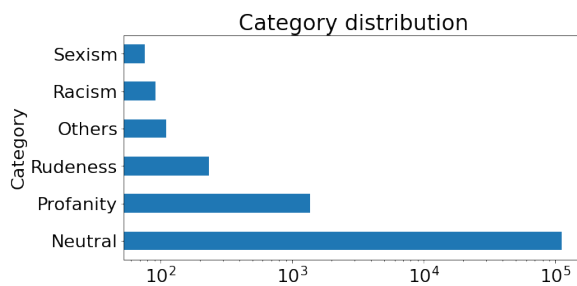


Figure 3: Distribution of each sentence class in the dataset. The x-axis is in a logarithmic scale.

The split made the imbalance even greater as you can see in Figure 3. The major disproportion in the dataset forced us to unify all hateful categories into one and perform a simple binary classification. This led to a dataset with 106,264 non-hateful and 1,878 hateful sentences.

### 3.2 Data preprocessing

Text data is one of the most disorganized and unstructured data types possible. That makes data preprocessing one of the deciding factors for the final quality of a model.

**Cleaning** the BG-Mamma dataset required the most time out of all the text gathered. Originally the text was in a BBCode-format<sup>7</sup>. BBCode tags and other format-specific syntax were removed to clean the text. An algorithm to eliminate posts appearing once as a standalone comment and a second time when they were being mentioned was also developed. Aside from this, HTML code, URLs, punctuation, stopwords, and all numbers were removed. The text was also made lowercase.

**Lemmatization** in linguistics is the act of grouping together different word forms so that a text processing algorithm can recognize them as a single word. In itself, lemmatization is complex because it has to identify the word on a part-of-speech level. For this project, lemmagen3<sup>8</sup> was used.

**Stemming** is usually considered a more naive version of lemmatization. That is due to the fact that stemming does not consider the context of the word, but only its morphology. Stemming removes or stems the last few characters of a word, often leading to incorrect meanings. In (Nakov, 1998) the author argues that stemming and lemmatization have achieved a similar performance in experiments. The stemmer described in that paper

was also used in this project. Along with the software package<sup>9</sup>, different rule sets are provided. All of them are included in the evaluation. Both the lemmatizer and the stemmer were evaluated and compared.

**Vectorization.** As already mentioned, text data is one of the most unstructured data types. One of the worst qualities is that it is of variable length. To offset that, the so-called vectorization is performed. Vectorization is the process of transforming unstructured text into a fixed-size numerical representation (usually a vector) that is easier to understand by a machine (Schütze et al., 2008). There are various ways to do this from simple bag-of-words or bag-of-characters methods and the famous TF-IDF to neural network embeddings (Bengio et al., 2000) using pre-trained models such as BERT (Devlin et al., 2018) or Word2Vec (Mikolov et al., 2013). We have primarily focused on TF-IDF, however, embeddings we have also evaluated some pre-trained embeddings such as the stacked embeddings from FlairNLP (Akbiik et al., 2018, 2019), FastText (Joulin et al., 2016) and others.

**Data imbalance** As previously mentioned, the dataset is significantly imbalanced (Günemann and Pfeffer, 2017). There are various approaches to offset this. For this, we focused on imbalanced-learn’s and scikit-learn’s implementations (Lemaître et al., 2017; Pedregosa et al., 2011). One can address this issue by oversampling the minority class, undersampling the majority class, or a mix of both. Two of the most basic approaches to handling imbalanced data consist of either replicating the minority class samples until the class distribution becomes uniform or providing class weights for each class to the classifier which will correct the loss function correspondingly.

A more advanced technique for oversampling is called ”Synthetic Minority Over-sampling Technique” or just SMOTE (Chawla et al., 2002). The algorithm works by finding the nearest neighbor of a sample point from the minority class in feature space. Then it chooses a random point between them, which is then added to the dataset. This algorithm’s effectiveness has been thoroughly evaluated and usually achieves a performance boost, although some authors suggest that the commonly accepted method for synthetic instance creation may not be the best one (Bajer et al., 2019).

<sup>7</sup><https://en.wikipedia.org/wiki/BBCode>

<sup>8</sup><https://github.com/vpdpodpecan/lemmagen3/>

<sup>9</sup><https://pypi.org/project/bulstem/>

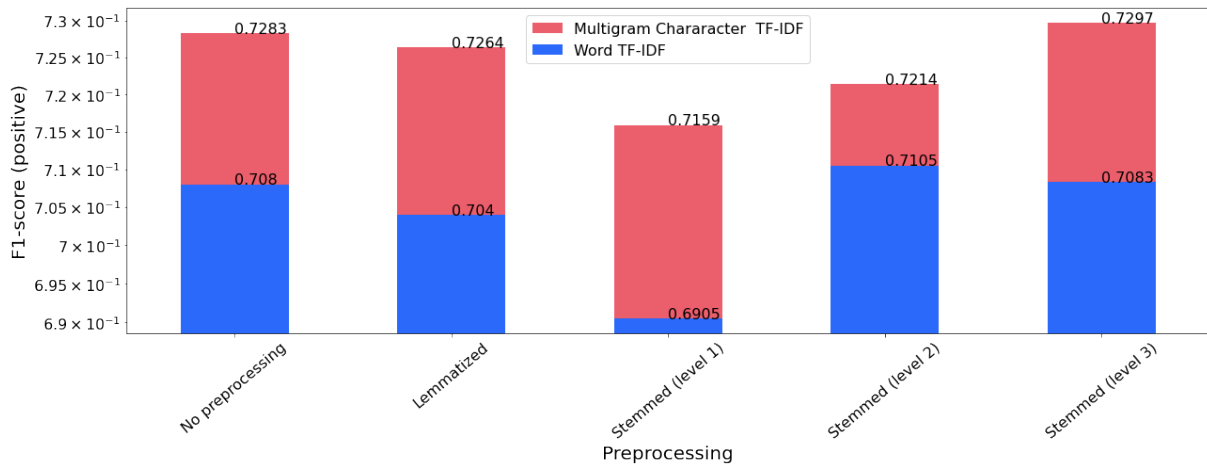


Figure 4: Comparison between the performance of preprocessors with two different vectorizers. (log-scale)

## 4 Models

For the classification, we trained classical machine learning models such as the logistic regression, support vector machines (Platt et al., 1999), decision trees (Breiman et al., 2017), random forests (Breiman, 2001) and naive bayes classifiers (Schütze et al., 2008).

We also evaluated the performance of several neural network architectures. The most basic of which is a shallow neural network with Keras’ (Chollet et al., 2015) embedding layer and TensorFlow’s (Abadi et al., 2015) TextVectorization layer. Another architecture we evaluated is the one discussed in (Zhang et al., 2015). We used another architecture that was based on pre-trained Word2Vec embeddings which was fine-tuned on the corpus, and its weight matrix was used to set the weights of a Keras embedding layer. After that, the architecture proposed above for the Character-Level-CNN was used again. Lastly, we also used the stacked embeddings model for text classification provided by the FlairNLP framework in a similar fashion as in (Marinova, 2019).

## 5 Analysis

The disparity in the distribution of the categories made us rethink how we should observe the classifiers’ performance. A dummy classifier predicting only one class has 98% accuracy. For this reason, the primary metric we used is macro F1 which is an arithmetic mean of the F1-Score for both classes and also the positive F1 score. The imbalance is ignored by using the mean of the two scores, and the two classes are equally weighted.

However, other metrics can also be chosen, such

as balanced accuracy, if the classifier is to be used in more practical settings (e.g., in the industry). *Balanced accuracy* is defined as the average of recall for all categories in the classification. It is used as a substitute for accuracy for imbalanced datasets. It is also much easier to interpret than F-Score.

For the evaluation, the dataset was split into a training and testing set (75% train, 25% test). A 75-25 proportion instead of 70-30 was used because it allowed for a better distribution of the main five categories in both datasets. Furthermore, although binary classification was performed, due to the data imbalance, the data was still split following a stratified approach for all five categories to achieve a similar distribution in both datasets. That was done to offset any additional bias towards one of the categories.

### 5.1 Comparing preprocessing techniques

**Stemming vs. Lemmatization** Before evaluating the performance differences in vectorizers, we wanted to see which preprocessing technique was the best. To do that, we prepared five pipelines: one without any preprocessing, one with lemmatization enabled, and three with stemming enabled, each with a different, more punishing, rule-set. At first, we used only the basic word TF-IDF vectorizer but later, after finding the best vectorizer (see following subsections), we decided to re-do this evaluation. The classifier used is scikit-learn’s support vector machine implementation with a linear kernel (also called LinearSVC), but similar outcomes were observed with other classifiers.

Figure 4 depicts the results. The results show that stemming outperforms lemmatization. Espe-

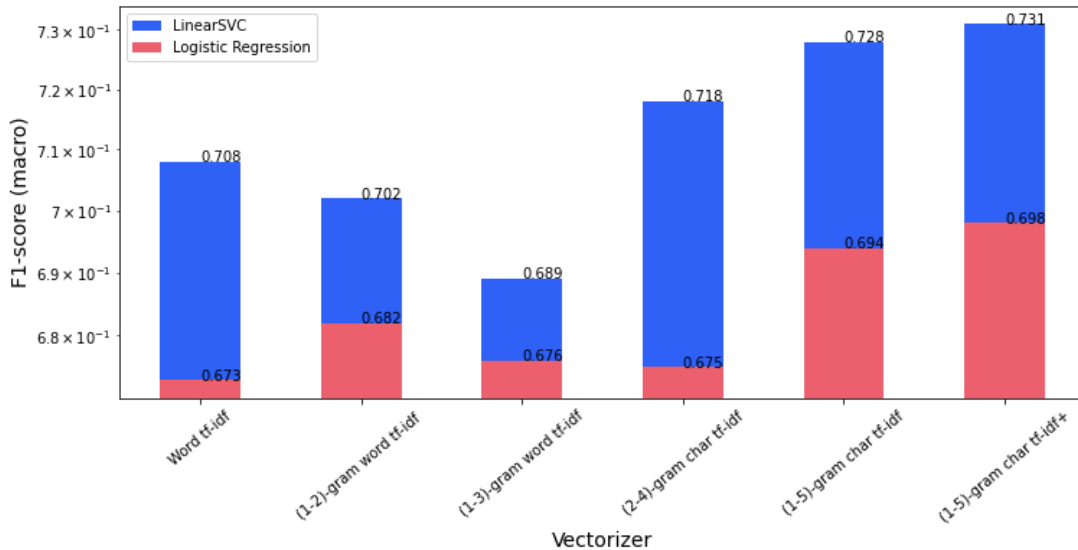


Figure 5: Performance comparison of vectorizers

Classifier	(macro) f1-score	precision	recall
SimpleNN	0.686	0.4246	0.347
Word2Vec	0.6523	0.2675	0.3821
Word2Vec+CNN	0.6850	0.3232	<b>0.4615</b>
FlairNLP	<b>0.7172</b>	<b>0.5193</b>	0.3860
Character-Level CNN	0.6359	0.2828	0.2808

Table 1: Performance comparison between all neural networks.

cially in a setting where words are used as features, level two stemming performs best. However, omitting to preprocess helps boost classifier performance when using character-level features, and despite the better performance, at first sight, we believe that omitting the stemming adds more robustness to the model when using a character-level vectorization. This is because the Bulgarian language is very rich in prefixes and suffixes and stemming at such a high level might disrupt the meaning of a word. Hence we have decided to stick to no preprocessing and unless otherwise specified, everything will be evaluated on a dataset without lemmatization or stemming in the following subsections.

**Vectorizers** In total, six vectorizers were evaluated on two different classifiers—a logistic regression and a linear support vector machine (LinearSVC). Three word-level and three character-level vectorizers were chosen. The classifiers’ macro F1-Score performances are visualized in Figure 5. From the Figure, it can be seen that for both classifiers, the character-level preprocessing tends to outperform the word-level vectorization.

Hence, unless otherwise specified, from this point onward, everything will be evaluated on data with character-level uni- to pentagrams.

Another key consideration is that some character n-grams that are too often seen in the dataset can be ignored due to the enormous class imbalance. This parameter is called maximum document frequency. The rightmost bar on the figure (“(1-5)-gram char tf-idf+”) is the same as the one before it but includes a maximum document frequency of 40% as well. As it can be seen, although it does offer an increase in performance, it is more or less negligible.

## 5.2 Models

**Neural networks** For the more basic network, TensorFlow’s Text Vectorization layer was used, again at a character level (but this time without TF-IDF enabled due to an immense increase in training times), followed by an embedding layer with an output dimension of 64. After that, the output of the embedding layer goes through a 1D max. pooling layer to reduce the dimensions and is consequently fed into a sequence of three dense layers, each with 64 neurons and a ReLU activation. We have not

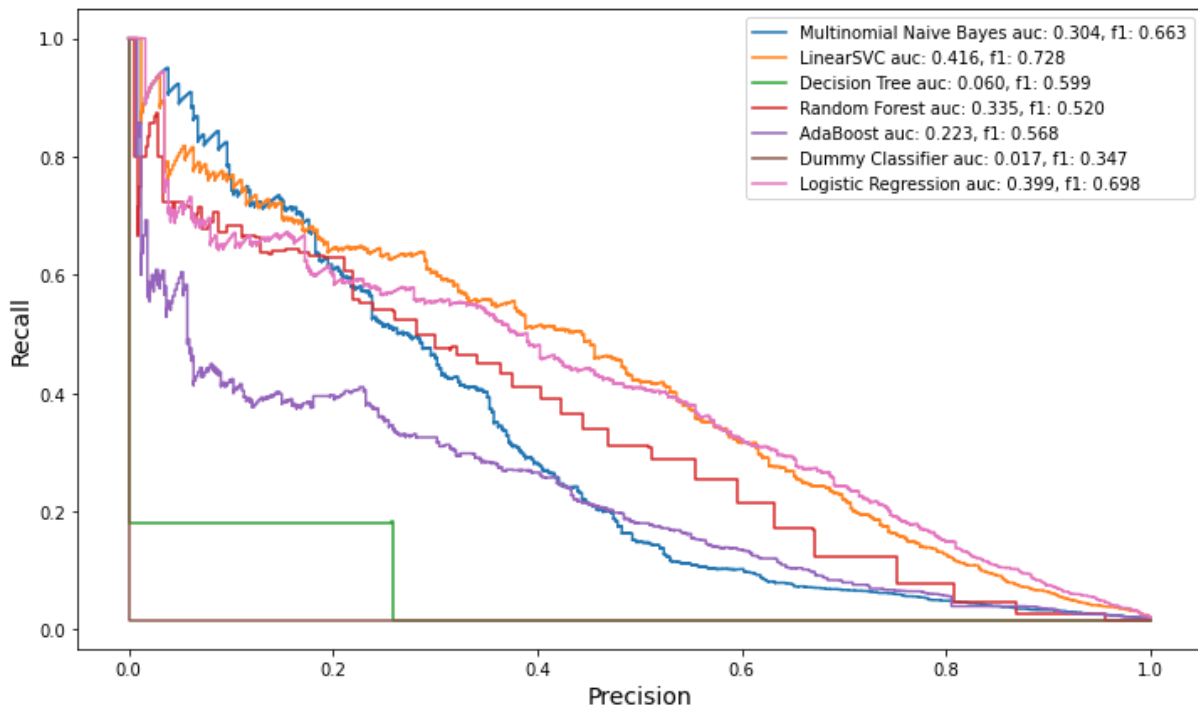


Figure 6: Precision-Recall-Curves for all classical machine learning classifiers without oversampling.

included any of the more simple neural networks with LSTM/Convolutional layers in this evaluation as they did not increase the performance of the model significantly enough.

The other networks are the FlairNLP stacked embeddings network, the Character-Level CNN from (Zhang et al., 2015), and the Word2Vec embeddings network (once with convolutional layers and once without). The results are summarized in Table 1.

All in all, the FlairNLP stacked embeddings model achieved the best performance. It is also the slowest model to train and uses the most pre-trained embeddings (four pre-trained models in total). The Word2Vec model with the CNN architecture and the simple neural network come in close second and third. An interesting note is that the Word2Vec+CNN model achieved the best recall score. A surprise was the performance of the Character-level CNN. It is the second-largest model on the list with a total of 96M parameters but it performed worse even than the simple neural network.

**Classical machine learning models** Firstly, a naive dummy classifier was created to benchmark the other models. The dummy classifier predicts using a uniform strategy, so each class has equal probability. After that, seven classifiers with default parameters were trained once on the dataset prepro-

Classifier	Balanced Accuracy
SimpleNN	0.6695
Word2Vec+CNN	<b>0.7224</b>
LinearSVC	0.4202
Logistic Regression	0.5350

Table 2: Performance comparison between selected models in terms of balanced accuracy.

cessed as discussed in previous sections and once on the same dataset but additionally with SMOTE oversampling enabled. Surprisingly, most of the classifiers either underperformed or showed no significant improvement on the oversampled dataset and were thus omitted for the sake of brevity.

The results are shown in Figure 6. The classifiers are compared based on their precision-recall curves, as well as the overall area under the curve (auPRC) and f1-score. The overarching winner in both setups was the linear support vector classifier with logistic regression coming in a close second. In the end, the LinearSVC managed to achieve an f1-score of **0.728**.

### 5.3 Balanced Accuracy

As previously mentioned, although F1 is the standard metric for comparing classifiers, in a more practical setting, better metrics can be found. The



main reason for this is that F1 is not as easy to interpret as other metrics may be. An excellent example of a suitable metric for a scenario like that is balanced accuracy.

Much to our surprise, some of the worst classifiers in terms of F1 are, in fact, the best ones in terms of balanced accuracy. In Table 2, we can see a selection of the previously evaluated models. As it becomes clear from the table, although the classical machine learning models are indeed the overall winners in terms of F1-score, they fall behind in terms of balanced accuracy.

## 6 Discussion

Hate speech has always been a problem in society. The internet revolution reinforced the problem by providing instant connectivity across social media and anonymity. There also exists mounting evidence of a connection between hate speech online and hate crime. All of this has led to increased attention towards hate speech not only from the general public, but also from governments and private organizations.

Because of its online nature, and hence the amount of data that is being constantly generated, hate speech lends itself very well to automatic detection by an artificial intelligence model. To do this, however, large and robust datasets are required, and although they do exist, most of them are focused on languages with a strong internet presence such as English. As a result, many of the not so well represented languages—such as Bulgarian—are mostly ignored.

Multiple reports have shown that hate speech is an even greater problem in Bulgaria than in other countries. For this reason, the scientific community in Bulgaria should follow in the footsteps of such communities in other countries and focus on the issue. A first step in doing that is to create datasets that can be used for training purposes of future research. We believe that by sharing our dataset consisting of 108,142 manually annotated sentences, we can contribute making that first step.

A further contribution of this paper is the evaluation of a variety of machine learning methods for the task of text classification in Bulgarian in an imbalanced setting, including some state-of-the-art approaches.

## 7 Future Work

Despite the continuing efforts of the scientific community, there are some fundamental issues with solving hate speech classification. For example, in (Arango et al., 2019), the authors argue that researchers have become overly optimistic about the results of their classifiers. That is because most research papers focus only on datasets coming from one source. That causes the models (usually deep neural networks) to overfit and are rarely able to generalize well on new datasets. Therefore, the creation of multiple datasets is compulsory for the development of a robust predictive model.

Another issue is annotator bias. In (Waseem, 2016), Waseem discusses how much the influence of annotators on the performance of classifiers and suggests that systems trained on data labeled by experts perform better than those labeled by amateurs. That leads us to another fundamental issue with hate speech classification: who defines what hate speech is? To mitigate any annotator bias future datasets should not only be annotated by experts but also, if possible, by different people.

Another important point that could be addressed in the context of imbalanced text classification is data augmentation (DA). Data augmentation is the process of creating artificial data to improve the performance of a classifier. One can argue that some aspects of data augmentation are already included by incorporating SMOTE into the preprocessing pipeline; however, SMOTE works on the feature space of the vectorized textual data. What might greatly impact the classifier’s performance would be to augment the data at the textual level. There are multiple ways of performing this, from using a thesaurus to substitute words on a synonym level to using model embeddings (for example, Word2Vec) to sample neighboring words. This approach has been shown to increase the performance of hate speech classifiers. (Rizos et al., 2019; Bayer et al., 2021)

A further unexplored method to improve the classifiers’ performance is employing additional feature engineering methods such as named entity recognition and part-of-speech taggers. These methods would enrich the feature space and result in a better classifier.

Lastly, if the context in which a model is to be used is social media, a further feature engineering idea would be to take user metadata such as age, location, or gender into account. Furthermore,

a "hate score" could be calculated for each user based on her or his past posts or her/his connections' past posts by utilizing social network analysis techniques.

## Acknowledgments

The authors are grateful to Marina Kuzmanova from BG-Mamma for providing the biggest portion of the dataset. The labeled dataset of this article can be found here: <http://www.pfeffer.at/data/bulgarian/>.

## References

- Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. *TensorFlow: Large-scale machine learning on heterogeneous systems*. Software available from tensorflow.org.
- Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.
- Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.
- Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54.
- Dražen Bajer, Bruno Zonć, Mario Dudjak, and Goran Martinović. 2019. Performance analysis of smote-based oversampling techniques when dealing with data imbalance. In *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 265–271. IEEE.
- Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2021. A survey on data augmentation for text classification. *arXiv preprint arXiv:2107.03158*.
- Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13.
- Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.
- Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. 2017. *Classification and regression trees*. Routledge.
- Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.
- François Chollet et al. 2015. Keras. <https://keras.io>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2019. Detecting toxicity in news articles: Application to bulgarian. *arXiv preprint arXiv:1908.09785*.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.
- Nikou Günnemann and Jürgen Pfeffer. 2017. Predicting defective engines using convolutional neural networks on temporal vibration signals. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, volume 74 of *Proceedings of Machine Learning Research*, pages 92–102. PMLR.
- Grigor Iliev, Nadezhda Borisova, Elena Karashtranova, and Dafina Kostadinova. 2015. [A publicly available cross-platform lemmatizer for bulgarian](#).
- Ivanka Ivanova. 2018. Public attitudes to hate speech in bulgaria in 2018. Technical report, Open Society Institute Sofia.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.
- Borislav Kapukaranov and Preslav Nakov. 2015. Fine-grained sentiment analysis for movie reviews in bulgarian. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 266–274.
- Svetla Koeva, Nikola Obreshkov, and Martin Yalamov. 2020. Natural language processing pipeline to annotate bulgarian legislative documents. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6988–6994.



- David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational social science. *Science*, 323(5915):721–723.
- Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.*, 18(1):559–563.
- Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. 2018. [Datasets of Slovene and Croatian moderated news comments](#). In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 124–131, Brussels, Belgium. Association for Computational Linguistics.
- Denitza Lozanova, Sevdalina Voynova, Snezhina Gabova, and Svetlana Lomeva. 2017. Mapping out the national context of online hate speech in bulgaria. Technical report, Coalition of Positive Messengers to Counter Online Hate Speech.
- Iva Marinova. 2019. [Evaluation of stacked embeddings for Bulgarian on the downstream tasks POS and NERC](#). In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 48–54, Varna, Bulgaria. INCOMA Ltd.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Mr Mohiyaddeen and Sifatullah Siddiqi. 2021. Automatic hate speech detection: A literature review. Available at SSRN 3887383.
- Angelina Mooseder, Momin M. Malik, Hemank Lamba, Earth Erowid, Sylvia Thyssen, and Jürgen Pfeffer. 2022. Glowing experience or bad trip? A quantitative analysis of user reported drug experiences on erowid.org. In *Proceedings of ICWSM 2022*.
- Karsten Müller and Carlo Schwarz. 2021. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167.
- Preslav Nakov. 1998. Bulstem: Design and evaluation of inflectional stemmer for bulgarian.
- Erida Nurce, Jorgel Keci, and Leon Derczynski. 2021. [Detecting abusive albanian](#). *CoRR*, abs/2107.13592.
- Alexis Okeowo. 2017. [Hate on the rise after trump’s election](#). *The New Yorker*.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.
- Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 991–1000.
- Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.
- Ravi Shekhar, Pranjić. Marko, Senja Pollak, Andraž Pelicon, and Matthew Purver. 2020. [Automating News Comment Moderation with Limited Resources: Benchmarking in Croatian and Estonian](#). *Journal for Language Technology and Computational Linguistics*, 34(1):49–79. [https://jcl.org/content/2-allissues/1-heft1-2020/jlcl\\_2020-1\\_3.pdf](https://jcl.org/content/2-allissues/1-heft1-2020/jlcl_2020-1_3.pdf).
- Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126(1):157–179.
- Alan Travis. 2017. [Anti-muslim hate crime surges after manchester and london bridge attacks](#). *The Guardian*.
- Zeerak Waseem. 2016. [Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter](#). In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.
- Zeerak Waseem and Dirk Hovy. 2016. [Hateful symbols or hateful people? predictive features for hate speech detection on Twitter](#). In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.
- Valentin Zhikov, Ivelina Nikolova, Laura Toloşi, Yavor Ivanov, Borislav Popov, and Georgi Georgiev. 2012. Enhancing social news media in bulgarian with natural language processing. *INFOthea*, 2(13):6–18.

# WordNet-Based Bulgarian Sign Language Dictionary of Crisis Management Terminology

**Slavina Lozanova**

Faculty of Educational Studies and the Arts  
Sofia University /  
Deaf Studies Institute, Bulgaria  
lozanovaslavina@gmail.com

**Ivelina Stoyanova**

DCL – IBL  
Bulgarian Academy of Sciences /  
Deaf Studies Institute, Bulgaria  
iva@dcl.bas.bg

## Abstract

This paper presents an online Bulgarian sign language dictionary covering terminology related to crisis management. The pressing need for such a resource became evident during the COVID pandemic when critical information regarding government measures was delivered on a regular basis to the public including Deaf citizens.

The dictionary is freely available on the internet and is aimed at the Deaf, sign language interpreters, learners of sign language, social workers and the wide public.

Each dictionary entry is supplied with synonyms in spoken Bulgarian, a definition, one or more signs corresponding to the concept in Bulgarian sign language, additional information about derivationally related words and similar signs with different meaning, as well as links to translations in other languages, including American sign language.

**Keywords:** Online dictionary, Bulgarian sign language, WordNet, crisis management, COVID.

## 1 Introduction

The Deaf community is a minority community characterised by its own history (history of the Deaf), original culture (culture of the Deaf) and social life, all of which are based on a specific territorial sign language (whether officially recognised in the country or not). The Bulgarian Sign Language (BGSL) was officially recognised in Bulgaria in January 2021 as the language of the Deaf community. The official recognition guaranteed Deaf people's right to access to information and education through sign language.

However, the Deaf community is heterogeneous and the individual specifics of language development, the modes of communication, etc. vary significantly between users. There is a group of sign

language users who acquire the language in the family at an early age and it becomes their primary mode of communication. They acquire spoken language (to a various degree depending on their hearing and spoken skills) through school and speech and language therapy.

When sign language is acquired at a later age, after relatively good verbal language skills have been developed, sign language competence is built on verbal competence, and in this case sign language is used as a second language. Over time, both languages can be used simultaneously, and in some cases sign language can also play a dominant role in the deaf person's daily communication. However, predominantly in this case the verbal language influences the sign language and we observe 'signed Bulgarian' rather than the authentic sign language.

This paper presents an online Bulgarian sign language dictionary covering terminology related to crisis management. The need for such a resource became very pressing during the COVID pandemic when critical information regarding government enforced measures was delivered on a regular basis to the public. Although government briefings were supplied with sign language interpreting, many sign language users faced difficulties in understanding properly and fully the information. There were words that had no signs known to the Deaf community at large, or such signs varied significantly between users and local Deaf communities. We attempted to collect and present variations of the signs, registering preferences among the users and raising discussion within the community with respect to particular signs and their meaning.

Our approach towards building the dictionary relies on linking it to WordNet as a large lexical-semantic resource. In this way we are able to employ all the descriptive information on the concepts that is available in WordNet and the Bulgarian

WordNet (BulNet), but also to use the numerous semantic relations between concepts.

The dictionary is available freely on the internet and is aimed at the Deaf community, Bulgarian sign language interpreters, as well as interpreters of other low resourced sign languages, learners of Bulgarian sign language, social workers, sign language researchers and the wide public.<sup>1</sup>

The structure of the paper is as follows. Section 2 discusses the challenges sign language communication poses to Deaf users in time of crisis and thus, presents the motivation behind the creation of the dictionary. Section 3 provides a brief overview of related works, mainly sign language dictionaries available online for different well resourced and studied sign languages. Section 4 outlines the steps in compiling the verbal side of the dictionary including the construction of the text corpus, its processing, keyword extraction, word sense assignment. The collection of signed speech, sign annotation and analysis of variations of signs is presented in section 5. It is followed by a description of the structure and components of the dictionary (section 6) focused on specific sign language features and their representation. The last part (section 8) gives some directions for future work both on expanding the coverage of the dictionary and improving the description of entries and the possible applications of the resources.

## 2 Specific features of sign language communication and standardisation with view to crisis management

When communicating through sign language, the following descriptive parameters of the performed signs are important (Valli et al., 2005; Baker, 2016):

- hadshape – the configuration of the hand(s) and the position of the fingers;
- palm orientation – the position of the palm(s) during signing;
- movement – the direction of movement or the fixed, stative position of the hand(s) during signing;
- location where the sign is performed relative to the body;
- non-manual expression – facial or body signals.

---

<sup>1</sup><https://study.deafstudiesinstitute.bg/course/view.php?id=8>

The different parameters and their combination change the meaning of the message, e.g. see examples of minimal pairs of signs<sup>2</sup>. These specifics need to be taken into account when building a dictionary of sign language and especially with a view to crisis management where the precision and punctuality of the delivered message is of paramount importance. Ambiguity of signs as well as signs with similar presentation, in particular with a view to the way they appear on screen (e.g., in TV broadcast, online video, etc.), need to be analysed and avoided, if at all possible.

Crisis management applies to different situations and in dynamic circumstances – situations of crisis, evacuation, emergency, natural disasters (earthquake, fire, flood), extreme weather such as heavy rain and snow, tornadoes, etc. (Manoj and Baker, 2007). The message should be delivered efficiently and clearly in sign language by an experienced interpreter. This raises the need for standardisation so that the language used is understood over the whole territory of Bulgaria and by all sign language users irrespective of their predominant mode of communication, sign language variety acquired and level of language skills. This in turn necessitates the comparative analysis of the variations in signs in order to facilitate the standardisation process.

Standardisation can be aimed at both the verbal and the sign language used in times of crisis when communication with Deaf citizens. Standard verbal language messages can be compiled and taught to Deaf school children as well as adults in order to familiarise them with common text patterns used in warning messages in crisis situations. This type of formulaic language is used in many areas such as airplane safety messages, traffic signs, etc.

Sign language standardisation is not a random choice of formal gestures, but a complex process that takes into account a number of linguistic, pragmatic and sociolinguistic factors related to the domains of communication, the diversity of territorial and social variations, the influence and acquisition of signs from foreign sign languages, the language needs of different groups of deaf people and many others. It is essential in this process that standardisation is not at the expense of linguistic diversity and richness, which deprives the users of linguistic means and productive models for expressing meanings and their nuances. The World Federation

---

<sup>2</sup><https://www.handspeak.com/learn/index.php?id=109>

of the Deaf has warned against negative trends in standardisation which in the long term alienate and deprive language communities of their authentic language<sup>3</sup>.

Standardisation is essential for the provision of quality interpreting services and is a long and controlled process based on language analysis and conscious attitude towards the language by its speakers, supported by sign linguists, interpreters and other professionals. In this sense, the standardisation of sign languages is a responsible activity, as much as the construction of literary verbal languages (for decades), as well as their enrichment and development through research, language training of native speakers and new learners.

### 3 Related work

There are many large dictionaries for sign languages across the world which have been made available online: American sign language (ASL)<sup>4</sup>, British sign language (BSL)<sup>5</sup>, Australian sign language (Auslan)<sup>6</sup>, German sign language (DGS, Langer et al.)<sup>7</sup>, Swedish sign language (STS)<sup>8</sup>, among others. Although these dictionaries are predominantly monolingual, in recent years there have been efforts to create some multilingual or linked dictionaries across several languages, either general such as Spread the Sign<sup>9</sup>, or domain-specific such as Hands in the Stars (specialised in astronomy)<sup>10</sup>.

For the Bulgarian sign language the largest modern dictionary is available only as a book both in printed and electronic format (Tisheva et al., 2017).

During the COVID pandemic many of the larger sign language dictionaries included the new concepts or those that gained popularity and were essential for the management of the crisis: coronavirus, COVID-19, pandemic, etc. Additional efforts have been focused on preparing informational

materials in many sign languages to inform the Deaf about the pandemic. Information materials have been developed for children as well. An example is the initiative of Rise e-books to present coronavirus stories for children<sup>11</sup>.

The development of the dictionary presented in this paper relies on its linking to Princeton WordNet (Miller et al., 1990; Fellbaum, 1999) and the Bulgarian counterpart, BulNet (Koeva, 2010), modelled after the Princeton WordNet. This approach facilitates the exploration of all semantic relations within the network (Ruppenhofer et al., 2016), as well as using the links to other languages (Vossen, 2002, 2004; Bond and Foster, 2013) and resources (Shi and Mihalcea, 2005; Leseva and Stoyanova, 2020) to expand the resource and its applications in both human-oriented products (e.g., resources for language learning for Deaf users) or natural language processing (e.g., in processing multimodal content such as sign language production, machine translation, question answering, etc.).

There have been limited attempts to link sign language dictionaries to WordNet (Lualdi et al., 2019, 2021; Wright, 2021). The mapping of WordNet senses to signs faces similar challenges as the development of WordNet for other minority languages with limited resources (Bella et al., 2020).

To the best of our knowledge, no efforts exist towards building a crisis management sign language dictionary which includes Bulgarian sign language. Also, there are no initiatives at present aiming at standardisation of crisis-related terminology in Bulgarian sign language or establishing any principles and considerations regarding standardisation.

### 4 Selection of concepts for the Bulgarian Sign Language Dictionary of Crisis Terminology

The selection process of the key concepts to be included in the Dictionary included the following steps. First, a large text corpus of briefings and COVID-related news was compiled and automatically processed. Secondly, a list of keywords were extracted. Thirdly, the keywords were matched to candidate WordNet senses and then manually disambiguated. This process resulted in a selection of over 4,000 concepts which are then filtered down to 500 most frequent concepts in the sign language data (see section 5).

<sup>3</sup><https://wfdeaf.org/news/wfd-statement-on-standardized-sign-language/>

<sup>4</sup><https://www.handspeak.com/word/>, <https://www.signasl.org/>

<sup>5</sup><https://www.british-sign.co.uk/british-sign-language/dictionary/>, <https://www.signbsl.com/>

<sup>6</sup><https://auslan.org.au/about/dictionary/>

<sup>7</sup><https://www.sign-lang.uni-hamburg.de/glex/intro/inhalt.html>

<sup>8</sup><https://teckensprakslexikon.su.se/>

<sup>9</sup><https://www.spreadthesign.com/en.gjb/search/>

<sup>10</sup><https://www.iau.org/news/pressreleases/detail/iaul706/>

<sup>11</sup><https://riseebooks.wixsite.com/access/copy-of-coronavirus-stories>



#### 4.1 Text Corpus: compilation and processing

The text corpus was automatically compiled by crawling the official website publishing regular briefings and news articles on COVID and the measures enforced by the government<sup>12</sup>. A set of televised video recordings have been automatically transcribed using Google Cloud Speech-to-text API<sup>13</sup>. Since this process was aimed at collecting preliminary material for analysis, precision of transcriptions was not considered and no manual evaluation or editing was performed.

The compiled text corpus included 158 official briefings and 282 news articles with a total of 365 thousand words. The texts have been tokenised, lemmatised and POS-tagged using the Bulgarian Language Processing Chain (Koeva and Genov, 2011)<sup>14</sup>.

#### 4.2 Keyword extraction and classification

For keyword extraction we apply the following procedure: (a) we filter out words from closed classes such as prepositions, pronouns, etc., as well as general stop-words with no domain specific meaning – the stoplist was compiled to include words that appear with high frequency in many different domains in the Bulgarian National Corpus (Koeva et al., 2010); (b) we use frequency ranking of full meaning words to identify keywords typical for the whole corpus; (c) we use the TF-IDF (term frequency-inverse document frequency) method to identify keywords at document level.

As a result, in the first stage we identify a list of 4,350 candidate keywords which are single words – nouns, verbs, adjectives and adverbs.

The identified keywords were manually validated and classified into six predetermined domains: (1) Healthcare, (2) Governance, (3) Statistics and data presentation, (4) Economy and finance, (5) Social care, and (6) Crisis. In about 15% of the cases words are assigned more than one domain (e.g., bg. *epidemiya* – *epidemic* is categorised both in the domains of Healthcare as well as Crisis).

Additionally, the list was expanded with 212 multiword expressions which appeared with high frequency in the text corpus and for which usually one of the components has been identified as a

keyword (e.g., we added bg. *bolnichno otdelenie* – *hospital ward* where only the adjective bg. *bolnichen* – *hospital* has been identified as a keyword).

For each selected keyword (single word or multiword expression) we compiled a list of usage examples from the text corpus allowing us to check the sense in which the word is used in the data.

#### 4.3 WordNet sense assignment and disambiguation

For each keyword we automatically identified all potential WordNet senses that apply to it – from the Bulgarian WordNet (Koeva, 2010) we found all synsets that the keyword appeared as a literal in. Then the appropriate sense was manually selected and assigned after analysing the examples from the text corpus.

In some cases more than one sense of the word appeared in the dataset (e.g., bg. *seriozen* is met both in the meaning of *serious*: bg. *seriozno sastoyanie* – *serious condition* and *strict*: bg. *seriozni merki* – *strict measures*).

After a unique WordNet sense has been assigned to the keyword, all its synonyms (if available in BulNet) and the definition were extracted and added to the description of the keyword.

There were also cases (around 9%) where no WordNet sense was a match, or the word was not found in BulNet. In those cases the definition was created manually.

### 5 Sign language data collection and processing

After the preliminary lists of keywords in the different domains have been prepared, we started collecting and processing the sign language material. Principles of work has been established after the first stages of the data collection since there is very limited experience nationwide in collecting linguistic data in Bulgarian sign language.

#### 5.1 Sign language data collection

Sign language data was collected during six online meetings with Deaf sign language users. Each meeting had a particular topic – one of the domains (see 4.2), and was lead by two Deaf moderators and was recorded in video format. All participants are displayed on the screen simultaneously (the speaker was not put in spotlight) since very often they spoke in sign language simultaneously and we wanted to collect as much data as possible. A

<sup>12</sup><https://coronavirus.bg/>

<sup>13</sup><https://cloud.google.com/speech-to-text>

<sup>14</sup><http://dcl.bas.bg/dclservices/index.php>

screenshot of a recording is shown in Fig. 1 where several signers sign simultaneously (top row second from the left, middle row rightmost signer, and bottom row leftmost and rightmost signers). Some of these signs express confirmation, rejection or other evaluation on the sign performed by a moderator, which is also relevant information although we have not used it at this stage.

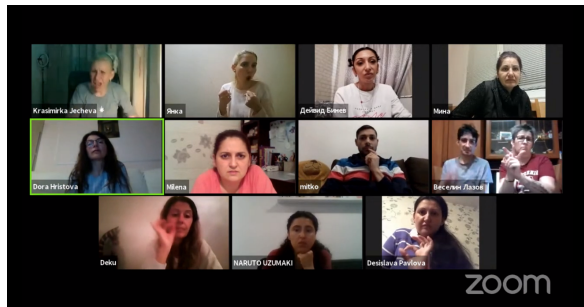


Figure 1: A screenshot of meeting recording

The participants (usually between 8 and 12) were from various cities across the country to ensure representativeness of the main regions formed around the large Deaf regional centres in Sofia, Plovdiv, Varna, Gorna Oryahovitza and Burgas.

For each meeting the moderators had prepared a list of discussion questions which involved the target concepts of the selected keywords. In some cases the concepts under observation were directly presented by the moderators using signs or in a written form, and the discussion was directly focused on the variations of the signs.

## 5.2 Sign language annotation

Sign language annotation of recorded meetings was performed on the ELAN platform (Crasborn and Sloetjes, 2008)<sup>15</sup> by the authors, who are fluent in Bulgarian sign language. Each participant in the recorded meeting was assigned a separate annotation layer since many participants signed at the same time.

Fig. 2 shows an annotated short excerpt of a recording. At present we have limited the annotation to cover only relevant lexical units (target signs) belonging to the target domains in order to make the annotation process more time-efficient and manageable. In some of the discussions interesting signs typical for the domains have emerged such as names of people, organisations or medical establishments – names of major hospitals, e.g. *Pirogov Multi-profile Active Treatment &*

<sup>15</sup><https://archive.mpi.nl/tla/elan>



Figure 2: A screenshot of ELAN annotation tool

*Emergency University Hospital*, newly established government structures, e.g. *National Operational Headquarters for Combating Covid-19*, or other organisations *World Health Organisation*). These signs have also been annotated and some of them included in the Dictionary.

The sign language material offers many other possibilities for annotation in future studies on Bulgarian sign language lexical system, structure of signs, communication and conversation patterns, etc.

## 5.3 Sign selection

After the recordings had been annotated, all occurrences of the target signs for each domain were automatically extracted and analysed in terms of frequency and variations. From them, the most representative sign variations for each keyword were selected. As representative were considered signs that: (a) were used by more than one signer; and (b) were used on more than one occasion. A single occurrence of a sign in the data does not necessarily mean that the sign is in use since it could be an occasional occurrence, individual invention or copied from a foreign sign language.

In some cases for very similar variations which are not questionable and would be understood by all sign language users (e.g., with slight variation of either handshape, palm orientation, movement or place of performance) only one of the variants was selected, usually the most specific, with the complete motion performed, or the most elaborate and thorough one. For example, the sign for *lekarstvo* – *medicine* can be performed with or without the supporting second hand (that stays in a fixed position with a flat palm up and only serves as a base for performing the sign with the main hand) – the full sign performed with both hands is recorded for the Dictionary while the simplified version is not

included, i.e. it is considered as a non-essential variation based only on simplification.

For some keywords no suitable signs were found in the data when: (a) the Deaf moderators deliberately excluded some keywords from the discussion if the signs were clear, well-established and frequently used in the language; (b) the signs were omitted from the discussion; (c) the participants did not know the sign for a given keyword; (d) the participants were not familiar with the concept under discussion. These words were not included in the Dictionary.

## 6 Structure of the Dictionary and components of the description

The first release of the Bulgarian Sign Language Dictionary of Crisis Management Terminology covers 600 concepts appearing with high frequency in the information regularly released by government officials and news agencies during the COVID pandemic. The entries are both single words and multiword expressions. Although primarily focused on the pandemic, the Dictionary also covers a variety of domains and terminology. In the future, the Dictionary can serve as a model for building language resources in Bulgarian sign language aimed at Deaf signers, sign language learners, interpreters, etc.

Each dictionary entry is supplied with extensive description. As a bilingual dictionary in spoken (verbal) Bulgarian and Bulgarian sign language, the Dictionary is also multimodal – it includes video presentation of the sign component and text description of the verbal component of the translational pairs.

In the description of each entry we also include information about the relation of the spoken word to other words, multilingual translational equivalents, including a translation into American sign language (ASL), text usage examples, etc. Most of the descriptive information of the verbal component is extracted from WordNet automatically. The description of the sign component is compiled manually since so far there are no available electronic and computationally processable resources for Bulgarian sign language, and there are still very limited processing tools for any sign language.

### 6.1 Information from WordNet

The Dictionary entries are linked to WordNet synsets (covering over 90% of the entries). From the Bulgarian wordnet we add the following compo-

nents of the description of the verbal components of the dictionary entry: (a) all synonyms of the identified keyword that appear in the synset; (b) the definition of the concept; (c) translational equivalents in other verbal languages.

Translational equivalents are extracted from various wordnets available through the Extendend Open Multilingual WordNet project (Bond and Foster, 2013)<sup>16</sup>. The wordnets are linked to the Princeton WordNet, and thus to each other and to the Bulgarian WordNet. Translations are provided wherever possible in up to 20 languages.

Moreover, the dataset is linked to one of the ASL online dictionaries – HandSpeak<sup>17</sup>. The mapping to ASL so far has been performed semi-automatically by processing the wordlist of the HandSpeak dictionary and matching it to the English translational equivalents of the Bulgarian word entry. The mapping was then verified manually.

The structure and organisation of the Dictionary allows linking to other languages as well through WordNet, and also to other sign languages through the links to ASL. However, research is still ongoing on mapping ASL to WordNet and to the best of our knowledge no data have been released so far.

### 6.2 Sign language specific features

Each sign is presented as a video recording and is performed by a skilled Deaf sign language user who is fluent in the language but also is experienced in presenting sign language in front of the camera. For each entry we also have recorded variants of the signs. There has been no research on sign formation in Bulgarian sign language. Although we call all signs that corresponding to a given concept 'variants', it is clear that in some cases these are new independent signs, thus we need to consider them as synonyms rather than variants.

Descriptive features (labels for handshape, palm orientation, direction of movement, etc.) of the signs have not been included in the present version of the Dictionary, but are envisaged for future releases.

A special part of the description of each dictionary entry are the relations to other words (in the verbal component) and to other signs (in the signed component). The derivationally related words, or

---

<sup>16</sup><http://compling.hss.ntu.edu.sg/omw/summx.html>

<sup>17</sup><https://www.handspeak.com/>



words that share the same root as the dictionary entry word, are relevant because very often they share the same sign. In particular, this is valid for a root word and its derivatives in other parts of speech.

For the purposes of the Dictionary we extract derivational relations from WordNet. We do not take into account the direction of derivation since it is not represented in WordNet. Derivationally related words often have similar meaning, and are often represented by the same or similar signs in the Bulgarian sign language (e.g., the sign for *bg. bolen – ill* is the same as the sign for *bg. bolest – illness*).

However, special attention should be paid to any exceptions:

- Different signs for derivationally related words with close semantics (e.g., there are different signs for *bg. lekar – doctor, medic* and *bg. lekarstvo – medicine*);
- The same signs for words that are only semantically and not derivationally related (e.g., we have the same signs for *bg. lekar – doctor* and *bg. bolnitsa – hospital*, as well as for *bg. aptekar – pharmacist* and *bg. lekarstvo – medicine*).

Similarly, attention should be paid to the cases where the same or very similar signs are used on semantically distant words. For example, the same sign is used for *bg. lineyka – ambulance* and *bg. politseyska kola – police car* (the sign is based on the siren and flashing lights of both vehicles). These also can cause confusion when used in delivering crucial information during crisis. Usually the disambiguation relies on the articulation of the signer (the signer mouths the word) or an additional sign (e.g., adding the sign for *medical* or *police*).

These irregularities pose a problem to interpreters and language learners, and this is why we consider the information relevant and beneficial to include in the dictionary. Moreover, since the main objective of the dictionary was to ensure the good quality and the high precision of the delivered information during crises, these relations provide a good starting point to investigate further and establish good practices for sign language presentation and interpreting.

### 6.3 Additional information

The additional information comprises:

- links to other lexical resources, most notably the online dictionaries of the Institute for Bulgarian Language where the users can find more information about the word, an alternative definition, as well as to seek information about multiword expressions;
- examples of the use of the word, excerpted from the text corpus of briefings and news articles;
- excerpts from the video recordings were added demonstrating the use of the sign in context. At present these examples apply to a small number of dictionary entries as they required manual processing and selection.

## 7 Online access

The Dictionary is freely available online on the educational platform of the Deaf Studies Institute<sup>18</sup> and is distributed under Creative Commons Attribution 4.0 License.<sup>19</sup>

The Dictionary entries can be listed in two ways – in alphabetic order of the keywords or by domain (see list of domains in section 4.2) for easier access to related terms. As some words are assigned more than one domain, they appear in more than one domain-specific list. A functionality to search by word or phrase is also added on each page of the Dictionary.

Fig. 3 shows the dictionary entry of *bg. bolnitsno otdelenie – hospital ward* with the components of its description.

Under each video of a sign there is a button to confirm or reject the validity of the sign. This feedback functionality can serve as crowdsourcing validation of dictionary entries. No efforts in the direction of the validation, testing sign language users preferences or standardisation have been made so far for the Bulgarian sign language.

## 8 Conclusions and future work

The present paper shows the compilation process of the Dictionary of Bulgarian Sign Language for Crisis Management. The dictionary is suitable to be used by Deaf people, sign language interpreters, learners of sign language, social workers and the

<sup>18</sup><https://study.deafstudiesinstitute.bg/course/view.php?id=8>

<sup>19</sup><https://creativecommons.org/licenses/by/4.0/>

ЖЕСТОВ РЕЧНИК ЗА КРИЗИСНИ СИТУАЦИИ

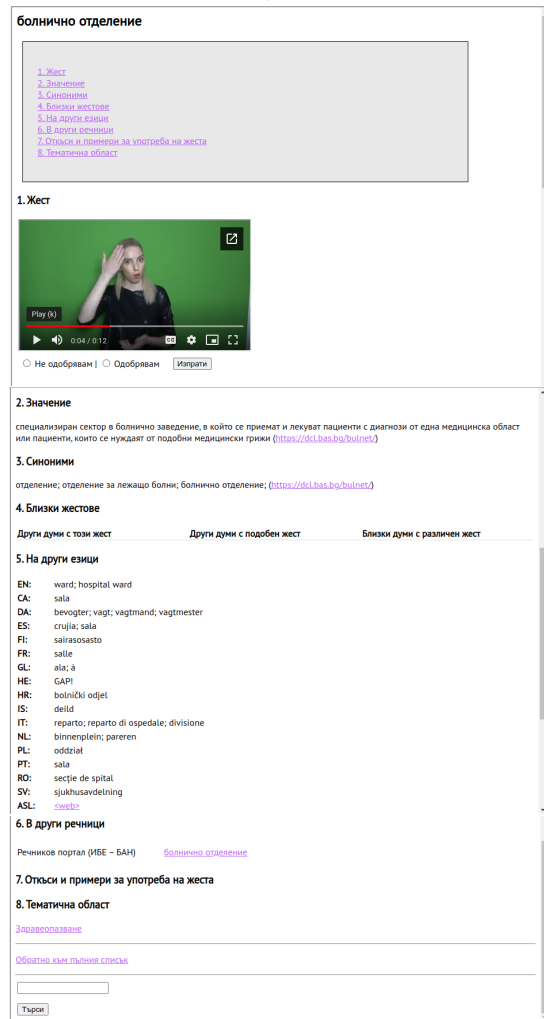


Figure 3: A screenshot of a dictionary entry (1: Sign, with Improve / Disapprove buttons underneath; 2: Definition; 3: Synonyms; 4: Similar signs; 5: Other languages, incl. ASL; 6: Information from other dictionaries; 7: Examples of usage; 8: Domain; Link to the word list; Search field)

wide public. It can accompany educational and information materials focused on crisis management. Although the selection of the concepts is based on a text corpus collected from COVID-related topics, the Dictionary covers six different domains. Moreover, the model of data collection and analysis can be applied to expand the dictionary in volume and in number of domains.

This work is also a first step towards the standardisation of Bulgarian sign language used in time of crisis which requires efficient and unambiguous information. In this respect we need more targeted efforts in collecting user feedback, observations on attitudes towards particular signs, investigating sign ambiguity, etc.

Future work will focus on adding new features to dictionary entries such as textual descriptors of sign components (handshape, palm orientation, motion, etc.). This will allow for searches by sign features (if a sign’s meaning is not known).

An interesting application of the sign language dictionary is in the field of language education for creating interactive materials and linked resources introducing new concepts and supporting the learning of Deaf children. An example of such interactive books for preschool and primary school children is shown on Fig. 4. For this purpose we need to expand the dictionary with more topics and to improve the description of dictionary entries.



Figure 4: Interactive book for language education accessing an online dictionary and its fields (image, sign, definition, etc.)

### Acknowledgments

This paper presents work carried out as part of the project *Efficient Communication for Deaf People in a Crisis* (2020 – 2021) funded by Sofia Municipality under the Crisis as an Opportunity programme (Developing Communities), managed by Sofia Development Association.

### References

Anne Baker. 2016. *The linguistics of sign languages: an introduction*. Amsterdam; Philadelphia: John Benjamins Pub. Company.

Gábor Bella, Fiona McNeill, Rody Gorman, Caoimhin O Donnaille, Kirsty MacDonald, Yamini Chandrashekar, Abed Alhakim Freihat, and Fausto Giunchiglia. 2020. *A major Wordnet for a minority language: Scottish Gaelic*. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2812–2818, Marseille, France. European Language Resources Association.

Francis Bond and Ryan Foster. 2013. *Linking and extending an open multilingual Wordnet*. In *Proceed-*

- ings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics.
- Onne Crasborn and Han Sloetjes. 2008. Enhanced ELAN functionality for sign language corpora. In *Proceedings of LREC 2008, Sixth International Conference on Language Resources and Evaluation*.
- Christiane Fellbaum, editor. 1999. *WordNet: an Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Svetla Koeva. 2010. Bulgarian Wordnet – current state, applications and prospects. *Bulgarian-American Dialogues*, pages 120–132.
- Svetla Koeva, Diana Blagoeva, and Siya Kolkovska. 2010. **Bulgarian National Corpus Project**. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).
- Svetla Koeva and Angel Genov. 2011. Bulgarian Language Processing Chain. In *Proceeding of the Integration of multilingual resources and tools in Web applications Workshop in conjunction with GSCL 2011*. University of Hamburg.
- Gabriele Langer, Susanne König, and Silke Matthes. Compiling a Basic Vocabulary for German Sign Language (DGS) – lexicographic issues with a focus on word senses. In *Proceedings of the XVI EURALEX International Congress: The User in Focus, July 15-19 2014 in Bolzano/Bozen – Italy*, pages 767–786.
- Svetlozara Leseva and Ivelina Stoyanova. 2020. Beyond lexical and semantic resources: Linking wordnet with framenet and enhancing synsets with conceptual frames. In S. Koeva, editor, *Towards a Semantic Network Enriched with a Variety of Semantic Relations*, pages 21–48. Sofia: Professor Marin Drinov Publishing House of BAS.
- Colin Lualdi, Jack Hudson, Christiane Fellbaum, and Noah Buchholz. 2019. **Building ASLNet, a Wordnet for American Sign Language**. In *Proceedings of the 10th Global Wordnet Conference*, pages 315–322, Wroclaw, Poland. Global Wordnet Association.
- Colin Lualdi, Elaine Wright, Jack Hudson, Naomi Caselli, and Christiane Fellbaum. 2021. **Implementing ASLNet v1.0: Progress and plans**. In *Proceedings of the 11th Global Wordnet Conference*, pages 63–72, University of South Africa (UNISA). Global Wordnet Association.
- Balakrishnan S. Manoj and Alexandra Hubenko Baker. 2007. Communication challenges in emergency response. *Communications of the ACM*, 50:51–53.
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to Wordnet: an on-line lexical database. *International journal of lexicography*, 3(4):235–244.
- Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Colin F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California.
- Lei Shi and Rada Mihalcea. 2005. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science*, volume 3406. Springer, Berlin, Heidelberg.
- Yovka Tisheva, Valentina Hristova, Vladimir Zhobov, Gergana Dacheva, Krasimira Aleksova, Petya Angelkova, Tzanka Popzlateva, and Yuliana Stoyanova. 2017. *Dictionary of the Bulgarian Sign Language*. Izkustvo i obrazovanie Publ., Bulgarian Ministry of Education and Science, Sofia. [https://mon.bg/upload/21132/Rechnik\\_bg\\_zhestov\\_ezik.pdf](https://mon.bg/upload/21132/Rechnik_bg_zhestov_ezik.pdf).
- Clayton Valli, Ceil Lucas, Kristin J. Mulrooney, and Miako Villanueva. 2005. *Linguistics of American Sign Language: An Introduction*. Washington, D.C.: Gallaudet University Press.
- Piek Vossen. 2002. WordNet, EuroWordNet and Global WordNet. *Revue Francaise de Linguistique Appliquee*. <https://research.vu.nl/ws/portalfiles/portal/74104438/rfla>.
- Piek Vossen. 2004. EuroWordNet: A Multilingual Database of Autonomous and Language-Specific Wordnets Connected via an Inter-Lingual Index. *International Journal of Lexicography*, 17(3):161–173.
- Elaine Wright. 2021. Signs in the Mind: Constructing ASLNet. <http://arks.princeton.edu/ark:/88435/dsp01hx11xj33r>.

# Raising and Control Constructions in a Bulgarian UD Parsebank of Parliament Sessions

Petya Osenova

Division of Bulgarian Language  
Sofia University “St. Kl. Ohridski”  
osenova@uni-sofia.bg

## Abstract

The paper discusses the raising and control syntactic structures (marked as ‘xcomp’) in a UD parsed corpus of Bulgarian Parliamentary Sessions. The idea is: to investigate the linguistic status of this phenomenon in an automatically parsed corpus, with a focus on verbal constructions of a head and its dependant together with the shared subject; to detect the errors and get insights on how to improve the annotation scheme and the automatic detection of this phenomenon realizations in Bulgarian.

**Keywords:** control and raising verbs, Bulgarian Parliamentary Corpus, Universal Dependencies.

## 1 Introduction

In the Universal Dependencies (UD) syntactic guidelines the dependency relation *xcomp* is viewed as a clause that belongs to the group of core arguments together with *csubj* and *ccomp*. It is used in two cases: a) in constructions with obligatory control (object-to-subject and subject-to-subject) and usually non-finite (for example, in the sentence ‘I want to sleep’, the non-overt subject of ‘sleep’ is determined by the overt subject ‘I’ of the higher predicate ‘want’), and b) for the respective types of secondary predication (for example, in the sentence ‘She declared the cake beautiful’ the predicates ‘declared’ and ‘beautiful’ are connected through *xcomp*). In this survey I am interested in the open clausal complements only, i.e. ‘a predicative or clausal complement without its own subject’. As the guidelines further say: ‘That is, there should be no available interpretation where the subject of the lower clause may be distinct from the specified role of the upper clause. In cases where the missing subject may or must be distinct from a fixed role in the higher clause, *ccomp* should be used instead [...]’. This includes cases of arbitrary subjects and

anaphoric control.’<sup>1</sup>

The aim in this paper is to observe the *xcomp* types of subject-to-subject control structures in an automatically parsed parliamentary corpus for Bulgarian. I am interested in the following questions: a) what kind of control syntactic structures were realized with respect to a main and a controlled predicate; b) what kind of subjects were realized in the control structures – both formally and semantically; c) were any error types detected within the observed structures; d) how do these observations contribute to the linguistic typology of Bulgarian control structures and to their better modeling and detection. I consider the linguistic investigations over parsebanks as a way to identifying real language problematic phenomena for parsing beyond the already modeled constructions in grammars, annotation schemes and manually annotated treebanks. I also believe that they give us hints on how to improve the coverage of a treebank (for example, through the means of active learning) for better linguistic research.

The paper is structured as follows: in the next section the details on the parsed corpus as well as on the used model are given. Section 3 focuses on the relation *xcomp* with respect to the above mentioned research questions. Section 4 concludes the paper.

## 2 The UD parsebank of Bulgarian Parliamentary sessions

This study was performed over the Bulgarian ParlaMint corpus<sup>2</sup> because it has been annotated with respect to the UD schema and is freely available for research. In future, the plan is to extend the texts in the parsebank with newsmedia and social

<sup>1</sup><https://universaldependencies.org/u/dep/xcomp.html>

<sup>2</sup><https://www.clarin.si/repository/xmlui/handle/11356/1431>



media corpora, among others.

ParlaMint<sup>3</sup> is a project supported by CLARIN-ERIC<sup>4</sup>. Its first phase - ParlaMint I - was completed in the period of years 2020 - 2021. Parliamentary data directly correspond to the most recent events with global impact on human health, social life and economics such as the current COVID-19 pandemic. The Bulgarian ParlaMint corpus contains plenary meetings from 2014-10-27 to 2020-07-31 and includes 717 documents, or 19,096,761 words. The data is publicly available from the project website. Now in the subsequent project phase - ParlaMint II (2022 - 2023) - more data have been compiled to the current corpora, and parliamentary corpora for new countries have been added.

The Bulgarian Parliamentary data was downloaded from the official website of the Bulgarian National Assembly<sup>5</sup>. The sessions for each day were represented in a single html file which was relatively easy to convert to XML. The conversion was performed in an incremental way. Initially, the data was converted into a basic TEI XML format and then uploaded into the CLaRK system — (Simov et al., 2004). Afterwards, the Parla-CLARIN format<sup>6</sup> was used for validation. However, this turned out to be too permissive, so an additional constraint schemata were applied. Within CLaRK system the conversion was done with the help of constraints (as implemented rules) and regular grammars for inserting some elements. The speaker information (such as date and year of birth, occupation, party memberships, personal web page, etc.) and incident data (such as applause, laughing, entering or leaving the plenary room, noise, etc.) were extracted, classified and returned back into the texts with the appropriate features added. Thus the present linguistic research can be extended in future with adding more society-oriented features from the available metadata – like which member of Parliament uses what control constructions and with what a reference, etc.

The created corpora were processed with the `classla-stanfordnlp` pipeline, which annotates text on the levels of morphosyntax, lemmas, dependency syntax and named entities for Bulgar-

ian, Croatian, Serbian, and Slovene.<sup>7</sup> This model is a CLASSLA Fork of the Official Stanford NLP Python Library for Many Human Languages. The Bulgarian part was trained with the UD Bultreebank model and on the provided big corpus of Bulgarian data. The resulting analyzed corpus of parliamentary sessions was uploaded into the CLaRK System where it was possible to search for respective subtrees related via *xcomp* within the UD syntactic structures. The extracted patterns include the control verb, the dependant verb and the subjects when they are explicit at the higher or lower verb level (although in *xcomp* constructions an explicit subject at the lower clause is not expected). In Figure 1 an example in XML of an extracted pattern is given from the CLaRK system. The sentence is as follows: But not can-1.PL to give-1.PL more money, ‘However, we cannot give more money either’. The *xcomp* relation connects the verb in the higher clause - ‘can’ - with the one in the lower clause - ‘give’. Both subjects are not overt.

In Figure 2 three examples are graphically visualized where the head and dependant verbs are related through *xcomp*.

In the tree on the top-left the following sentence is given (here glossed, and all that follow are also glossed): *Can-2.PL to check-2.PL (You can check)*. In this subject-to-subject control both subjects are null since Bulgarian is a pro-drop language. We consider this structure as a true control one because the subject of the verb in the lower clause - ‘check’ - is the same as the one of the verb in the higher clause - ‘can’.

In the tree on the top-right the following sentence is given: *Raynov will come to them. CLITIC take (Raynov will come to take them)*. Here the main verb ‘come’ has an explicit subject – the surname Raynov – in contrast to its dependant verb ‘take’. I do not consider such a structure a true control one, since the verb ‘come’ can take dependant verbs with a different subject. One test that can be used here is the possible substitution of the marker да (to) with the subordinator за да (‘for to’, in order to). In the example the subjects of the two verbs are the same. We would like to have a way to distinguish such cases in parsebanks.

In the tree in the bottom-middle, the following sentence is given: *How would could to happen this? (How could this happen?)*. Here the explicit subject is realized to the dependant verb ‘happen’

<sup>3</sup><https://www.clarin.eu/parlamint>

<sup>4</sup><https://www.clarin.eu/>

<sup>5</sup><https://www.parliament.bg/bg/plenaryst>

<sup>6</sup><https://github.com/clarin-eric/parla-clarin>

<sup>7</sup><https://pypi.org/project/classla/>

```

s :: :
├─ linkGrp :
│   ├─ link : seg220.5.1: Но      : seg220.5.3 : cc
│   ├─ link : seg220.5.2: не      : seg220.5.3 : advmod
│   ├─ link : seg220.5.3: можен   : seg220.5   : root
│   ├─ link : seg220.5.4: да      : seg220.5.5 : aux
│   ├─ link : seg220.5.5: дадем   : seg220.5.3 : xcomp
│   ├─ link : seg220.5.6: и       : seg220.5.8 : cc
│   ├─ link : seg220.5.7: повече  : seg220.5.8 : advmod
│   ├─ link : seg220.5.8: пари    : seg220.5.5 : obj
│   └─ link : seg220.5.9: .       : seg220.5.3 : punct

```

Figure 1: An extracted pattern from the CLaRK system.

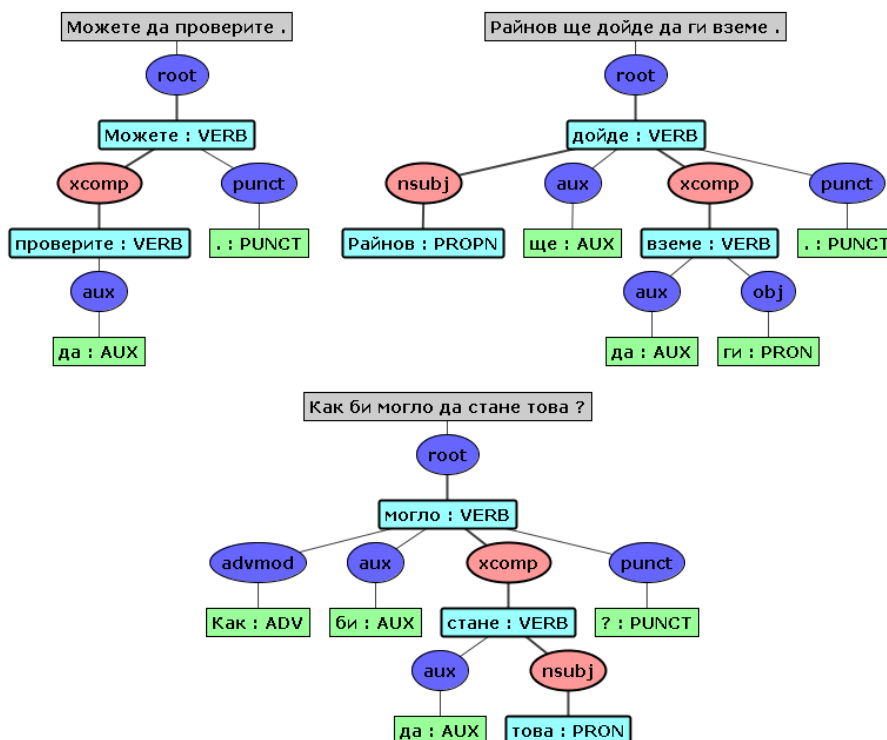


Figure 2: Visualized patterns with the *xcomp* relation.

in contrast to the main modal verb. However, here many other factors play a role. For example, the adjacency of the pronominal subject either to the main or to the dependant verb with respect to the illocutionary force - interrogative in this case. I view such patterns as formally controlling.

All the patterns presented here were used as templates in the process of extracting a subcorpus for the current study.

### 3 The *xcomp* realisations and their analysis

The control verbs are usually discussed on a par with the raising ones. The literature on control

and raising verbs from a theoretical or a specific language point of view is very rich and sometimes controversial. For that reason I will mention now only the work on control and semantic resource sensitivity by (Asudeh, 2005). The author gives an overview of the main approaches and proposes a structure sharing alternative for both – non-finite and finite control. The analysis is based on Glue Semantics and is performed within the framework of LFG.

In the original constituency Bultreebank (later converted into the UD style), the control structures were not specially marked as such. There was a mechanism to indicate the same subject in the syn-

tactic structures via co-reference links. However, these links reflected the contextual usages of same-subject-hood, not the real control. Thus, they can be viewed as overgenerating. This means that no real distinction was made between structures of control where the predicate imposes on its dependent the same subject in all contexts, and structures where the same subject is not obligatory and thus would allow the appearance of different subjects. Making such a differentiation is not a trivial task per se. At the same time, the fact that raising verbs do not impose any restrictions to their subjects (expletive as a rule) has been reflected by assigning the referential subject to the lower clause verb.

### 3.1 Structures of control in Bulgarian: a brief overview

In the traditional Bulgarian grammar literature the control verbs are viewed as imposing argument sharing. These verbs are modal (with some exceptions) or phasal. They are considered auxiliaries and thus constitute the so-called ‘complex verbal predicate’ forming a simple sentence where both verbal subjects are co-indexed. See an overview of the various points of view in (Viktorova, 2005). These verbs are: (мога (*can*), трябва (*have to*), започвам (*start*), продължавам (*continue*), спирам (*stop*)) with their synonyms. The exceptions include the verb искам (*want*) because it can take various subjects.

Among the modals there exist also raising verbs such as the impersonal verbs with expletive subjects like трябва (*have to*) and може (*to be possible to*).

In cases where the modal verb allows for a different subject of the dependent verb, the sentence is considered not simple but complex. Such a verb, as mentioned above, is искам (*want*). Compare Искам ти да дойдеш. (*Want-I you to come, I want you to come.*)

(Penchev, 1993) mentions the control structures of types subject-to-subject (p. 169) and object-to-subject (p. 87, p. 169). For the first type the example is Ти<sub>1</sub> забрави про<sub>1</sub> да дойдеш. (*You<sub>1</sub> forgot pro<sub>1</sub> to come.*) For the second type the example is Принудиха ги<sub>1</sub> про<sub>1</sub> да заминат. (*Forced-they them to go, They were forced to go.*)

In (Boyadjiev et al., 1998) (pp. 550-551) Penchev also shows that control is not related to modality only, since some modal verbs behave like content verbs while there are also non-modals

that exhibit control characteristics. The author promotes a unifying analysis where both control structures – with modals and non-modals – form a complex sentence.

### 3.2 Realisations of control structures in the corpus

First, let us look at the heads of the control structures and their frequency. The most frequent one is the modal verb мога (*can*) with 47514 occurrences. In the UD version of Bultreebank modals were treated as full verbs, not as auxiliaries.

In the top 20 lemmas the following types have been observed: other modal verbs ща (*want*); verbs of phases (продължа/продължавам (*continue*), започна/започвам (*start, begin*), спира/спирам и престана (*stop*)); other verbs (успеея/успявам (*succeed*), опитам се/опитвам се (*try*), пропуска/пропускам (*miss*), отида/отивам (*go somewhere*), откажа/отказвам (*deny*)). Also in the top part come other modal or modal-like verbs like: стремя се (*aim*), възнамерявам (*intend*), умея (*be able*), смея (*dare*).

At the same time some verbs seem to be out of place here because they either express adverbial semantics or allow a non-controlled subject. Such verbs are: изляза/излизам (*go out*), бързам (*hurry*) with adverbial semantics and thus the expected relation would be *advcl* or призова/призовавам (*call for*), предлага/предлагам (*suggest*) and thus the expected relation would be *ccomp*. This fact is not surprising because – as mentioned above – such verbs could also share the same subject in some of their realizations.

Let me now turn to the structures with controlling and controlled predicates. I am interested in three questions: a) which are the typical controlling predicates, b) which are the structures that are not really controlling and c) which are the linguistic tests that show the non-controlling usages of the detected verbs in b).

Concerning the modal verbs, the most frequent structure is мога да кажа (*can-I.SG to say-I.SG, I can say*). It has 2230 occurrences. Overall, the perfective verbs are preferred: мога да разбера / приема / дам / направя (*can-I.SG to understand-I.SG / accept-I.SG / give-I.SG / do-I.SG, I can understand/accept/give/do*). One remark should be done here. The third person of the verb can have also an impersonal usage, i.e. mean-

ing that something is possible. Such cases of two possible readings for convenience were annotated in Bultreebank as preferably personal verbs. Thus, many of the examples in the parliamentary corpus also bear this inherited ambiguity.

### 3.3 ‘True’ control verbs

Here come the ‘true’ control verbs, or in other words, verbs that would not allow for a different subject of the lower clause verb. Apart from the modal, phase and other verbs, mentioned above, some other verbs are listed below. Please note that some of them are used in their reflexive forms. The semantic classification is made with respect to the lexicographic classes in Princeton WordNet (in contrast to (Henri and Laurens, 2011) where another type of semantic classification is given for Mauritian):

- verbs of cognition: уча се (learn), пропуска/пропускам (miss)
- social verbs: опитам се/опитвам се (*try*), принудя се/принуждавам се (*force*), задължа се/задължавам се (*oblige*), рискувам (*risk*)
- verbs of change: готвя се (*prepare*)
- verbs of communication: откажа/отказвам (*refuse*)

It would be interesting to investigate further the relation between control structures and reflexivity. In general, the reflexive marker *се* ‘se’ ensures the intransitive use (thus – subject-to-subject control) of a transitive verb that provides an object-to-subject control. For example, Учих го да чете (*Taught-1.SG him to read-3.MASC.SG, I taught him to read*) vs. Учих се да чета (*Taught-1.SG REFL to read-1.SG, I taught myself to read*).

Some insights with respect to the usage and blocking of impersonal and passive se-constructions have been considered in (Penchev, 2001). For example, when a reflexive control verb is used in an impersonal-passive, then either such an usage is semantically blocked (ex. try) or its dependant has to share the same form, and the subject becomes arbitrary (ex. forget): Забравя се да се звъни (*Forget-IMPERS.REFL to REFL ring-IMPERS, Ringing is (being) forgotten*). It should be noted that such usages are rare.

Another issue that became evident is the role of diathesis. It can be detected in the examples of

the verb задължа се (*oblige oneself*). In all the examples these usages are in reflexive passive. Let us see one: Общината се задължава да извърши проверка (*Municipality-the REFL oblige to perform check, The municipality is obliged to perform the checks*). Such cases are also considered control structures – not from a lexical but from a syntactic point of view. The role of the reflexive passives is discussed in (Dzhonova and Mihaylova, 2021) where it is mentioned that these forms can have modal meanings when used in a generic way. The reflexive passives can be placed also in the diathesis typology, presented in (Koeva, 2022).

Here it would be also interesting to observe the combinations of a control verb with types of dependant verbs as well as their common subject characteristics.

The modal verb мога (*can*) as the most frequent one has many collocations, thus we will ignore it here. In the group of the phase verb започна/започвам (*start*) the following clusters can be identified: започвам да функционирам (*start functioning*) where the dependant verbs are in active voice and subjects refer to the government, software, assembly, law, portal; започвам да тека (*start to run*) where the dependant verbs are in active voice and subjects refer to mandate, process, deadline, intership; започвам да работя (*start to work*) where the dependant verbs are in active voice and subjects refer to institution, system, assembly, power. There are also structures where the dependant verb is preferred in se-passive. Here are some examples: започва да се прави компромис/реформа; започва да се гледа бюджет/закон; започва да се говори истина/неистина (*start to REFL do compromise/reform; start to REFL look budget/law; start to REFL speak truth/non-truth*).

In the group of the phase verbs продължа/продължавам (*continue*) the following clusters can be identified: продължавам да действам (*continue to hold/be in effect*) where the dependant verbs are in active voice and subjects refer to rule, practice, formula, criterion; продължавам да съществувам (*continue to exist*) where the dependant verbs are in active voice and refer to nation, threat, problem, inequality, tension, possibility.

The cognitive verb пропуска/пропускам (*miss*) has a preference to speech-related active dependant verbs like say, note, mention, remind, give an answer.



The social verb *принудя се/принуждавам се* (*force oneself*) prefers dependant verbs of activities like ‘*to be forced to come (for a prime-minister); to co-finance (for a municipality); to resort to (for the state)*’.

It turned out that the control verbs other than modal and phase ones are not so frequent in the data.

On the basis of the statistical information about the distribution of these constructions - the combination of the head verbs, the dependent verbs and the grammatical features of the subjects, rules can be formulated to classify the candidate control structures. These are based on grammar characteristics such as shared number and gender where applicable. Then manual evaluation over 3951 examples was performed. From these 3100 were classified as control structures while only 5 cases happened to be misclassified. From the rest there were 651 cases which were classified as structures with non-shared subjects, and 200 that were considered as quasi control structures presented in the next section.

### 3.4 Quasi control verbs

Some examples were given above with verbs that can take not only the inherited infinitive particle *да (to)*, but also the subordinator *за да (in order to)*. This fact can be used as a test for classifying such verbs as quasi control ones because it allows a structure with different subjects. This group mostly consists of verbs of action. For example, *дойда да гласувам (come to vote)*, *излизам/отивам да говоря (go to speak)*, *чакам да видя (wait to see)*, *работя да осигуря (work to ensure)*, etc.

There is one verb that is ambiguous between a control and quasi control interpretation. This is *спра/спирам (stop)*. In the first meaning – the phase one – it is a verb of control: *Спях да пуша (Stopped-I to smoke, I quitted smoking)*. In the second meaning – the action verb – it is a verb of quasi control: *Спях да купя мляко (Stopped-I to buy milk, I stopped to buy milk)*. In the parliamentary data only the phase verb has been detected.

There is another group of quasi control verbs that allows for the dependant verb to take a subject in a different number. These verbs belong preferably to the groups of verbs of communication and cognition. For example: *предложа/предлагам (suggest)*, *ангажирам се (engage oneself)*, *апелирам (apel)*, *избера/избирам (choose)*, *плани-*

*рам (plan)*. For example, *Предлагам да дойдем по-късно, Suggest-1.SG to come-1.PL later, I would suggest we to come later*.

As a result from these observations, a number of tests were created for the classification of control vs. quasi control usages like the one with the subordinator substitution, and some based on the lexical properties of the verbs like their valency and agreement potential. In addition to using them as features when training parsing models, such tests might be implemented as filters over the search in parsebanks.

## 4 Conclusions

In this paper some focused observations were shown on the behaviour of Bulgarian structures of raising/control in an automatically parsed UD corpus of parliamentary sessions. The manual checks over the extracted data confirmed the high quality of the UD parser on these data. Thus, it became possible to detect for example the ‘true’ control structures vs. quasi control structures. The over-generation seems to be inherited from the Bultreebank model where all cases of shared subjects were marked as coindexed. Due to the distinction between active (*nsubj*) and passive (*nsubjpass*) subjects in the UD schema, it was possible to survey the internal structure of control and observe the preferences of dependant predicates with respect to their control heads to active or passive usages.

One of my goals in this study was also to detect weaknesses in the Bulgarian UD treebank which needs some extensions of the annotation patterns in order to provide better parsed corpora for linguistic research. I think that these analyses of control constructions in the current version of the corpus show the following directions of future work: extension of the treebank coverage with new texts that would demonstrate some of the problematic cases for the parser.

My observations showed that it is difficult to distinguish between similarly presented phenomena in texts, such as control and quasi control structures. These phenomena might be approached by using lexical lists with both types of verbs. However, this is not enough because their contextual realizations also have to be taken into account. In my view the challenge behind the automatic annotation is to find the best balance between lexicon and grammar. If such a balance was achieved, then the parser would be more linguistically informed and would classify

the presented phenomena in a better way.

## References

- Ash Asudeh. 2005. *Control and semantic resource sensitivity*. *Linguistics*, 41. Cambridge University Press.
- Todor Boyadjiev, Ivan Kutsarov, and Yordan Penchev. 1998. *Contemporary Bulgarian*, name of chapter: Syntax. Petar Beron Publishing House, Sofia.
- Marina Dzhonova and Bilyana Mihaylova. 2021. Reflexive passive in Bulgarian and Romanian - forms and uses. *Contrastive Linguistics*, 2-3:25–36. St. Kliment Ohridski University Press.
- Fabiola Henri and Frédéric Laurens. 2011. The complementation of raising and control verbs in mauritian. In O. Bonami and P. Cabredo Hofherr, editors, *Empirical Issues in Syntax and Semantics*, 8, pages 195—219. [Http://www.cssp.cnrs.fr/eiss8](http://www.cssp.cnrs.fr/eiss8).
- Svetla Koeva. 2022. System of Diatheses in Bulgarian. *Proceedings of the International Annual Conference of the Institute for Bulgarian Language*, pages 80–91. Prof. Marin Drinov Publishing House of Bulgarian Academy of Sciences.
- Yordan Penchev. 1993. *Bulgarian Syntax: Government and Binding*. Plovdiv University Press, Plovdiv.
- Yordan Penchev. 2001. One component sentences. In Svetla Koeva, editor, *Contemporary Linguistic Theories*, pages 86–93. Plovdiv University Publishing House, Sofia.
- Kiril Simov, Alexander Simov, Hristo Ganev, Krasimira Ivanova, and Ilko Grigorov. 2004. *The CLaRK System: XML-based Corpora Development System for Rapid Prototyping*. *Proceedings of LREC 2004*, pages 235–238.
- Kalina Viktorova. 2005. Functional development of da-construction in contemporary bulgarian. In Svetla Koeva, editor, *Argument structure: Problems of the simple and the complex sentence*, pages 185–224. SEMA RSH, Sofia.

# Syntactic characteristics of emotive predicates in Bulgarian A corpus-based study

Yovka Tisheva

Sofia University “St. Kliment Ohridski”  
tisheva@uni-sofia.bg

Marina Dzhonova

Sofia University “St. Kliment Ohridski”  
djonova@slav.uni-sofia.bg

## Abstract

The paper presents a corpus-based study of emotive predicates (verbs and predicative constructions with adjectival, adverbial or noun phrases) in Bulgarian with respect to their syntactic characteristics. The sources of empirical data analyzed here are Bulgarian National Corpus, Corpus of Bulgarian Political and Journalistic Speech and Bulgarian part of Multilingual Comparable Corpora of Parliamentary Debates ParlaMint. The analyzes are organized in terms of morpho-syntactic features of emotive predicates, transitivity, syntactic functions and theta-roles of their arguments. Emotive predicates denote a state or an event involving an affective experience. As part of the special semantic class of psychological/Experiencer verbs, they have been studied in relation to the interaction between lexical semantics and argument realization. Bulgarian data confirm the well-established division of Psych predicates into three classes: Subject Experiencer (*fear* type verbs), Object Experiencer (*frighten* type verbs), Dative Experiencer. The third class is mostly represented by adverbial predicates.

**Keywords:** Psychological predicates, Emotive predicates, Experiencer, Argument structure

## 1 Introduction

The main topic of this study is the syntactic realization of arguments to verbal predicates and predicative constructions in Bulgarian expressing positive emotions. The analysis will not be restricted to emotive verbs only, but will represent adjectives, adverbs, or nouns used in constructions which meaning corresponds to the category of the

positive emotions. The objectives of empirical data analyzes are to compare the syntactic structure of two types of sentences - with verbal or with adjectival, adverbial or nominal predicates. The focus of our observations is related to the question whether the argument structure of emotional verbs is "inherited" by the corresponding adjectives, adverbs or nouns. Special attention will be paid to the syntactic realization of the central participant in the emotional scenario marked by the semantic role of experiencer.

The sources of empirical data analyzed in this paper are Bulgarian National Corpus (<http://dcl.bas.bg/bulnc/>; Koeva et al., 2012), Corpus of Bulgarian Political and Journalistic Speech (<http://political.webclark.org>; Osenova and Simov, 2012) and Bulgarian part of Multilingual Comparable Corpora of Parliamentary Debates ParlaMint (<https://www.clarin.eu/resource-families/parliamentary-corpora>; Erjavec et al., 2022). In this article we provide statistic data only from Bulgarian National Corpus. The observations are organized in terms of morpho-syntactic features of emotive predicates, transitivity, syntactic functions and theta-roles of their arguments. First, the structure of sentences with emotive verbs *veselya* (rejoice), *zabavlyavam* (entertain), *radvam* (make someone happy; glad), and their reflexive counterparts *veselya se*, *zabavlyavam se*, *radvam se* will be discussed. Then the results of analyses will be compared with the features of sentences with adjectival, adverbial and nominal constructions with *vesel* (joyful), *zabaven* (amusing), *radosten* (joyful; happy); *veselo* (joyfully), *zabavno* (funny), *radostno* (happily); *veselba* (merriment), *zabava* (entertainment), *radost* (joy). The choice of these particular lexemes is motivated by the fact that two verbs and not just one signify the feeling, as is the case with *strahuvam se* (fear). On the other hand, the group of emotive predicates includes adjectives, adverbs, and nouns corresponding to the verbs of emotion.

Verbs like *plasha* (frighten), *strahuvam se* (fear), *valnuvam* (excite someone) or *valnuvam se* (get excited) have no corresponding adjectives.

## 2 Emotional scenario

Emotions are mental processes reflecting the experiences, perceptions, and evaluations associated with a particular object or specific stimulus. According to [Wierzbicka](#) (1999), all natural languages have lexical means for expressing conceptualized notions of emotional states, evaluations and attitudes. Lexical semantics of the elements from the emotional lexicon provides the relational and semantic frameworks for syntactic structures used to denote different types of emotions.

Apart from subject who can feel or sense something (experiencer), an element of evaluation is present in the emotional scenario. For the predicates under consideration in our work, it is an evaluation of what is happening by the experiencer as something positive for him or her. This evaluation, in turn, is a stimulus for the positive emotion; stimulus affects the experiencer, changing or maintaining his/her emotions. This general scenario specifies the possible syntactic structures of the sentences with emotive predicates. Causative verbs like *veselya* (rejoice), *zabavlyavam* (entertain), *radvam* (make someone happy) are two-argument predicates. The stimulus (cause) and the experiencer must be presented in the sentence. With reflexives *veselya se*, *zabavlyavam se*, *radvam se* only one element of the emotional scenario is necessary to be expressed. Since the emotion is conceptualized and separated from its stimulus this argument will represent the experiencer.

## 3 Psychological (Experiencer) verbs

Emotive predicates are part of a larger group of predicates called mental predicates, affective verbs ([Belletti and Rizzi](#), 1988), psychological verbs (psych-verbs; [Levin](#), 1993), experiencer verbs (experiencer verbs; [Pesetsky](#), 1995). Psych verbs are a class of verbs defined not only by their lexical semantics, but also by the semantic properties of the sentences they function in. As [Belletti and Rizzi](#) (1988) first stated, “verbs expressing psychological states have a uniform  $\theta$ -grid, involving an EXPERIENCER, the individual experiencing the mental state, and a THEME, the

content or object of the mental state” ([Belletti and Rizzi](#), 1988: 291). The second role is more often called stimulus.

Three subtypes of psych verbs are defined based on their lexical semantics: verbs of perception (*see*, *hear*), verbs of cognition (*know*) and verbs of emotion (*fear*, *frighten*). Emotive predicates, on the other hand, “fall into two grammatically distinct classes: those whose subject is the animate Experiencer and whose object (if there is one) is the Source (*fear*, *miss*, *adore*, *love*, *despise*); and those whose object is the animate Experiencer and whose subject is the Source (*amuse*, *charm*, *encourage*, *anger*)” ([Fellbaum](#), 1999: 297).

Most of the emotive verbs in Bulgarian can be used with short reflexive pronoun *se* (self), e.g. *radvam – radvam se*, *plasha – plasha se*. In this case, *se* is marker for middle voice construction and does not indicate reflexiveness (cf. [Asenova and Guentchéva](#) 2022), it occupies the direct object position and those verbs could have only PP or a complement clause as their second argument. In these cases, the difference between verb groups (fear-type with subject experiencer and frighten-type with object experiencer) is also marked by the use of short reflexive pronoun *se*.

## 4 Types of verbal constructions

### 4.1. Transitive constructions

Verbal expressions with psych transitive verbs *radvam*, *zabavlyavam*, *veselya* display similarities in their argument structure and realizations of experiencer and stimulus of emotion. Usually, both arguments are expressed. NPs in subject position display the features of stimulus (rather than an effector or pseudoagent). Subject may be either animate or inanimate. If the stimulus is animate, it may get agent-like interpretation; if it is inanimate, it will be source of the emotion.

Subject is explicitly expressed mainly by a nominal phrase whose referent is a person. If inanimate nouns with specific reference (object or proposition) are used, they generally denote the result of a person's activity by which an emotional impact is achieved. It is also possible subject to be expressed by nominalizations. The only difference in syntactic patterns concerns the use of complement clauses. *Radvam* and *zabavlyavam* allow complement clauses with *che*,

*da, kak, deto*, while *veselya* can have only NPs in subject position.

*Радва ни, че си оценил нашата търпимост.*

*We are glad you appreciated our tolerance.*

*Близко три часа групите "Сигнал" и Б. Т. Р. веселяха гостите.*

*For nearly three hours, the groups Signal and B. T. R. entertained the guests.*

*Radvam, zabavlyavam* and *veselya* are (direct; accusative) object-experiencer verbs. Our observations show more limited possibilities for syntactic representation of this argument. Experiencer argument of *veselya* is expressed by NPs denoting an animate object in 77 occurrences and by a pronoun in 35. In comparison, there are 493 occurrences of *radvam* with NP denoting animate object vs. 176 with a pronoun in object position.

If this argument is inanimate the examples can be interpreted as metonymic or metaphorical transfer (*syrceto* ‘heart’, *ochite* ‘eyes’, *dushata* ‘soul’).

*Съзнанието за това не веселеше сърцата им както преди.*

*The consciousness of it did not rejoice their hearts as before.*

No complement clauses are allowed in object position. Another essential feature of these verbs is that experiencer is always explicit. There are no examples with implicit (null) experiencer.

Along with the nominal phrases representing the experiencer and the stimulus, a prepositional phrase with *s* can also be part of the sentences with *radvam, zabavlyavam* and *veselya*. PPs introduce a means, most often with a specific referential interpretation, by which the animated stimulus achieves the effect on the experiencer. The PP is an adjunct of the predicate, always instrumental and non-animate.

*Безобидните артисти, които радват народа с уменията си.*

*The innocent artists who entertain the people with their skills.*

Our observations are represented briefly in the following table.

	stimulus	adjunct
<i>radvam</i>	NP or <i>che, da, kak,</i>	s-PP
<i>zabavlyavam</i>	<i>deto</i> complement clause	
<i>veselya</i>	NP	

Table 1: Object-experiencer verbs

The corpus data confirms those properties of object-experiencer verbs. The corpus data statistics shows interesting results in respect to the frequency of each type of complement clause. For the verb *radvam* we have 43 examples with *che*-complement clause vs. 9 examples with *da*-complement clause. *Deto* as a complementizer has no occurrences with object-experiencer verbs in corpus data. This result for *deto* is expected due to its colloquial status in contemporary Bulgarian. For *zabavlyavam* we observe almost equal number of occurrences in respect to the complementizers: 9 examples with *da* and 7 examples with *che*.

The corpus data confirms our hypothesis concerning the adjunct s-PP, which are always instrumental and non-animate.

#### 4.1 Intransitive constructions

*Radvam se, zabavlyavam se* and *veselya se* are subject-experiencer psych verbs. As pro-drop language, Bulgarian allows subject position to be empty. If subject is explicit, syntactic realizations of experiencer include nominal phrases only. There are no examples with complement clauses in subject position.

The intransitive verbs are formally reflexive. Stimulus of the emotion can be syntactically unexpressed. If this element of the emotional scenario is also expressed, a prepositional phrase with *s* or subordinate clauses with *che, da, kak* denote the instrument, effector or situation evaluated by the subject experiencer. *Radvam se* takes these subordinated clauses as complements. The subordinate clause alters with an argument PP with *na* or *za*. On the other hand, *zabavlyavam se* and *veselya se* could have only s-PP in adjunct

position. *Zabavlyavam se* allows also an adjunct instrumental clause with *che, da, kak*.

	stimulus	adjunct
<i>radvam se</i>	<i>na-PP, za-PP che, da, kak</i>	
<i>zabavlyavam se</i>		s-PP <i>che, da, kak</i>
<i>veselya se</i>		s-PP

Table 2: Subject-experiencer verbs

The corpus data shows prevalence of *che*-clauses with *radvam se* – 8336 vs. 3996 occurrences with *da*-clauses. We found very few examples with *deto* as a complementizer – only 24, and even less with *kak* – 8 occurrences.

As the subordinate clause is an adjunct for *zabavlyavam se*, we found much less examples, most of them with *da*-clauses – 292 occurrences vs. only 20 with *che*-clauses. The hypothesis that *kak* and *deto* could also introduce the subordinate clause is not strongly supported by corpus data – we found only one example with *deto* as a subordinator.

No examples with clausal stimulus to *veselya se* were found in the data.

Concerning the adjunct PPs, the corpus data shows predominance of the committative PP with animate noun (40 examples for *zabavlyavam se*) comparing to the instrumental PP (18 examples for *zabavlyavam se*).

## 5 Types of constructions with adjectives, adverbs, or nouns

The constructions whose meaning correspond to the meaning of the verbs for positive emotions denote an emotional state. They have the same argument structure as the verbs of emotion – the experiencer and the stimulus.

### 5.1. Constructions with subject experiencer

The first type of constructions form by an adjective and an auxiliary verb: *radosten sam, vesel sam*. The experiencer argument is obligatory, though it is not always explicit. These two constructions show differences in respect to the realization of the second argument. The stimulus argument for *radosten sam* is PP with *za* or *na*, or a complement clause with *che, da, kak, deto*. As for *vesel sam*, it could only have a complement clause with *che* as stimulus argument.

	stimulus
<i>radosten sam</i>	<i>za/na-PP</i> or <i>che, da, kak, deto</i> complement clause
<i>vesel sam</i>	<i>che</i> complement clause

Table 3: Subject-experiencer constructions

The corpus data shows for *radosten sam* the same tendency as shown for *radvam se* for the predominance of *che*-complement clauses – 145 vs. 60 occurrences with *da*-clause. The data confirms the possibility for *vesel sam* to have *che*-complement clause, but those examples are very rare – we found only two. Concerning *deto*-clauses, we found only one example for each construction.

### 5.2. Constructions with dative experiencer

The respective constructions with dative experiencer are *radostno mi e, veselo mi e, zabavno mi e*. They can only have a complement clause with *che* or *da* as a stimulus argument. With *radostno mi e, veselo mi e* we also found complement clauses with *deto*, while *zabavno mi e* can have a complement clause with *kak*.

	stimulus
<i>radostno mi e</i>	<i>che, da, deto</i> complement clause
<i>veselo mi e</i>	clause
<i>zabavno mi e</i>	<i>che, da, kak</i> complement clause

Table 4: Dative-experiencer constructions

The corpus data shows very few examples for those two constructions with a complement clause – 3 examples for *che*-clauses with *radostno mi e* and for *veselo mi e*, 13 with *zabavno mi e*. *Da*-complement clauses are also very rare: 5 with *radostno mi e*, 12 with *veselo mi e*. With *zabavno mi e* we have much more examples with *da*-complement clause – 104.

### 5.2. Constructions with implicit experiencer

There are also two types of constructions denoting emotion, but with an implicit, generic experiencer. The first of them corresponds to the constructions with dative experiencer – *radostno e, veselo e, zabavno e*. We analyze them separately due to the fact they show differences in respect to the stimulus argument. It could be a complement clause with *che* or *da* (for *veselo e* – only with *da*) or a nominalization – an NP in subject position. In



both cases, a PP with *za* could appear in order to specify the generic experiencer. This is also true for the second construction with generic experiencer with a predicative noun: *radost e*, *veselba e*, *zabava e*. Only the first one *radost e* could also have a stimulus argument – a complement clause with *che* or *da* or an NP.

	stimulus	adjunct
<i>radostno e</i>	NP or <i>che</i> , <i>da</i> complement clause	za-PP
<i>zabavno e</i>		
<i>radost e</i>		
<i>veselo e</i>	NP or <i>da</i> complement clause	

Table 4: Implicit-experiencer constructions

The corpus data shows predominance of the examples with *da*-complement clauses in comparison with *che*-clauses: 19 vs. 11 for *radostno e*, 15 vs. 1 for *veselo e* and 238 vs. 15 for *zabavno e*. We observe the same tendencies in the constructions with dative experiencer. *Radost e* could have either *che* or *da* clauses as their complement, again with more occurrences found with *da* as a complementizer (82 vs. 10 with *che*).

As for the adjunct *za*-phrase, there are single examples with *radostno e* and *veselo e*, 9 with *zabavno e* and 82 with *razost e*. As *za*-PP refers to an animate entity, it competes with dative experiencer, which is possible with *radostno e*, *zabavno e*, *veselo e*. The construction *radost e* has no corresponding construction with an explicit experiencer and *za*-PP is the only animate participant, which could possibly appear with that construction.

## 6. Conclusion

Analyzes on experiencer verbs and constructions based on corpus data show that the experiencer argument is obligatory in the semantic and syntactic structure except for the constructions with *nous* or adverbials, which could have an implicit experiencer. Only the causative object experiencer verbs have always two-argument structure. The stimulus argument could be an NP, a PP or a complement clause. The verbs and the constructions expressing positive emotion vary in the extent to which they accept all those possibilities for the stimulus argument. The data confirms the observations [Becker and Naranjo \(2020\)](#) for the high degree of variation in the

expression of psychological predicates depending on the concept.

## Acknowledgements

This research is carried out as part of the project An Ontology of Stative Situations in the Models of Language: a Contrastive Analysis of Bulgarian and Russian funded by the Bulgarian National Science Fund under the Programme for Bilateral Cooperation, Bulgaria – Russia 2019 – 2020, Grant Agreement No. КП-06-РУСИЯ/23 from 2020.

## References

- Adriana Belletti and Luigi Rizzi. 1988. Psych-verbs and  $\theta$ -theory. *Natural Language & Linguistic Theory*, Vol. 6, № 3, 291–352.
- Anna Wierzbicka. 1999. *Emotions Across Languages and Cultures: Diversity and Universals*. Cambridge: Cambridge University Press.
- Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago, IL: University of Chicago Press.
- Christiane Fellbaum. 1999. The Organization of Verbs and Verb Concepts in a Semantic Net. In P. Saint-Dizier, editor, *Predicative Forms in Natural Language and in Lexical Knowledge Bases*, volume 6 of *Text, Speech and Language Technology*. Springer, Dordrecht.
- David Pesetsky. 1995. *Zero syntax: Experiencers and Cascades*. Cambridge, Massachusetts: The MIT Press.
- Laura Becker and Matías Guzmán Naranjo. 2020. Psych predicates in European languages. A parallel corpus study. *Language Typology and Universals* 73 (4): 483–523.
- Petya Asenova and Zlatka Guentchéva. 2022. Reflexive and Middle Voice. *Glagolati. Balkan Verb Typology*. Sofia: Sofia University Press, 438–462.
- Petya Osenova and Kiril Simov. 2012. The Political Speech Corpus of Bulgarian. *LREC*.
- Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Tsvetana Dimitrova, Rositsa Dekova, and Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, 1: 65–110.



Tomaž Erjavec, Maciej Ogródniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Dargis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2022. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*.

# Evidential strategies and grammatical marking in clauses governed by *verba dicendi* in Bulgarian

**Ekaterina Tarpomanova**  
Sofia University  
St. Kliment Ohridski  
katya@slav.uni-  
sofia.bg

**Krasimira Aleksova**  
Sofia University  
St. Kliment Ohridski  
krasimira\_aleksova@  
slav.uni-sofia.bg

## Abstract

The study explores the interaction between the participants in the communication process with respect to their knowledge about the situation presented in the utterance when transforming direct into indirect speech using a *verbum dicendi*. The speaker has a choice between firsthand (indicative tenses) which by definition denotes a witnessed situation and non-firsthand which presents the situation as non-witnessed. The interplay between the grammatical marking and the speaker's evidential strategy is analyzed by applying a corpus method. The data of the Bulgarian National Corpus are used to detect the preferences for a given strategy considering also the grammatical person which indicates the level of knowledge of the communicants about the situation: the 1<sup>st</sup> person shows the strong knowledge of the speaker, the 2<sup>nd</sup> person is related to the strong knowledge of the listener, and the 3<sup>rd</sup> person is associated with a weak knowledge of both participants. Illustrative examples representative for a given situation are extracted from the corpus and subjected to a context analysis.

**Keywords:** evidentiality, evidential strategy, grammatical marking, knowledge

## 1 Introduction

Bulgarian is among languages with grammaticalized evidentiality, but in sentences with a strong lexical marker such as a verb of utterance, the use of a non-firsthand evidential is not obligatory. Context-induced variability may be viewed as a deviation of the prototypical grammatical category, a manifestation of the grammatical periphery, i.e., obligatory features

whose realization is blocked by the context (Plungian, 2011). In this case, the grammatical category is not entirely blocked by the context (the verb of utterance), but there are several options due to the possibility of realization of the different values (grammemes) of the category.

## 2 Objectives

Our objective is to explore the interaction of the grammatical person in the main and the dependent clause when converting direct to indirect speech after a *verbum dicendi* and the evidential strategy used in the dependent clause. We analyze sentences with the following structure: in the main clause, there is a verb of utterance (we use the verb *казвам/казва* 'say imperfective/perfective' as it is the most frequent and with a generalized semantics to denote an utterance), and the dependent clause – a content clause serving as direct object of the verb of utterance introduced by the complementizer *че* 'that', comprises the converted speech.

(1) *They said that she was ill.*

We analyze two possible strategies in the dependent clause – firsthand and non-firsthand, and how they are motivated by the grammatical person, which relates to the knowledge of the speaker about the situation. We hypothesize that there is a strong relation between the grammatical person and the evidential strategy, as the grammatical person implies the level of knowledge of the participants in the speech act: the 1<sup>st</sup> person shows the strong knowledge of the speaker, the 2<sup>nd</sup> person is related to the strong knowledge of the listener, and the 3<sup>rd</sup> person indicates a weak knowledge of both participants. Our main goal is to find out which strategy is preferred depending on the grammatical person and the tense (considering the opposition between past and non-past tenses). To achieve this goal, we apply corpus-

based methods providing statistical information and analysis of sentences both extracted from the Bulgarian National Corpus (Koeva et al., 2012).

### 3 Evidentiality system and grammatical homonymy in Bulgarian

The evidentiality system of Bulgarian is classified by Aikhenvald (2004) as A1 type (i.e., firsthand vs. non-firsthand), given that the indicative is marked for firsthand, but in fact there are three morphologically marked non-firsthand evidentials: reported, marked by the omission of the auxiliary in the 3<sup>rd</sup> person; inferential, marked by the presence of the auxiliary in the 3<sup>rd</sup> person; dubitative, marked by the auxiliary *бил* in all persons. The non-firsthand evidentials arose from the perfect tense and further developed temporal paradigms (cf. Gerdzhikov, 2003: 214). An important feature of the evidentiality in Bulgarian is the appearance of the imperfect active participle – an innovation that does not exist in the other Slavic languages. It is used in the non-firsthand evidentials and cannot form the perfect indicative.

In the process of paradigm formation, several cases of grammatical homonymy emerged:

- Perfect indicative and aorist inferential (*чел е*). The disambiguation is very difficult, even in the context there are often multiple readings. There is an ongoing debate in the Bulgarian linguistics which form is used in dependent clauses after *verba dicendi* (Gerdzhikov, 2003: 233; Aleksova 2003; Aleksova 2004; Moskova 2019, among others).

- Inferential and reportative in the 1<sup>st</sup> and 2<sup>nd</sup> person (*четял съм, четял си*) – the grammatical marking by the auxiliary applies for the 3<sup>rd</sup> person only;

- Reportative and dubitative: the reportative can express doubt (another point of view is that the auxiliary of the dubitative is omitted and it coincides with the reported);

- Perfect/pluperfect reportative and aorist dubitative (*чел бил*).

### 4 Statistical data

The first step of the present study is to provide statistical information about the evidential strategies in the relevant context. We use the Bulgarian National Corpus to obtain the number of occurrences of the firsthand and the non-firsthand evidentials after *verba dicendi* using as a search

method a regular expression for the following pattern:

- 1) verb of utterance (*казвам/каза* ‘say’) in the respective person in all tenses

- 2) the complementizer *че* ‘that’

- 3) a distance of 0-2 words between the complementizer and the verb in the dependent clause

- 4) firsthand evidential (all tenses of the indicative) / non-firsthand evidential (*l*-participle)

As the disambiguation of the perfect indicative and the aorist inferential is impossible, the perfect has been sorted as an indirect evidential.

The results are presented in Table 1.

Person in the main clause	Person in the dependent clause	Evidential	Number of occurrences (%)
1	1	Firsthand	9305 (91,96%)
		Non-firsthand	813 (8,04%)
1	2	Firsthand	2834 (90,2%)
		Non-firsthand	308 (9,8%)
1	3	Firsthand	11599 (90,8%)
		Non-firsthand	1175 (9,2%)
2	1	Firsthand	924 (90,15%)
		Non-firsthand	101 (9,85%)
2	2	Firsthand	9465 (95%)
		Non-firsthand	492 (5%)
2	3	Firsthand	3810 (84,72%)
		Non-firsthand	687 (15,28%)
3	1	Firsthand	5088 (91,23%)
		Non-firsthand	489 (8,77%)
3	2	Firsthand	2515 (65,36%)
		Non-firsthand	1333 (34,64%)
3	3	Firsthand	34106 (66,04%)
		Non-firsthand	17537 (33,965)

Table 1. Number of occurrences and ratio between firsthand and non-firsthand according to the

configuration of the grammatical person in the main and the dependent clause.

## 5 Analysis of the results

The total number of the sentences with the 3<sup>rd</sup> person in the main clause is the biggest one, i.e., the indirect speech is most often used to transmit the utterance of a non-participant in the speech act. Furthermore, among the sentences with the 3<sup>rd</sup> person in the main clause, most are those with the 3<sup>rd</sup> person in the dependent clause, too (the referent could be the same or different).

In all kinds of combinations of grammatical persons in the main and the dependent clause, sentences with the firsthand in the dependent clause prevails. This fact can be explained with the frequent use of the present indicative in the dependent clause, as in Bulgarian there is no tense agreement.

As it can be seen in the table, in the majority of configurations, the use of the firsthand is more than 90%. There are three combinations that increase the percentage of the non-firsthand evidentials:

2<sup>nd</sup> person – 3<sup>rd</sup> person (*you said that he did something*): 84% vs. 16%;

3<sup>rd</sup> person – 2<sup>nd</sup> person (*he said that you did something*): 65% vs. 35%;

3<sup>rd</sup> person – 3<sup>rd</sup> person (*he said that he did something*): 66% vs. 34%.

The common point of the three cases is the lack of the 1<sup>st</sup> person both in the main and the dependent clause. The combination of the 2<sup>nd</sup> person in the main and the dependent clause does not cause the raise of the percentage of the non-firsthand. The biggest increase of the non-firsthand may be seen in sentences with the 3<sup>rd</sup> person in the main clause – 35% and 34%. These cases imply the weakest knowledge of the situation by the speaker.

## 6 Two evidential strategies: general trends

In sentences with a verb of utterance in the main clause, both firsthand and non-firsthand may occur in the dependent clause, but with the opposite distribution when combined with past and non-past tenses.

### 6.1 Strategy 1: firsthand in the dependent clause (the converted speech)

In the non-past, the verb of utterance in the main clause appears to be sufficient to convey an indirect information (often associated with non-witness position). The use of the firsthand, i.e., the indicative tenses, does not necessarily imply firsthand information, having the potential to indicate both firsthand and non-firsthand.

(2) *Тя каза, че идва.* / *Тя каза, че ще дойде.*

‘She said she is coming. / She said she will come.’

On the contrary, in the past the use of the indicative tenses is restricted; we hypothesize that they emphasize the witness position.

### 6.2 Strategy 2: non-firsthand in the dependent clause

In the non-past the use of the non-firsthand evidentials is optional, they emphasize the non-firsthand information.

(3) *Тя каза, че идвала.*

‘She said she is coming-REP.’

In the past the use of the non-firsthand evidentials is regular with their respective values, except the inferential which rarely expresses inferred information, but rather is a neutral (non-emphatic) means to denote a non-witness position.

## 7 Analysis of instances of the evidential strategies

In what follows, we make qualitative analysis of sentences extracted from the BulNC and sorted by the person in the main clause. We aim at establishing how the choice of a given strategy is motivated by the grammatical person, at the same time considering the abovementioned relation between evidential strategy and tense (past or non-past).

## 8 1<sup>st</sup> person in the main clause

With the 1<sup>st</sup> person in the main clause the speaker reports their own information.

### 8.1 Firsthand

As the 1<sup>st</sup> person is associated with the actual speaker, the information in the utterance is presented as strong knowledge. The firsthand in the dependent clause occurs regularly in the non-past, but it is not unusual even if the event has a past

reference – as in (5), emphasizing the witness position.

(4) *Казах, че ти не разбираш.*

‘I said that you don’t understand.’

(5) *Казах, че беше така. Лъжец ли ме наричаш?*

‘I said it was like that. Are you calling me a liar?’

## 8.2 Non-firsthand

Using the non-firsthand the speaker focuses on their non-witness position about the situation in the dependent clause. In fact, a good number of the sentences with such interpretation contain a negative form of the verb ‘say’, by which the speaker distances him/herself from his/her own words.

(6) *Не казвам, че си искал да убиваш.*

‘I’m not saying that you intended to kill.’

Another group of instances of the non-firsthand combined with the 1<sup>st</sup> person is associated with an unusual situation: the speaker simulates that the information is indirectly acquired to underline that it is a false statement (a lie).

(7) *Казах им, че една ръждясала решетка се е строшила под вас. Казах, че случайно сте паднал и сте пропъзлял в укритие. ... Те приеха честната ми дума и си тръгнаха.*

‘I told them a rusty grille had broken under you. I said you accidentally fell and crawled into hiding. ... They accepted my word of honor and left.’

In some sentences the verb form composed of the auxiliary ‘be’ and the aorist active participle has a perfect reading and therefore should not be interpreted as non-firsthand. The perfect reading is often supported by the typical adverbials that collocate with the perfect, the so-called reference time adverbials, such as *already, always, ever, never*, etc., as opposed to the event time adverbials that denote the time point in which the event occurs and collocate with the aorist (after Reichenbach 1947).

(8) *Нали ти казах, че никога не съм изпитвала такива чувства спрямо някого.*

‘Didn’t I tell you that I have never felt like that about anybody.’

## 9 2<sup>nd</sup> person in the main clause

With the 2<sup>nd</sup> person in the main clause the speaker quotes the utterance of their interlocutor.

## 9.1 Firsthand

The firsthand in the dependent clause emphasizes the witness position of the actual speaker especially with the 1<sup>st</sup> person in the dependent clause.

(9) *Значи мойта идея ви допадна? – Та нали вече каза, че и сам бях стигнал до нея.*

‘So, you liked my idea? – But you already said that I came up with it myself.’

The witness position is possible also with the 2<sup>nd</sup> and the 3<sup>rd</sup> person in the dependent clause. In the sentences below the speaker presents his/herself as a witness to underline his/her strong knowledge. Interestingly enough, the two sentences contain a verb of mental activity so the speaker could not be a witness in the strict sense and the firsthand evidential is rather a means to demonstrate a strong knowledge.

(10) *Кажете, че излъгахте и още сега ще ви бъде простено.*

‘Say that you lied, and you will be forgiven right now.’

(11) *И не ми казвайте, че не знаехте, че пътят е забранен.*

‘And don’t tell me you didn’t know that this road was forbidden.’

In many cases the verb of utterance in the main clause implies that the information is non-firsthand and the use of a non-firsthand evidential is not necessary. This holds especially for non-past tenses.

(12) *Казваш, че те преследва чудовище.*

‘You say you are being chased by a monster.’

## 9.2 Non-firsthand

With the 2<sup>nd</sup> person of the verb of utterance, the non-firsthand strategy in the dependent clause has various manifestations.

The number of sentences where the *l*-form could be interpreted as a perfect remains unidentified, we consider lexical features and the general context.

(13) *Колко казахте, че **сте сътворили** досега?*

‘How many you said you have created up to now?’

(14) *Казваш, че **съм пораснал** ли... аз съм остарял!*

‘You say I have grown up... but I have grown older!’

A regular instance of the non-firsthand is the non-witness position of the speaker who quotes the listener’s words.

(15) *Каза, че си я познавал.*

‘You said you knew her.’

With the 1<sup>st</sup> person in the dependent clause, the non-witness position means that the speaker does not remember the situation described in it.

(16) *Казваш, че съм прекарал тук около три хиляди години. Може и така да е.*

‘You are saying that I spent about three thousand years here. That may be so.’

With the 3<sup>rd</sup> person in the dependent clause the inferential and the reportative differ by the presence or the omission of the auxiliary, the reportative focusing on the fact that the speaker quotes the listener’s words.

(17) *Казваш, че носела пистолет.*

‘You are saying that she had a gun.’

To express the present with a non-witness position, only the reportative is possible, as the inferential cannot have a present value.

(18) *Ти каза, че имало неща, които трябва да видя.*

‘You said there are three thing I have to see.’

Dubitative interpretation is possible too, expressed with either dubitative or reportative.

(19) *Да живееш Негово Царско Височество! Виждаш ли как викам да живееш, пък ти си взел да казваш, че съм бил против.*

‘Live His Majesty! You see, I’m saying “live!”, and you say that I’m against.’

(20) *Хмм! А казваш, че били страхливци!*

‘Hmm! And you say they are cowards.’

### 9.3 Imperative

A special case are sentences with the imperative in the main clause by which the speaker wants the interlocutor to make a particular statement. In such context the future has the same function. The firsthand has not any specificity.

(21) *Кажу, че изпълняваш заповед на принца.*

‘Say you’re following the prince’s orders.’

In the majority of the sentences with non-firsthand in the dependent clause the speaker wants the interlocutor to make a false statement, i.e., to utter a lie.

(22) *Ако е някой за мен, кажи, че съм си легнал.*

‘If it’s for me, say I’m in bed.’

(23) *После за съда аз ще намеря добър адвокат. Ще отречеши признанието. Ще кажеш, че си бил пиан.*

‘Then I’ll find a good lawyer for the court. You will deny the confession. You’ll say you were drunk.’

A specific interpretation is found in sentences with negative form of the non-firsthand in the dependent clause – the speaker takes a non-witness position and asks the interlocutor to deny their assumption about the situation

(24) *Но ти нали не можеш да говориш! Не живееш в този свят, не знаеш, че се казвам Вероника! Снощи не си бил с мен, моля те, кажи, че не си бил! – Бях. Тя взе ръката му.*

‘But you can’t talk, can you! You don’t live in this world, you don’t know my name is Veronica! You were not with me last night, please say you were not! – I was. She took his hand.’

(25) *Шон, погледни ме в очите и ми кажи, че не си взел тези пари!*

‘Sean, look me in the eye and say you didn’t take that money!’

## 10 3<sup>rd</sup> person in the main clause

Using the 3<sup>rd</sup> person in the main clause, the speaker reports somebody else’s utterance.

### 10.1 1<sup>st</sup> person in the dependent clause

In sentences with the 1<sup>st</sup> person in the dependent clause there is no change in the ratio between firsthand and non-firsthand, i.e., the firsthand is the predominant strategy expressing strong knowledge of the speaker often resulting from their witness position.

(26) *Гералт казва, че вече съм много добра на махалото. Казва, че имам такова, във... Усет.*

‘Geralt says I am already very good on the pendulum. He says I have... uuuh... flair.’

(27) *Не може да се каже, че разговаряхме.*

‘It can’t be said that we talked.’

The use of the non-firsthand is associated with the emphasis of the reported speech.

(28) *Казва, че съм имала опашката на някакъв бог саламандър.*

‘He says I have the tail of some salamander god.’

(29) *Чисто и просто казва, че много съм пиел.*

‘She just says I drink a lot.’



## 10.2 2<sup>nd</sup> or 3<sup>rd</sup> person in the dependent clause

In sentences with the 3<sup>rd</sup> person in the main clause and the 2<sup>nd</sup> or the 3<sup>rd</sup> in the dependent clause we found the biggest increase of the non-firsthand in the dependent, because they exhibit the weakest knowledge about the situation.

In the non-past, it is still possible to express non-witness position by the firsthand, i.e., the lexical item (the verb ‘say’) is the only evidential marker.

(30) *Хем ми казаха, че не **нараняваш** хората.*  
‘But they told me you don’t hurt people.’

As for the past, the non-firsthand is preferred. In the Bulgarian linguistics there is a widespread opinion that the past indicative (especially the aorist) cannot occur after a verb of utterance. In fact, we found a few examples in which the past indicative is used to emphasize the speaker’s strong knowledge usually associated with a witness position.

(31) *Интересува ме кой е убил жената на Ленъкс. – Боже мой, Гренц не ви ли каза, че той **написа** пълно признание? Дори вестниците го публикуваха. Вие не четете ли пресата?*

‘I wonder who killed Lennox's wife. – My God, didn't Grenz tell you he wrote a full confession? Even the newspapers published it. Don't you read the press?’

(32) *Тад ѝ каза, че не **бе успял** да запише номера.*

‘Tad told her he hadn’t been able to write down the number.’

Although the non-firsthand is the prevailing strategy in the past, there are, however, sentences with a possible perfect interpretation.

(34) *Казваха, че **си загинал**.*

‘They said you were dead.’

Most often the non-firsthand denotes non-witness position when the information is reported. When the verb in the dependent clause is in the third person, the differentiation of the reportative and the inferential is possible.

(35) *Един шофьор ми каза, че **е видял** колата.*

‘A driver told me he saw the car.’

(36) *Каза, че **можело** да означава само едно – магия.*

‘He said it could only mean one thing – magic.’

Provided that the 3<sup>rd</sup> person allows for grammatical disambiguation between the non-firsthand evidentials (reported, inferential and dubitative) based on the auxiliary (omission, presence, *бил*, respectively), it is possible to verify

which non-firsthand strategy is preferred. To find out the ratio of the three non-firsthand evidentials, we searched for the following strings:

- reported: *каза* ‘he/she said’ + *че* ‘that’ + aorist/imperfect active participle;
- inferential: *каза* ‘he/she said’ + *че* ‘that’ + auxiliary *е* ‘is’ + aorist/imperfect active participle;
- dubitative: *каза* ‘he/she said’ + *че* ‘that’ + auxiliary *бил* ‘is DUB’ + aorist/imperfect active participle.

	with aorist active participle	with imperfect active participle	total
inferential	2396	464	2860 (54%)
reported	1449	934	2383 (45%)
dubitative	77	0	77 (1%)

Table 2. Ratio of the non-firsthand evidentials after *каза* ‘he/she said’.

The inferential appears to be predominant although after a verb of utterance reportative meaning is expected. On the other hand, the grammatical homonymy between the aorist inferential and the perfect indicative, both consisting of the auxiliary ‘be’ and the aorist active participle, is difficult to resolve in this context. Yet the imperfect inferential is distinguishable from the perfect indicative as it is formed with the imperfect active participle. Subsequently the instances with imperfect active participle should be interpreted only as non-firsthand. Here another type of grammatical homonymy impedes the analysis – the formal coincidence of the aorist and the imperfect active participles of verbs of the 3<sup>rd</sup> conjugation. The manual review of the search results showed there are only six instances of the sequence auxiliary + imperfect active participle of verbs of 1<sup>st</sup> or 2<sup>nd</sup> conjugation (out of 464) that could be unambiguously interpreted as imperfect inferential. The rest are ambiguous – a perfect indicative reading is possible.

## 11 Aorist inferential and perfect indicative – disambiguation impossible?

In the Bulgarian linguistics there are two opposite opinions about the grammatical form in

the dependent clause after a verb of utterance consisting of the auxiliary ‘be’ and the aorist active participle – it is interpreted either as aorist inferential or as perfect indicative with the respective arguments.

### 11.1 Arguments for aorist inferential

If we assume that in the original utterance as direct speech a past indicative tense (aorist or imperfect) is used, then in the converted indirect speech after the verb of utterance the respective non-firsthand (inferential) tenses would appear (Moskova 2019).

(37) *Иван: Аз пристигнах вчера. > Иван каза, че е пристигнал вчера.*

‘John: I arrived (AOR IND) yesterday. > John said he arrived (AOR INF) yesterday.’

On the other hand, the context implies a reported semantics and there is a specialized reportative evidential in Bulgarian.

### 11.2 Arguments for perfect indicative

The perfect has taxis use after *verba dicendi, sentiendi, cogitandi*. The perfect has been generalized as a universal tense to express an event which is prior to the event in the main clause regardless of the tense in the main clause, presenting the viewpoint of the cognitive subject (Nitsolova 2008: 298). In sentences with verbs of perception, there is often firsthand semantics.

(38) *Погледай ме на какво съм заприличала* (А. Каралийчев).

‘Look at what I have become.’

### 11.3 Contamination

Another possible interpretation is that a contamination of the perfect indicative and aorist inferential took place in contexts that support past and non-firsthand reading simultaneously.

## 12 Conclusions

Bulgarian is a language with grammaticalized evidentiality but displays complicated strategies in communicative acts with converted speech after verbs of utterance involving both firsthand and non-firsthand evidentials. Some problems are difficult to resolve due to the grammatical homonymy. However, conclusions about evidential strategies in the described context can be made.

The main viewpoint for the choice of evidential strategy is the knowledge of the speaker about the information they communicate. The 1<sup>st</sup> person in the main and/or in the dependent clause is connected to the predominance of the firsthand strategy. The non-firsthand evidentials combined with the 1<sup>st</sup> person are often associated with a false statement. The same function may have the 2<sup>nd</sup> person imperative or future of the verb ‘say’ in the main clause followed by non-firsthand in the dependent clause, with which the speaker expresses their wish the false statement to be made by the addressee.

The weakest knowledge of the speaker is encoded in the 3<sup>rd</sup> person and results in the increase of the non-firsthand in the dependent clause. The grammatical marking of the non-firsthand evidentials in the 3<sup>rd</sup> person allows for the differentiation of the inferential and the reported, but the homonymy between the aorist inferential and the perfect indicative remains difficult to resolve. The dubitative is marked in all persons and even in cases of homonymy with the perfect/pluperfect reportative, the disambiguation is easy in the context.

Despite the grammaticalization of the evidentiality, the verb ‘say’ is a strong evidential marker, and in some contexts, it is sufficient to indicate the non-firsthand.

## References

- Krasimira Aleksova. 2003. Udostoveriteln perfekt ili umozaklyuchiteln aorist – ot teoretichnite osnovi kam prepodavaneto na chuzhdentsi. *40 godini IChS. Yubileyna nauchno-prakticheska sesiya*, Sofia, 60 – 66.
- Krasimira Aleksova. 2004. Otnovo za otnoshenieto indikativen perfekt – konkluziven aorist (nyakoi teoretichni aspekti s ogled i na obuchenieto po balgarski ezik na chuzhdentsi). – *Treta mezhdunarodna konferentsiya “Ezikat – sredstvo za obrazovanie, nauka, profesionalna realizatsiya”*, 183 – 191. Varna: Steno.
- Georgi Gerdzhikov. 2003. *Prezikazvaneto na glagolното deystvie v balgarskia ezik*. Sofia: UI “Sv. Kliment Ohridski”.
- Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova, Ekaterina Tarpomanova. 2012. *The Bulgarian National Corpus: Theory and Practice in Corpus*

Design. *Journal of Language Modelling*, (1), pages 65–110.

Mihaela Moskova. 2019. Upotrebi na konkluziva pri predavane na ruzhda rech v podchinenoto izrechenie s glagol za predavane na chuzhda rech v glavnoto. – *Savremenna lingvistika*, 2, 19 – 30. [https://slav.uni-sofia.bg/images/bf/SPISANIE\\_LINGVISTIKA\\_2\\_2019.pdf](https://slav.uni-sofia.bg/images/bf/SPISANIE_LINGVISTIKA_2_2019.pdf)

Ruselina Nitsolova. 2008. *Balgarska gramatika. Morfologiya*. Sofia: UI “Sv. Kliment Ohridski”.

Vladimir Plungyan. 2011. *Vvedenie v grammaticheskuyu semantiku: grammaticheskie znacheniya i grammaticheskie sistemi yazikov mira*. Moskva: RGGU.

Hans Reichenbach. 1947. *Elements of Symbolic Logic*. New York: Macmillan & Co.

# Corpus-Based Research into Verb-Forming Suffixes in English: Its Empirical and Theoretical Consequences

Junya Morita

Kinjo Gakuin University/ 2-1723 Omori Moriyama-ku, Nagoya, Japan 463-8521  
morita@kinjo-u.ac.jp

## Abstract

The present study explores the semantic and structural aspects of word formation processes in English, focusing on how verbs are derived by the suffixes *-ize*, *-ify*, *-en*, and *-ate*. Based on relevant derivatives extracted from the British National Corpus, their detailed observation is made from semantic and formal viewpoints. Then their theoretical analysis is carried out in the framework of generative theory. The BNC survey demonstrates that (i) the meanings of derived verbs are largely divided into five types and the submeanings are closely related to each other, (ii) the well-formedness of derived verbs is primarily determined by the semantic and formal features of their bases, and (iii) *-ize* suffixation is creative enough to provide a constant supply for new labels. To account for these empirical observations, the mechanism for forming *-ize* derivatives is proposed in which the semantic properties and creativity of *-ize* derivation stem solely from the underlying structure and the formal properties of the bases derive from the lexical entry of *-ize*.

**Keywords:** corpus research, derived verbs, semantics, morphosyntax, word formation mechanism

## 1 Introduction

The central issue of generative morphology is how to account for children's lexical acquisition: they acquire the vocabulary rapidly and accurately based on limited and often degenerated data. The most promising way of achieving this is to

establish a general mechanism to generate an infinite number of possible words so that words to remember are greatly reduced in number. In addition, the mechanism itself needs to be of wide application and elegant in linguistic terms to minimize information specified in the grammar.

As part of the above enterprise, the present study attempts to construct a system which automatically produce well-formed derived verbs, as in “she has been hopelessly *sentimentalized* and hopelessly *magicalized* by tradition (BNC: ACL 1522).” This study is organized as follows: after outlining the method of research (section 2.1), we closely examine the derivation of verbs in English and illuminate its semantic features (sections 2.2 and 2.3). Then we elucidate its morphological properties—the formal restrictions of the bases and productivity (sections 2.4-2.6). Section 3 presents theoretical implications for the results of our research. A summary of the main arguments is presented in section 4.

## 2 Observation and Generalizations

### 2.1 Target and Methodology

In this section, we will make systematic observations of English derived verbs and present generalizations based on it. We now begin a brief description of the method of research and the resulting data. By repeatedly using the “wild card” function of a research engine, words ending in *-ize*, *-ify*, *-en*, and *-ate* are extracted from the British National Corpus (BNC), a 100-million-word corpus.<sup>1</sup> In particular, their frequency is checked to identify the hapax legomena (token frequency 1). As a result of the research, we have obtained 381 word types in *-ize*, 68 word types in

---

<sup>1</sup>I am indebted to the research engine of [www.english-corpora.org](http://www.english-corpora.org) (BNC).

-ify, 58 word types in -en, and 447 word types in -ate, including 123, 5, 2, and 26 hapaxes, respectively.

## 2.2 The Meanings of Derived Verbs

This section and the next deal with the semantic aspects of verb derivation. The semantic aspects of derived verbs have been well observed in the literature from a descriptive perspective (Jespersen, 1949; Marchand, 1969) and from a generative perspective (Plag, 1999; Lieber, 2005). According to Plag (1999: 125), the meanings of derived verbs can be divided into seven classes: 1 resultative ‘make into x’; 2 locative ‘put (in)to x’; 3 ornative ‘provide with x’; 4 performative ‘perform x’; 5 similitive ‘act like x’; 6 causative ‘make x’; 7 inchoative ‘become x.’

We will revise this classification in the following way. First, ‘resultative’ and ‘locative’ may be classed together as ‘result’; *atomize* denotes ‘put sth in a state of an atom’ and *hospitalize* signifies ‘put sb into a hospital,’ and thus both of them are associated with the change of state or place.

Second, two more submeanings join the classification, ‘agentive’ and ‘instrumental’; *patronize* and *cauterize* are interpreted as ‘act as patron’ and ‘do sth with cauter,’ respectively. Then we may group ‘ornative,’ ‘performative,’

‘agentive,’ and ‘instrumental’ under the heading of ‘providing or giving.’ This is because they are all interpretable as ‘make sb/sth provided with x; *chemicalize* (ornative) means ‘give chemical to sb/sth,’ *dichotomize policemen* (performative) signifies ‘give or apply the process of dichotomy to policemen,’ *patronize the shop* (agentive) represents ‘(in a widened sense) assign a patron to the shop,’ and *cauterize* (instrumental) denotes ‘provide sb with cauter.’ Finally, there is one other addition to the traditional classification; the submeaning “function,” referring to ‘make sth be as x,’ should be set up to interpret *canonize the texts* as ‘make the texts be as canon.’<sup>2</sup>

Table 1 shows the results of our research into the semantics of derived verbs.<sup>3</sup> Deadjectival derived verbs are essentially transitive verbs and have the meaning of ‘make sth x’ (causative), as in *circularize* ‘make sth circular.’ We see in Table 1 that the submeaning ‘causative’ is the highest in number of word types and hapaxes, showing that this is the central meaning of derived verbs. Part of these derivatives can be used as intransitive verbs and they mean ‘become x’ (inchoative). This shift has been well studied from a generative perspective; it is frequently treated as an alternation between transitives and inchoatives (Levin and Rappaport Hovav, 1995). We will not, though, deal with the issue of how they are related to each other.<sup>4</sup>

base	the meaning of derivative	-ize	-ify	-en	-ate	total
Adj	causative	215 (67)	22 (3)	51 (2)	19 (0)	307 (72)
N	(a) result					
	(i) resultative	51 (13)	17 (2)	4 (0)	7 (0)	79 (15)
	(ii) locative	3 (0)	0	0	0	3 (0)
	(b) providing					
	(i) ornative	35 (16)	3 (0)	1 (0)	8 (1)	47 (17)
	(ii) performative	35 (8)	3 (0)	1 (0)	4 (1)	43 (9)
	(iii) agentive	7 (4)	0	0	2 (0)	9 (4)
	(iv) instrumental	5 (2)	0	0	0	5 (2)
	(c) similitive	15 (9)	1 (0)	0	0	16 (9)
	(d) function	3 (2)	1 (0)	0	1 (1)	5 (3)
	purpose	1 (1)	0	0	0	1 (1)
	bound stems	11 (1)	21 (0)	1 (0)	406 (23)	439 (24)
	total number of types (hapaxes)	381 (123)	68 (5)	58 (2)	447 (26)	954 (156)

Table 1: The submeanings of -ize, -ify, -en, and -ate words

<sup>2</sup>We can find the submeaning ‘purpose’ (e.g. *winterize*), but this is quite exceptional.

<sup>3</sup>If a derived verb can be interpreted in two ways (e.g. *capitalize* ‘resultative/ornative’), it is separately counted. Cases of this kind are very few indeed—only 2 cases. Additionally, when the base can be an adjective or noun, the one which is naturally interpretable is chosen. For example, *editorialize* is denominal, since it means ‘to express an opinion in an editorial.’

<sup>4</sup>Levin and Rappaport Hovav (1995: 102-104) comment as follows: transitive verbs implying the intervention of an intentional agent do not have intransitive forms, as exemplified in (i), and -ize and -ify verbs are generally of this kind so that they cannot normally detransitivize, as illustrated in (ii).

- (i) a. The terrorist *assassinated* the senator.
- b. \*The senator *assassinated*.
- (ii) a. The farmer *homogenized* the milk.

As for denominal derived verbs, the submeanings of ‘resultative,’ ‘ornative,’ and ‘performative’ form a majority in number of word types and hapaxes, and so they are the central meanings of denominal derived verbs. The productivity of each derived verb will be discussed in section 2.6.<sup>5</sup>

### 2.3 Correlation between Derived Verbs and Their Bases

This section shows how, and to what extent, derived verbs’ meanings are predicted from their bases. This issue has received relatively little attention in previous morphological analyses. We have already stated that deadjectival verbs generally have the causative sense ‘make sth x’ and part of them may become inchoatives. The meanings of denominal verbs are largely divided into four classes and each class is closely related to the meanings of the base. Importantly, the three main subclasses of meanings—‘resultative,’ ‘ornative,’ and ‘performative’—are generally predictable from the bases’ meanings. The meaning correspondence is offered in Table 2.<sup>6</sup>

#### ‘resultative’

- (a) state/quality: *dimerize, fossilize, harmonize, isomerise, mylonitize, pauperize*, (19)
- (b) language: *capitalize<sub>1</sub>, diarize, editorialize, journalize, novelize, poetize, rhapsodize* (10)
- (c) basic element: *atomize, carbonize, oxidize, unitize* (4)
- (d) system/classification: *categorize, factorize, computerize, memorize, organize* (21)
- (e) one’s status: *deputize<sub>1</sub>, idolize* (2)

#### ‘ornative’

- (a) (bio)chemical substance: *chemicalize, heparinize, narcotize, siliconize, trypsinize* (8)
- (b) presentable thing: *accessorize, capitalize<sub>2</sub>, caramelize, deodorize, subsidize* (12)
- (c) academic matter: *anthropologize, biologize, botanize, philosophize, symbologize* (7)
- (d) format: *alphabetize, rasterize, tokenize* (3)
- (e) rights: *autonomize, hegemonize, prioritize* (3)

#### ‘performative’

action/process: *anatomize, apologize, eulogize, assassinize, dichotomize, economize* (32)

Table 2: Meaning correlation between derived verbs and their bases

Three points are worth noting here. First, a derived verb denotes ‘resultative’ when the base noun has one of the five meanings; if an underlying noun (*dimer*) expresses a state or quality, the derivative (*dimerize*) can naturally be taken as meaning ‘put sth in a state/quality.’ From nouns related to language are derived verbs that denote ‘put sth into a verbal form,’ as in *diarize*. Nouns indicating basic elements and those conveying system/classification are verbalized to mean ‘put sth into a basic element’ (*atomize*) and ‘put sth into a system/classification’ (*categorize*), respectively.

Second, the given meanings of base nouns lead to the meaning ‘ornative’ of the derivatives; from the names of (bio)chemical substance are derived verbs that signify ‘give the substance,’ as in *chemicalize*. This correlation is reasonable, since chemical substance is usually given to somebody or something to cause chemical action. Examples such as *accessorize* and *capitalize<sub>2</sub>* ‘provide (a company) with capital’ can be treated similarly; accessory and capital are presentable, that is, suitable to be presented. In addition, the ‘ornative’ meanings of derived verbs are commonly expected from the underlying nouns indicating academic matter, format, or rights. For instance, *anthropologize* and *alphabetize* imply ‘provide an anthropologic view’ and ‘provide an alphabet format,’ respectively. Finally, we can easily assign the ‘performative’ meaning to a derived verb when the base noun entails an action or process, as in *anatomize*.

Turning to other submeanings, we can easily understand that a verbal suffix combines with a noun expressing a place to produce a verb with a sense of locative (*palletize*) and a verbalizer is added to an agentive noun to form a verb that refers to the related action (*burglarize*). From nouns of instruments are derived verbs that denote the action for which the instruments are meant (*catheterize*) and ‘similative’ verbs are built from proper nouns (*Beethovenize*) and animal names (*serpentize*).

b. \*The milk *homogenized*.

<sup>5</sup>As evidenced in *refer, remit, and resume*, bound stems generally have no fixed meanings; only words may have constant meanings. Thus, “all regular word-formation processes are word-based” (Aronoff, 1976: 21). According

to this thesis, a case in which a verbal suffix attaches to a stem is left out of consideration here.

<sup>6</sup>The number in a parenthesis indicates the total number of word types.



## 2.4 Formal Restrictions on the Bases

The previous sections have examined semantic facets of verbalization. In this section and the following two sections, we will demonstrate the morphological facts on derived verbs. We will concentrate here on the internal structure of the bases and their vocabulary strata. Restrictions are imposed on the size and composition of the bases. First, as is pointed out in Marchand (1969: 100), verbal suffixes do not combine with compounds and this is attributed to the general inhibition of direct verb compounding (cf. *\*to rock-throw*). Our research supports this view; there are no such verbs in BNC (*\*rock-crystalize*, *\*rock-solidify*, *\*knife-sharpen*). Second, as a result of the same research, we find that a verb-forming suffix generally does not attach to prefixed bases. Thus, *-ize*, *-ify*, *-en*, and *-ate* do not combine with words including prefixes such as *a-*, *trans-*, and *ultra-* (*\*atypicalize*, *\*transcontinentalize*, *\*ultratrendify*). We have only three counterexamples to this: *immobilize*, *impersonalize*, and *internationalize*.

Third, Lieber (2005: 412) states that the verbalizers *-ize* and *-ify* normally do not attach to suffixed words, excepting those ending in *-al*, *-ian*, and *-ic*. However, our research demonstrates that there is considerable variation in the combination of suffixed words between verbal suffixes. *-Ize* attaches to words ending in *-able* (*permeabilize*), *-ive* (*passivize*), *-er* (*computerize*), *-(a)(t)ion* (*revolutionize*) in addition to *-al*, *-(i)an*, and *-ic* bases (*commercialize*, *Christianize*, *classicize*). Some suffixes in verbal bases are truncated when combined with *-ize*, as exemplified by *-ous* in *anonymize*. Morpheme truncation will be discussed in the next section. Contrastively, other verbal suffixes can attach to suffixed bases in a very limited way. *-Ate* can attach to *-al*, *-ant*, *-ic* and *-ous* bases (*liberate*, *resonate*, *rubricate*, *stimulate*), *-ify* can be added to *-ic* and *-ity* bases (*mystify*, *commodify*), and *-en* can affix to *-(i)an* and *-th* bases (*Christen*, *strengthen*). Most of these base-internal suffixes are truncated in combination with the verbal suffixes. We can say then that *-ize* affixation is a major verb-forming process in the sense that it may attach to various suffixed bases to produce a variety of verbs.

Let us now turn to the issue of vocabulary strata. It has been well observed that an affix chooses an

item of a specific vocabulary stratum; *-ize*, *-ify*, and *-ate* typically combine with words of Latinate origin, while *-en* normally combines with words of native origin (Jespersen, 1949; Marchand, 1969). Our BNC research has identified the vocabulary strata of words with which each suffix combines: (i) [Latinate] (354 word types), [Greek] (18), [Native] (7), the others (2) for *-ize*; (ii) [Latinate] (64), [Native] (3), the others (1) for *-ify*; (iii) [Native] (53), [Latinate] (5) for *-en*; (iv) [Latinate] (447) for *-ate*. The result leads us to conclude that *-ize* mostly takes [Latinate] or [Greek] bases, *-ify* and *-ate* predominantly or exclusively take [Latinate] bases, while *-en* mainly takes [Native] bases. Thus, the previous observations have been confirmed by our BNC research.

It is widely accepted that affixes can be divided into two classes: one may cause phonological change of the base (class I), while the other is phonologically neutral (class II). Additionally, their ordering is recognized: class I affixes cannot appear outside class II affixes. *-Ize* may be considered as a class I affix, since it may change the phonological quality of the base (cf. *stable* and *stabilize*). According to Selkirk (1982: 81), the suffixes *-ful*, *-less*, *-ly*, *-y*, *-ish*, *-en*, *-ed*, *-some*, *-able*, *-er* are all class II suffixes, and hence they are predicted not to occur in *-ize* derivatives. This prediction is confirmed by the ill-formedness of words such as *\*harmfulize*, *\*powerlessize*, and *\*friendlyze*, which are never found in BNC. It is worth noting here that all the suffixes except *-able* are of native origin. The co-occurrence restriction is then deduced from the requirement that a base be largely Latinate or Greek, and therefore the present ordering will be unnecessary for *-ize* verbalization.<sup>7</sup>

## 2.5 Truncation of a Word-Internal Suffix

This section deals with the truncation of a word-internal suffix concerning verbalization, focusing on *-ize* affixation. There are good reasons for the truncation of an intra-word suffix. One is that the underlying form of *[X-suffix]-ize* is well suited to the meaning of the whole word. For example, *systematize* means ‘make sth systematic’ and so the meaning is assigned easily and naturally to the word if *-ic* is underlyingly involved in the word. Another is that we can get rid of an unnecessary bound base; the lexicon would be redundant if the

<sup>7</sup>Selkirk (1982: 81) points out that *-able* has dual status, that is, it may be a member of both classes. It might be argued

then that the type of the affix *-able* involved in *-ize* derivatives belongs to class I.

bound stem *systemat-* were listed only for *-ize* affixation. The strongest reason of all is the fact that there exists a doublet of truncated form and untruncated form, as exemplified in *digitize/digitalise* and *monetize/monetarize*. There seems to be no significant meaning difference between both forms, and hence their relationship can be described clearly by the relevant truncation.

With respect to suffix-containing *-ize* verbs, some internal suffixes are truncated while others are not. The results of our BNC survey are shown in Table 3.

**truncated suffixes:** *-ic* (20 types), *-ous* (4), *-al* (3), *-ity* (3), *-ant* (1), *-ism* (1), *-ive* (1)  
**untruncated suffixes:** *-al* (81), *-ic* (13), *-(i)an* (8), *-able* (3), *-(a)(t)ion* (2), *-ary* (1), *-er* (1)

Table 3: Truncation of a suffix in *-ize* words

Seven suffixes are deleted in *-ize* verbs: *-ic* (e.g. *anaesthetize*), *-ous* (*anonymize*), *-al* (*attitudinize*) *-ity* (*authorize*), *-ant* (*deodorize*), *-ism* (*ostracize*), and *-ive* (*sensitize*). By contrast, seven suffixes prove to be intact in *-ize* verbs: *-al* (e.g. *centralize*), *-ic* (*classicize*), *-(i)an* (*Americanize*), *-able* (*respectabilize*), *-(a)(t)ion* (*productionize*), *-ary* (*militarize*), and *-er* (*computerize*).<sup>8</sup>

*-Ic* truncation deserves special mention. *Ic-* is essentially deletable in the position at issue; twenty word types of such derivatives are identified in BNC. However, we detect thirteen word types of derivatives whose internal *-ic* is not deleted: (i) *classicize*, *ethicize*, *Gallicize*, *Gothicize*, *poeticize*, *publicize*, (ii) *romanticize*, *geometricize*, (iii) *aestheticize*, *cosmeticize*, *eroticize*, *hermeticize*, *phonemicize*. Looking closely at these examples, we notice that the base of the internal suffix *-ic* is monosyllabic as in (i) and it ends in two consonants as in (ii). Then, a generalization emerges: when the base of the internal suffix *-ic* is polysyllabic or ends in a single consonant, *-ic* truncation applies. Although the examples in (iii) remain unaccounted for, the generalization applies to the *-ive* truncation as well (cf. *passivize* and *\*passize*).

To conclude this section, the internal suffix *-al* is generally intact in *-ize* derivatives while suffixes like *-ous* and *-ity* are truncated. The suffix *-ic* may

be either truncated or untruncated and a generalization can be made about the truncation process at work.

## 2.6 Productivity

As the last morphological facet, we will discuss the productivity of verb-forming suffixes. A hapax-centered productivity measure for derivation is applied to data collections to calculate the productivity value of verb-forming process. We accept a hapax-based productivity measure, which gives a key role to hapax legomena of a large-scale corpus (Baayen and Renouf, 1996). This rests on the view that the capacity of an affix to create new forms crucially involves the degree to which the affix yields words of ultra-low frequency (Hay, 2003).

We propose a productivity measure:  $Productivity (P) = n_i/V$ , where  $n_i$  is the number of hapaxes and  $V$  is the total number of word types.<sup>9</sup> Our BNC research detects 123 hapaxes and 381 word types of *-ize* derivatives, giving its productivity value of 0.323 (cf. Table 1). In this measure, the productivity of *-ize* affixation is defined as the potentiality of creating 123 kinds of new words when 381 kinds of *-ize* derivatives are used; nearly one-third of the attested *-ize* types are innovated verbs. According to the same measure, the productivity values of *-ify*, *-en*, and *-ate* verbalization are, respectively, 0.074, 0.035, and 0.058. The results of the research then demonstrate that while *-ify*, *-en*, and *-ate* are not productive affixes, *-ize* affixation is fairly productive to promote the creation of neologisms.

Additionally, *-ize* derivatives may be created depending on context. In example (1), the process of making worms into arthropods is momentarily lexicalized with the verb *arthropodize*, relying on the preceding noun *arthropods*. Example (2) illustrates how a complex word is created in the enumerative or listing environment; a series of comparable activities are enumerated by the use of three *-ize* final verbs, with *moronised* and *lobotomised* being innovated. Online word formation at issue is largely determined by the functions of “naming” (to conceptualize a property by giving it a name) and “brevity” (to construct a concise and sensible word) (Clark and Clark, 1979;

<sup>8</sup>We confine our attention to well-established and recognized suffixes, that is, those listed in Quirk et al. (1985: 1548-1555). Hence we leave out of consideration suffixes like *familiar*, *alkaline*, and *maximum*.

<sup>9</sup>This productivity measure is a revised version of the one proposed by Baayen and Renouf (1996), who place the total number of tokens in the denominator of the productivity formula.

Rice and Prideaux, 1991). That *-ize* words may be constructed wherever there exist such functional requirements confirms the derivational potentials of the verbalizer investigated.

- (1) ... different *arthropods* may have come from different and separate worms, independently, which became “*arthropodized*” by acquiring an external skeleton. (BNC AMM: 953)
- (2) She describes women, for example, as “*moronised*,” “*robotised*,” “*lobotomised*,” as “the puppets of Papa.” (BNC ECV: 1405)

### 3 Theoretical Perspectives

#### 3.1 Antilexical Approach

Our task in this section is to formalize *-ize* affixation, a major verb-forming process. Specifically, we will present pertinent syntactic structures, lexical entries, and subsidiary rules. Before proposing a new analysis, let us sketch a grammatical model on which our analysis relies.

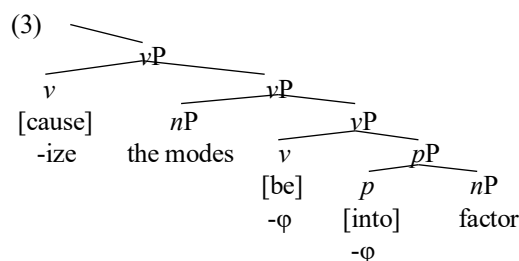
The properties of *-ize* derivatives observed above are best dealt with in the framework of antilexicalism. This thesis holds that major word formation processes take place outside the lexicon so that the creative aspects of sentence and word construction are uniformly captured in syntax (Halle and Marantz, 1994; Matushansky and Marantz, 2013). The creativity of *-ize* derivation substantiates the view that it is treated in syntax, but not in the lexicon, which is generally defined as a set of listed items. Thus, such a view has the merit of maintaining the homogeneity of the modules.

The present thesis also has the theoretical advantage of removing the *-ize* word formation rule from the lexicon by the independently established syntactic devices, whereby the related redundancy is expelled from the grammar completely. Moreover, an empirical advantage can be gained by adopting the antilexical approach. As indicated in section 2.6, the syntactic environments, anaphoric and enumerative, can be a major source of word creation. The spur-of-the-moment word composition in the syntactic contexts provides a constant supply for new labels like *arthropodize* and *moronise* and thus backs up the thesis of syntactic word formation.

#### 3.2 Underlying Structures

Let us consider the underlying structures concerning *-ize* words in the framework of antilexicalism. We follow Hale and Keyser’s view that the meaning of a complex word is primarily determined by the syntactic structure (Hale and Keyser, 1993). Thus, the converted verb *to shelve* is arguably derived by the head-movement of a noun (*shelf*) from an underlying structure such as  $[_{vP} -\phi [_{vP} \text{the book } [_{vP} -\phi(=\text{put}) [_{pP} -\phi(=\text{on}) \text{shelf}]]]]]$ , with abstract *v* and *p*. According to this view, the deadjectival *-ize* construction has the basic structure of  $[_{vP} v_{[\text{cause}]-\text{ize}} [_{vP} [_{nP} \text{the cell}] [_{vP} v_{[\text{be}]-\phi} [_{pP} \text{alkaline}]]]]]$ , where the underlying adjective *alkaline* is a predicative of the surface object *the cell* and the “small clause” is dominated by the causative *-ize*. Thus, the causative meaning of *they alkalinize the cell* can be readily obtained from the underlying configuration.

As observed in section 2.2, the meanings of denominal verbs are broadly divided into four types: (i) ‘make sth be into x’ (result, e.g. *factorize*), (ii) ‘make sth provided with x’ (providing/giving, *chemicalize*), (iii) ‘make sth be like x’ (simulative, *Beethovenize*), and (iv) ‘make sth be as x’ (function, *canonize*). Accordingly, the structure of the result-type will be as shown in (3):



The sentence *they factorize the modes* is then to be interpreted as meaning that they make the modes broken down into factors. The structure of the providing-type is essentially the same as that of the result-type:  $[_{vP} v_{[\text{cause}]-\text{ize}} [_{vP} [_{nP} \text{the dress}] [_{vP} v_{[\text{be}]-\phi} [_{pP} \text{with}-\phi [_{nP} \text{accessory}]]]]]$ . The only difference between the two types is that the providing-type involves the preposition *with* as opposed to *into*, so that the meaning of *they accessorize the dress* is something like ‘they make the dress accompanied by an accessory.’ Note that one of the main meanings of *with* is ‘accompanied by another person or thing.’

Similar remarks apply to the simulative-type and function-type of *-ize* derivatives. The former example *he Beethovenized Haydn’s minuet* has a

syntactic structure:  $[_{VP} v_{[cause]-ize} [_{VP}[_{NP} \text{Haydn's minuet}][[_{VP} v_{[be]-\phi}[_{PP} p_{[like]-\phi} [_{NP} \text{Beethoven}]]]]]]]$ , where the preposition *like* makes a difference in the way the base noun is characterized. From this follows the meaning: ‘he made Haydn’s minute like a work of Beethoven.’ To take the latter example, *they canonized the texts* has a configuration:  $[_{VP} v_{[cause]-ize} [_{VP}[_{NP} \text{the texts}][[_{VP} v_{[be]-\phi}[_{PP} p_{[as]-\phi} [_{NP} \text{canon}]]]]]]]$ . This type differs from others in that the preposition *as* is involved, so that the reading ‘they made the texts as a canon’ can readily be obtained.

There are two advantages of the present analysis. First, it can account for the meaning properties of derived verbs observed in section 2.2; the basic meaning and additional meanings of *-ize* verbs can be distinguished accurately. The basic one is ‘make y (be) in the state of x’ and this meaning is attributed to the core part of the *-ize* construction. The additional meanings are divided into five types according to what condition the surface object y is in. This is typically represented by the spatial and functional relations that are expressed by specific prepositions. Thus, the difference between the submeanings originates in the different prepositions in the core layer, whereby the submeanings can be related to each other.

The second advantage is that possible classes of *-ize* verbs can be predicted from our analysis: *-ize* verbs can only be transitives and ergative intransitives (inchoatives). Two cases in point can be recognized: unergative (intransitive) verbs do not engage in *-ize* affixation, as in *\*they dancize to rap music/I must journe(y)ize there*. This is because unergatives typically signify movement of animate entities and such a movement/action construction is not fitted to the predicative nature that *-ize* affixation involves. Note that converted verb may be a verb of this type (*they dance to rap music/I must journey there*), since verbal conversion does not necessarily involve predicative construction. Additionally, *-ize* derivatives of unaccusatives are illicit, as in *\*ethical problems will surfacize (=ethical problems will rise to the surface)/\*lower level of pollution will resultize*. An unaccusative (intransitive) verb expresses a phenomenon that happens spontaneously without the intervention of any causer, which is incompatible with the intentionality that *-ize* verbs imply.

### 3.3 Vocabulary Insertion

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). From the semantic and morphological properties identified in section 2, we can describe the internal features and selectional conditions of *-ize*: all five types of the suffix *-ize* have a common feature as verbalizer, yet each requires the base with a distinct feature. These descriptions can be formalized into the lexical entry on the basis of an underspecified model, as seen in (4).

- (4) *-ize*: (a)  $[V][[cause]$ , (b)  $+< vP_{[be]}$ ,  
 $a/p_{[into]}/p_{[with]}/p_{[like]}/p_{[as]}$ , Latinate/Greek>  
 Condition: predicative=[root (suf)]

The internal features of the affix are listed in (a) and its license environment is specified in (b). 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 and complement (cf. Emonds, 2000). The lexical entry *-ize* in (4) then designates something like ‘*-ize* makes a causative verb, adjoining to a “small clause” consisting of a subject and a predicative; the predicative is divided into five groups and they are all of Latinate or Greek origin.’ For instance, when *-ize* is inserted under the *v* node in the environment of predicative including  $p_{[into]}$ , the result-class of *-ize* derivative is obtained. The condition of predicative entails that compounds and prefixed words are ruled out as the base of *-ize*. The crucial point is that *-ize* verbs are freely coined as long as the affixation meets the licensing conditions, particularly those on the structure of the bases and their vocabulary strata.<sup>10</sup>

### 3.4 Subsidiary Rules

This section focuses on two kinds of auxiliary rules for *-ize* derivation. The first one is a “redundancy rule,” which eliminates the redundancy of item-by-item specification. As shown in section 2.3, there is an essential meaning correlation between *-ize* verbs and their bases. Confining discussion below to the resultative-type and ornative-type, we

<sup>10</sup>How to construct a word form from the corresponding syntactic representation will not be explored here. There are

two ways in which such a word is constructed: one is to use syntactic head-movement (Harley, 2009); the other is to use morphological merger (Marantz, 1996).

observe that *-ize* verbs with a sense of ‘resultative’ show a systematic tendency to be derived from nouns that designate <state>, <language>, <fundamental>, <system>, and <status>. <sup>11</sup> Similarly, ‘ornative’ verbs tend to stem from nouns indicating <(bio)chemical>, <presentable>, <academic>, <format>, and <rights>.

These generalizations can be formalized into the redundancy rules on vocabulary insertion, as demonstrated in (5) and (6). These rules essentially signify that a noun indicating state or quality and a noun expressing (bio)chemical are inserted under the sister node of  $p_{[into]}$  and that of  $p_{[with]}$ , respectively. Accordingly, the noun *harmony* is correctly inserted into the sister position of  $p_{[into]}$ , without having to specify that *harmony* is connected to  $p_{[into]}$ .

(5)  $n \rightarrow$  <state>, <language>, <fundamental>, <system>, <status> /  $p_{[into]}$  —

(6)  $n \rightarrow$  <(bio)chemical>, <presentable>, <academic>, <format>, <rights> /  $p_{[with]}$  —

The second subsidiary rule involves the truncation of a word-internal suffix. We have seen in section 2.5 that *-ize* affixation triggers the truncation of an intra-word suffix in the cases of *-ic*, *-ous*, *-ity*, *-ant*, *-ism*, and *-ive* while it may not trigger the truncation in the cases of *-al*, *-(i)an*, *-able*, *-(a)(t)ion*, *-ary*, and *-er*. Moreover, the suffix *-ic* proves to be intact in specific circumstances.

To adjust the morphological structure of *-ize* words, we propose a truncation rule in (7), which is operative in the PF component. <sup>12</sup> This morpheme-truncation rule entails that *-ic*, *-ous*, *-ity*, *-ant*, *-ism*, and *-ive* are deleted in *-ize* suffixation (cf. *aromatize*) but each of them is not deleted when its base is monosyllabic (cf. *classicize*) or ends in two consonants (cf. *romanticize*). <sup>13</sup>

(7) *-ic*, *-ous*, *-ity*, *-ant*, *-ism*, *-ive*  $\rightarrow$   $\emptyset$   
/ X\_\_ *-ize*

Condition: X=polysyllabic or ending in a single consonant

## 4 Conclusion

Based on detailed observation of the derived verbs discerned in a large-scale corpus, we have revealed the essential properties of verb derivation. Semantically, derived verbs are divided into five main groups and each submeaning is correlated with a base’s meaning. Formal restrictions are placed on the internal structures and vocabulary strata of the bases. As regards productivity, *-ize* affixation is creative in its construction of numerous innovated verbs. The above properties of derived verbs are theoretically accounted for; basic features common to all five submeanings follow naturally from a core part of their underlying structures. The productivity of *-ize* derivation also arises from its underlying syntactic configuration. Finally, formal restrictions on the bases and the base-derivative meaning correlation originate in the insertion conditions of vocabulary items.

A rigorous analysis of the formal restrictions and the semantic correlation awaits further investigation. Hopefully, we have shown that the study of word formation mechanism can be widely promoted by “corpus-based investigation.”

## Acknowledgments

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. 22K00562) from the Japan Society for the Promotion of Science.

## References

- Mark Aronoff. 1976. *Word Formation in Generative Grammar*. MIT Press, Cambridge, MA.
- Harald R. Baayen and Antoinette Renouf. 1996. Chronicling *the Times*: productive lexical innovations in an English newspaper. *Language*, 72:69-96.
- Eve V. Clark and Herbert H. Clark. 1979. When nouns surface as verbs. *Language*, 55:767-811.
- David Embick. 2010. *Localism versus Globalism in Morphology and Phonology*. MIT Press, Cambridge, MA.

<sup>11</sup>Angle brackets are used here for referring to semantic categories; <language> is intended to mean ‘something related to language.’

<sup>12</sup>See Aronoff (1976: 88-98) for arguments for truncation rules.

<sup>13</sup>Exceptional cases which do not seem to follow the rule are specified on an item-by-item basis. For example, that *-ic* in *aestheticize* is untruncated is specified in the lexical entry of *aesthetic*.

- Joseph E. Emonds. 2000. *Lexicon and Grammar: The English Syntacticon*. Mouton de Gruyter, Berlin.
- Ken Hale and Samuel J. Keyser. 1993. On Argument Structure and the Lexical Expression of Syntactic Relations. In K. Hale and S. J. Keyser, editors. *The View from Building 20: Essays in Linguistics in Honor of Sylvain Bromberger*, pages 53-109. MIT Press, Cambridge, MA.
- Morris Halle and Alec Marantz. 1994. Some key features of Distributed Morphology. *MIT Working Papers in Linguistics*, 21:275-288.
- Heidi Harley. 2009. Compounding in Distributed Morphology. In R. Lieber and P. Štekauer, editors. *The Oxford Handbook of Compounding*, pages 129-144. Oxford University Press, Oxford.
- Heidi Harley and Ralf Noyer. 2000. Formal versus Encyclopedic Properties of Vocabulary: Evidence from Nominalisations. In B. Peeters, editor. *The Lexicon-Encyclopedia Interface*, pages 349-374. Elsevier, Amsterdam.
- Jennifer Hay. 2003. *Causes and Consequences of Word Structure*. Routledge, New York.
- Otto Jespersen. 1949. *A Modern English Grammar on Historical Principles*, volume 6. George Allen and Unwin, London.
- Beth Levin and Malka Rappaport Hovav. 1995. *Unaccusativity: At the Syntax-Lexical Semantics Interface*. MIT Press, Cambridge, MA.
- Rochelle Lieber. 2005. English Word-Formation Processes. In P. Štekauer and R. Lieber, editors. *Handbook of Word-Formation*, pages 375-427. Springer, Dordrecht.
- Alec Marantz. 1996. 'Cat' as a phrasal idiom: consequences of late insertion in Distributed Morphology. ms., MIT.
- Hans Marchand. 1969. *The Categories and Types of Present-Day English Word-Formation: A Synchronic-Diachronic Approach*, 2<sup>nd</sup> ed. C. H. Beck, München.
- Ora Matushansky and Alec Marantz. 2013. *Distributed Morphology Today: Morphemes for Morris Halle*. MIT Press, Cambridge, MA.
- Ingo Plag. 1999. *Morphological Productivity: Structural Constraints in English Derivation*. Mouton de Gruyter, Berlin.
- Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.
- Sally Rice and Gary Prideaux. 1991. Event-packing: the case of object incorporation in English. *BLS*, 17:283-298.
- Elisabeth O. Selkirk. 1982. *The Syntax of Words*. MIT Press, Cambridge, MA.

# Some Notes on *p(e)re*-Reduplication in Bulgarian and Ukrainian: A Corpus-based Study

Ivan Derzhanski, Olena Siruk

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences

[iad58g@gmail.com](mailto:iad58g@gmail.com), [olebosi@gmail.com](mailto:olebosi@gmail.com)

## Abstract

We present a comparative study of *p(e)re*-reduplication in Bulgarian and Ukrainian, based on material from a parallel corpus of bilingual texts. We analyse all occurrences found in the corpus of close sequences and conjunctions of two cognate words, the second of which features the intensive and recursive prefix *pre-* (Bulgarian) or *pere-* (Ukrainian). We find that in Bulgarian this construction occurs more frequently with finite verb forms, and in Ukrainian with participles and nouns. There is also a correlation with the mode of action denoted by the prefix: in its intensive meaning it turns up more often in Bulgarian, in its recursive meaning in the two languages equally, and in Ukrainian there are more occasions where it cannot be identified as either intensive or recursive. Finally, in both languages instances of *p(e)re*-reduplication are most common, by a wide marge, in texts with Ukrainian originals.

**Keywords:** reduplication, intensive prefix, prefix *pre-*, prefix *pere-*, parallel corpus, Bulgarian language, Ukrainian language.

## 1 Introduction

The Proto-Indo-European root *\*per* ‘take, carry across or through’ (Pokorny 1959: 810) gave rise in Slavic to a preverb with a fundamental spatial meaning as well as a variety of derived meanings, all related to crossing a boundary or surpassing a degree, with the potential of combining with parts of speech other than the verb, too, as an elative marker:

*пръ-* expresses the idea of going beyond, surpassing: *пръити* ‘to cross, get over’, *пръстѣжити* ‘to transgress’, *прълитати* ‘to overflow’; and of transporting, transforming:

*пръселити* ‘to resettle’, *пръобразити* ‘to transfigure’. At the same time it is an intensifier which adjoins to adjectives, to nouns: *пръпогъбѣль* ‘complete perdition, πανωλεθρία’ and to verbs: *пръчюдиѣ сѧ* ‘being excessively astonished, ὑπερθαυμάσας’. (Vaillant 1948: 323)

These meanings persevere, by and large, in the contemporary Slavic languages. The details vary. In Bulgarian the recursive mode of action (‘redo, do again or in a new way’) appears to be the most prominent among the ones marked by the preverb *пре-*, followed by the majorative-resultative (or intensive: ‘do to a degree higher than the norm’) and the transgressive (‘do across an area’), with 96, 67 and 41 examples listed in (Ivanova 1974: 49ff), respectively. Bulgarian *пре-* does not mark the resultative-pancursive-distributive mode of action (‘do upon all available objects’), which is often expressed by its Ukrainian cognate *непе-* (Zhovtobrjukh 1979: 262f). On the other hand, in Ukrainian the inherited preverb *непе-* cedes the elative function almost entirely to the borrowed prefix *пре-* (ESUM 2003: 558), which operates mainly on adjectives and adverbs. It is also noted (Šerech 1959: 291f) that *непе-* tends to denote motion across and above, contrasting with the similar preverb *про-*, which indicates motion through the inside of an object, and this motivates its further evolution to a pancursive, majorative (intensive) and recursive marker.

The Bulgarian prefix *пре-* and the Ukrainian *непе-* play a key part in a phenomenon which we will call *p(e)re*-reduplication. It consists of the use in close succession of two cognate words (as a rule of the same part of speech and in the same grammatical form), the second of which is formed with the prefix *пре-* or *непе-* (in the two languages respectively), where the first has no prefix or has another. In general this pursues a rhetorical effect:



a concept is expressed twice with added emphasis the second time, which results in intensification:

- (1) [Bg] *Но тези роти вече, както личи, не са формирани от патилите и препатили войници, които текат закърпени от болниците* (O. Honchar, *The Standard Bearers*) ‘But now these companies are evidently not composed of those seasoned and overseasoned soldiers who stream, patched, from the hospitals.’
- (2) [Uk] *Це досить відверта посмішка жінки, яка бачила-перебачила.* (B. Raynov, *Don't Make Me Laugh*) ‘This is the rather brazen grin of a woman who has seen, and seen a lot’.

The device is especially typical of the language of folklore and of colloquial speech influenced thereby.

It may be tempting to say that this is simply the same construction serving the same purpose in two closely related languages. But this does not mean that its use is identical: there may be differences in the lexical categories most commonly involved, the details of the morphology and the syntax and perhaps other parameters. Such differences can only be established or disproven with the help of evidence drawn from corpora.

In this paper we present the results of a comparative study of *p(e)re*-reduplication in Bulgarian and Ukrainian based on material from a parallel bilingual corpus.

## 2 On the Corpus

The bilingual Bulgarian–Ukrainian corpus (CUB) (Siruk and Derzhanski, 2013; 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 version of CUB used in this research 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, Polish, Russian-1 (stories about the past and present) and Russian-2 (stories about the future)—approx. 2 million words in each of the ten sectors (in Bulgarian and Ukrainian counted together; for various reasons the ratio tends to be about 53:47);
- the Bible, in canonical translations from Church Slavonic into Bulgarian and from Hebrew, Aramaic and Greek into Ukrainian—1.1 million words.

The total size of the corpus is 21 million words (11.2 million in Bulgarian and 9.9 in Ukrainian). The Bible is aligned by verse, and the other texts (mostly) by sentence.

## 3 The results

A total of 130 instances of *p(e)re*-reduplication in one or both languages were found in the corpus, including 48 in Bulgarian only, 19 in both Bulgarian and Ukrainian and 63 in Ukrainian only.

We disregarded most occasions of *npe*-reduplication of adjectives or adverbs in Bulgarian, because we are interested in comparing Bulgarian *npe*- to its regular etymological counterpart in Ukrainian, which is *nepe*-, and for this particular purpose, as was said above, Ukrainian tends to also use *npe*-.

### 3.1 Distribution by part of speech

The items which compose the construction include finite verb forms or gerunds, participles, adjectives, nouns or pronouns. In Ukrainian it is expedient to handle invectives as a separate category: these are formally adverbs, pronouns or nouns, but used with no regard to their part of speech and original semantics: *Розтуди тебе перетуди* ‘And unprint thyself’ (E. Hemingway, *For Whom the Bell Tolls*), *Голій, таку-перетаку, коня прибери з вагу!* ‘Holiy, you so and so, take your horse off the platform!’ (V. Zemlyak, *Green Mills*), *Мать-перемать!* (A. and B. Strugatsky, *Roadside Picnic*; this invective is borrowed from Russian, which is why it involves Ru *мать* in lieu of Uk *мату* ‘mother’, but the pattern is the same).

The two words are of the same part of speech, except for a few instances where the first is an adjective and the second a participle; the

grammatical form is likewise the same, except for one occasion in Ukrainian when a gerund is combined with a finite verb form (*вибираючи, перебирає* ‘chose choicely’ in F. Nietzsche’s *Thus Spoke Zarathustra*).

Apart from invectives, the use of pronouns in *pere*-reduplication is also restricted to Ukrainian:

- (3) [Uk] *Я роду такого й перетаккого, мої предки те й перете зробили!* (G. Vossaccio, *The Decameron*) ‘I belong to the So-and-so family and my ancestors did such-and-such!’.

In addition, Table 1 attests that in Bulgarian this construction occurs more often with finite verb forms and in Ukrainian with participles and nouns.

	Bg only	Bg, Uk	Uk only	total
verbs	39	10	20	<b>69</b>
participles	3	7	13	<b>23</b>
adj. : part.	5	—	1	<b>6</b>
adjectives	—	—	1	<b>1</b>
nouns	1	2	17	<b>20</b>
pronouns	—	—	2	<b>2</b>
invectives	—	—	9	<b>9</b>
total	<b>48</b>	<b>19</b>	<b>63</b>	130

Table 1: Distribution by part of speech

Somewhat controversially, we have counted as an instance of *pere*-reduplication the Bulgarian adjective-participle compound *бяла-пребледняла* ‘white-blached’ (found in O. Kobylanska’s *On Sunday Morning She Gathered Herbs*); it is not one *stricto sensu*, as its parts are not even etymologically related, but they are phonetically and semantically similar, and also the writing of the whole as a hyphenated word, akin to *бледна-пребледняла* ‘pale-blached’ from the same book, argues in favour of such treatment.

With virtually identical frequency in the two languages – about 47.66% – the items forming the couple only differ in that the second one has the prefix *n(e)pe-* (notated as *p* in the formulae in Table 2). Alternatively, *n(e)pe-* can replace a prefix present only in the first item (*p°*); this is more common in Ukrainian (*закуска-перекуска* ‘hors d’œuvre snack’, *розказано й переказано* ‘told and retold’). Conversely, it is more common in Bulgarian for the items to differ in suffixes (*s*). In both languages the latter happens mostly because of the suffix it takes to reconvert the verb which has been perfectivised by the addition of

*n(e)pe-* back to the imperfective aspect (Bg *топлені и претопляни супи* ‘soups heated and reheated’, Uk *читає і перечитує* ‘reads and rereads’), but also when an adjective is coupled with a participle (Uk *старе-перестаріле* ‘old and overaged’). The co-occurrence of the two differences is predictably rare; there is only one example of this in our data, in Ukrainian: *Часті й тривалі перекури, розмови, перемовки* (V. Shishkov, *Gloomy River*) ‘Frequent and long smoking breaks, chats, talks’.

With the verb ‘read’ in Bulgarian another prefix (*про-*, notated as *p’* here) is also added (*четените и препрочетени книжки* ‘the books read and reread’, *чете и препрочита* ‘reads and rereads’). This happens 7 times in the corpus.

	Bg	Uk
<i>p°R-pR</i>	5 (7.46%)	23 (28.05%)
<i>R-p(p’)R</i>	32 (47.76%)	39 (47.56%)
<i>p°R-pRs</i>	—	1 (1.22%)
<i>R-p(p’)Rs</i>	30 (44.78%)	19 (23.17%)

Table 2: Derivational relationship between the two items in the couple

In Bulgarian in the absolute majority of cases the two items are linked by a conjunction; it is significantly rarer for them to be juxtaposed (or appear in juxtaposed phrases), which materialises as a comma in writing; and there are only three hyphenated compounds in our data, all of the adjective-participle type (*бледна-пребледняла* ‘pale-blached’ and *бяла-пребледняла* ‘white-blached’, mentioned above, and *пълно-препълнено* ‘full-overfilled’). In Ukrainian the distribution among the three categories is more balanced, but in both languages the preference is for the two items to be connected syntactically rather than morphologically:

	Bg	Uk
conj.	54 (81.82%)	46 (56.79%)
juxt.	9 (13.64%)	23 (25.93%)
hyph.	3 (4.55%)	14 (17.28%)

Table 3: Grammatical link between the two items

The first two of the options formulated here do not necessarily imply that the words need to be adjacent or only separated by a conjunction: there may be functional words interfering (up to three in

our material), less often content words, or the construction may appear in direct speech and be broken by the author's words:

- (4) [Bg] *Четох ги и ги препрочитах чак до сутринта* (P. Zahrebelnyi, *Let's Come to Love*) 'I read them and reread them until the very morning'.
- (5) [Uk] *А от ми зараз подивимося, хто кого дожене, хто кого пережене!* (A. and B. Strugatsky, *Roadside Picnic*) 'Now we'll see who catches up and who gets ahead!'
- (6) [Uk] — *Розтуди, — коротко сказав Агустін. — I перетуди.* (E. Hemingway, *For Whom the Bell Tolls*) "'Milk," Agustín said simply. "And milk again."

There is one example which doesn't fall easily into any of the three categories in either language, and is not counted in Table 3:

- (7) [Uk] *А я скочив — Дунай перескочив* (M. Stelmakh, *The Four Fords*) 'And I jumped and vaulted over the Danube' || [Bg] *Кога скочи — Дунава прескочи.*

### 3.2 Distribution by meaning of the prefix

The semantic relation between the two parts of the construction varies. By far most frequently, the meaning of the prefix is intensive or recursive, so the whole adds up to, literally, 'do and overdo' or 'do and redo', in either case conveying emphasis or intensity. Occasionally, however, the second (prefixed) word does not exist at all outside of this construction:

- (8) [Uk] *нехай вона в тебе буде і чесна, й перечесна — не зарікайся, що вона одна з усіх того не зробить* (G. Vossaccio, *The Decameron*) "'tis at least possible, that, however honest she be [*lit.* let her be honest and overhonest], she will do as others do',
- (9) [Uk] *Добре, туди їх перетуди, усіх фашистів* (E. Hemingway, *For Whom the Bell Tolls*) 'To obscenity with all fascism good' (*lit.* 'Well, thither and re-thither with them, with all fascists');

or is a close synonym of the first word:

- (10) [Uk] *Та конкуренція, конкуренція... нові винаходи, новіші винаходи... зміни, переміни. Світ мене обігнав* (C. Dickens, *Dombey and Son*) 'But competition, competition—new invention, new[er] invention—alteration, alteration—the world's

gone past me' (the original has three exact repetitions; the translator introduces gradation into two of them, one by a comparative degree and one by a *пере*-derivative which means the same as the word with *з*-, but the two together create an impression of waxing intensity);

or a less close synonym, so that the gradation is more clearly felt:

- (11) [Bg] *Струвалих ми се, че някой ме следи, че ме преследва, опитва се да ме хване...* (A. Christie, *They Do It with Mirrors*) 'I thought people were spying on me, watching me [*lit.* following me, pursuing me], trying to hound me down';<sup>1</sup>

or bears some other relation to the first word, such as being a transgressive derivative ('do from place to place'), a supergressive-resultative ('outdo someone else') or a finitive one ('finish doing')—modes of action which are also typical of the prefix *n(e)pe*- in one or both languages:

- (12) [Uk] *Четверо коліс каронади прокочувалося й перекочувалося по вбитих нею людях, шматуючи їх, кришачи й розриваючи* (V. Hugo, *Ninety-Three*) 'The four wheels of the carronade passed back and forth [*lit.* over and across] over the men it had killed, cutting, crushing and rending them' (the French original features the formally similar, but different in content, *passaient et repassaient* 'passed and passed again');
- (13) [Uk] *Люди дотримуються свого звичного побутового ритму, поки ми отут безглуздо наздоганяємо й переганяємо один одного* (B. Raynov, *Typhoons with Gentle Names*) 'People follow their usual schedule, while we here mindlessly overtake and surpass one another';
- (14) [Bg] *Люлякът в градинката на райкома цъфтя и прецъфтя, а нея все я няма и няма от Велики Устюг...* (V. Zemlyak, *The Swan Flock*) 'The lilac in the District Committee garden had shed its blossoms [*lit.* bloomed and finished blooming], but still she did not return from Velikiy Ustyug'.

There is a single example, in Ukrainian, of a non-deverbal noun with a derivative in which the

<sup>1</sup> The meanings of the verb *преследвам* range from 'follow, pursue' (shared with *слідя*) to 'persecute, haunt'; here the context argues that the more ominous meanings are not the ones intended (because a victim of persecution is very much aware of it), but the hearer is aware of their existence in the language, so they can contribute to the effect.

prefix *непе-* has a spatial meaning: *лісами та перелісками* ‘through forests and thickets’ (M. Stelmakh, *The Four Fords*).

Exceptionally the second item may bear no synchronically detectable relation to the first:

(15) [Bg] — *Намерила, та премерила — прихна той* (M. Stelmakh, *The Four Fords*) “‘She is insatiable,’ he snorted’ (*lit.* ‘She has found and measured’;<sup>2</sup> originally the words share a root, as per (Georgiev and Duridanov 1995: 484), but at present they are not perceived as being semantically akin);

or has a separate lexical (or even terminological) meaning, so that the use of the two words in succession is not a rhetorical device, but – because of the similarity to a familiar one – may have a similar effect:

(16) [Uk] *Деякий час маленький загін ішов піскуватими ґрунтами, що утворились із скалок двійчастих черепашок і висхлих кісток, з великою домішкою закису й перекису заліза* (J. Verne, *In Search of the Castaways*) ‘For a part of the day, the little troop trod a sand composed of debris of bivalve shells and cuttlefish bones, and mixed in a great proportion of iron protoxide and peroxide’ (the French original has *une grande proportion de peroxyde et de protoxyde de fer*, but the translator has reversed the order, thus achieving, consciously or otherwise, outward similarity with the *p(e)re*-construction),

(17) [Uk] *Були ще й інші сходи та переходи, якими ніхто не ходив цілими тижнями* (C. Dickens, *Dombey and Son*) ‘There were other staircases and passages where no one went for weeks together’.<sup>3</sup>

The frequency of the construction in the two corpus languages correlates with the semantics of the prefix: in its intensive meaning it turns up more often in Bulgarian (which harmonises with

the fact that in Ukrainian this meaning has been partly taken over by the South Slavic loan *непе-*), in its recursive meaning approximately equally in the two languages, and in Ukrainian there are more cases where it cannot be identified as either intensive or recursive. This is summarised in Table 4.

	Bg only	Bg, Uk	Uk only	total
intensive	19	5	1	<b>25</b>
recursive	19	13	25	<b>57</b>
miscell.	10	1	37	<b>48</b>
	<b>48</b>	<b>19</b>	<b>63</b>	130

Table 4: Distribution by semantics of the prefix

### 3.3 Distribution by source language

It is known that in their choice of wording translators are prone to being influenced by constructions used in the original. Since the use of reduplication for emphasis is universal, this can be expected to happen here as well.

Table 4 attests that *p(e)re*-reduplication is much more frequent in original Ukrainian texts and their Bulgarian translations than in any other texts in the corpus.

	Bg only	Bg, Uk	Uk only	total
Bg	3	2	3	<b>8</b>
De	2	1	2	<b>5</b>
E1	2	—	7	<b>9</b>
E2	1	—	5	<b>6</b>
Fr	2	2	8	<b>12</b>
It	8	—	6	<b>14</b>
Pl	2	—	1	<b>3</b>
R1	4	—	9	<b>13</b>
R2	1	1	6	<b>8</b>
Uk	14	13	15	<b>42</b>
Bible	9	—	1	<b>10</b>
	<b>48</b>	<b>19</b>	<b>63</b>	130

Table 5: Distribution by source language

When *p(e)re*-reduplication appears in a corpus text, the original (if different) may

- (I) use an analogous reduplicative construction with a prefix with similar semantics on the second item. Such are German intensive *über-*

<sup>2</sup> Along with the idiom *намерил съм, та съм премерил* ‘to have found and measured’ there exists the similar one *намерил съм, та съм се прехласнал* ‘to have found and become entranced’ (Nicheva et al. 1974: 644f); the latter makes more literal sense and so is likely to be the original variant, from which the former is derived by copying the root of the first word into the second, giving the whole the shape of a *pre*-reduplicated construction.

<sup>3</sup> Apart from meaning ‘staircase’, *сходи* means ‘ascents; descents’, *перехід* (pl. *переходи*) likewise means ‘passing’ as well as ‘passage, corridor’, so in the translation there are two ways in which the words are cohyponyms; this enhances their perception as more than two words with their regular meanings which happen to occur in sequence in the text.

and recursive *wieder*-,<sup>4</sup> French *re*-, Italian *ri*-, Russian *пере*-,

- (II) repeat a word exactly or with a different kind of modification (as when Bg *питаха, разпитаха* ‘they asked and inquired’ in Elin Pelin’s *Yan Bibiyan on the Moon* is translated as Uk *питали й перепитували*, or Bg *бледна-пребледняла* ‘pale-blanché’ and *бяла-пребледняла* ‘white-blanché’ serve to render Uk *біла-біліська* ‘white-white[diminutive]’ in O. Kobylianska’s *On Sunday Morning She Gathered Herbs*),
- (III) not involve repetition at all.

Table 6 demonstrates that Bulgarian translators from Ukrainian use *p(e)re*-reduplication nearly as eagerly as Ukrainian writers: of the 28 occurrences of the phenomenon in original Ukrainian prose they have only kept a little less than half (13), but have contributed a little more than that (4+10=14), ending up with approximately the same number. (Curiously, the same can be said to have happened in the translations in the opposite direction, only the numbers are smaller there.)

	Bg			Uk				
	I	II	III	I	II	III		
Bg	5	—	—	5	2	1	2	5
De	2	—	1	3	1	1	1	3
E1	—	—	2	2	—	1	6	7
E2	—	—	1	1	—	—	5	5
Fr	4	—	—	4	5	2	3	10
It	2	1	5	8	—	—	6	6
Pl	—	—	2	2	—	—	1	1
R1	—	—	4	4	5	—	4	9
R2	1	—	1	2	6	—	1	7
Uk	13	4	10	27	28	—	—	28
Bible	—	7	2	9	—	1	—	1
	27	12	28	67	47	6	29	82

Table 6: Distribution by the presence of reduplication in the original language

On 4 occasions in Bulgarian translations from French and on 6 in Ukrainian ones, the original features a similar construction with the prefix *re*-.

<sup>4</sup> There is one occurrence of each of these in F. Nietzsche’s *Thus Spoke Zarathustra: Wie er sie schlingt und kaut und wiederkaut!* ‘How it swalloweth and cheweth and recheweth them!’ > Bg *Как само ги налага и дъвче, и предъвква!* || Uk *Як вона душить її, жує й пережовує!*; *sie schwellen und überschwellen von Mitleiden* ‘they swelled and o’erswelled with pity’ > Bg *те се издуваха и преиздуваха от състрадание.*

Also, on 11 occasions the Ukrainian construction renders its materially identical Russian analogue. In Bulgarian this only happens once, but on 7 occasions in the translation of the Bible there is a kind of reduplication (albeit not of the same form) in the Church Slavonic (as well as the Ukrainian) text, which in turn follows literally the Hebrew or Greek original:

(18) [Bg] *Аз ще благословя и преблагословя, ще размножа и преумножа твоето семе* || [Uk] *благословляючи, Я поблагословлю тебе, і розмножуючи, розмножу потомство твоє* || [He] *kī-bārēk ’ābārēkkā, wə-harbāh ’arbeh ’et-zar’ākā* ‘in blessing I will bless thee, and in multiplying I will multiply thy seed’ (Gn 22:17);

(19) [Bg] *наистина ще те благословя и преблагословя, ще те размножа и преумножа* || [Uk] *Поблагословити Я конче тебе поблагословлю, та розмножити розмножу тебе!* || [Gk] *ἤ μὴν εὐλογῶν εὐλογήσω σε καὶ πληθύνων πληθυνῶ σε* ‘Surely blessing I will bless thee, and multiplying I will multiply thee’ (Heb 6:14).

Finally, it is remarkable that none of the few uses of *p(e)re*-reduplication in translations from Polish reflect a similar construction in the original; expressions such as *myślał i przemyślał* ‘thought and rethought’ (cf. Bg *мислил и премислял*) are not totally alien to that language, but evidently are much less used than in the other Slavic languages in the corpus.

#### 4 Conclusions

The constructions are similar indeed, but when it comes to actual use, they differ in many points, as we have seen: the parts of speech involved most commonly (predominantly verbs in Bulgarian and nouns more often – and exclusively, pronouns and a separable category of invectives – in Ukrainian), the interpretation of the prefix (intensive mostly in Bulgarian, transgressive etc. in Ukrainian), the derivational models (a distinctive prefix on the first item being more typical of Ukrainian), the grammatical link between the two items (with strong preference for a conjunction in Bulgarian). These can be explained in part by the presence of the borrowed prefix *npe*- in Ukrainian, which has relieved *npe*- of some of its functions, especially in the literary language. But since we work with fiction, and mostly with translated texts, there is an

occasion for examining the impact of the original languages and the translators' attitudes to using the target languages' vernacular constructions.

The material for this study was collected by a semi-automatic search in a bilingual corpus of aligned text. As the corpus is continually evolving, this raises the question of enriching it with appropriate alignment which would facilitate such research.

## References

Ivan Derzhanski and Olena Siruk. 2019. The Intensifying Prefix *pre-* in a Corpus of Bulgarian and Ukrainian Parallel Texts. In R. Pavlov and P. Stanchev (eds), *Digital Presentation and Preservation of Cultural and Scientific Heritage*, vol. 9, pages 177–188. Sofia: Institute of Mathematics and Informatics—BAS. [http://dipp.math.bas.bg/images/2019/177-188\\_12\\_2.10\\_fDiPP2019-67\\_f\\_v.1a.F\\_20190908.pdf](http://dipp.math.bas.bg/images/2019/177-188_12_2.10_fDiPP2019-67_f_v.1a.F_20190908.pdf).

ESUM. 2003. *Etymolohičnyj slovnyk ukrajins'koji movy. T. 4*. Kyjiv: Naukova dumka.

Vladimir Georgiev and Ivan Duridanov (eds). 1995. *Bâlgarski etimologičen rechnik, t. 4*. Sofiya: Akademichno izdatelstvo „Marin Drinov”.

Kalina Ivanova. 1974. *Nachini na glagolnoto deystvie v sâvremenniya bâlgarski ezik*. Sofiya: Izdatelstvo na BAN.

Keti Nicheva, Siyka Spasova-Mihaylova, Kristalina Cholakova. 1974. *Frazeologičen rechnik na bâlgarskiya knizhoven ezik*. Sofiya: Izdatelstvo na BAN.

Julius Pokorny. 1959. *Indogermanisches etymologisches Wörterbuch. B. III*. Bern, München: Francke Verlag.

Olena Siruk and Ivan Derzhanski. 2013. Linguistic Corpora as International Cultural Heritage: The Corpus of Bulgarian and Ukrainian Parallel Texts. In R. Pavlov and P. Stanchev (eds), *Digital Presentation and Preservation of Cultural and Scientific Heritage*, vol. 3, pages 91–98. Sofia: Institute of Mathematics and Informatics—BAS. <https://dipp.math.bas.bg/images/2013/091-098-f1-DiPP2013-6-Siruk-et-al.pdf>.

Jury Šerech. 1951. *Narys sučasnoji ukrajins'koji literaturnoji movy*. Mjunxen: Molode žyttja.

André Vaillant. 1948. *Manuel du vieux slave, t. I: Grammaire*. Paris: Institut d'Études slaves.

Mykhajlo Zhovtobryukh (ed.). 1979. *Slovotvir suchasnoji ukrajins'koji literaturnoji movy*. Kyjiv: Naukova dumka.

# An Open-Domain QA System for e-Governance

Radu Ion, Andrei-Marius Avram, Vasile Păiș, Maria Mitrofan,  
Verginica Barbu Mititelu, Elena Irimia and Valentin Badea

Research Institute for AI “Mihai Drăgănescu”

13 “Calea 13 Septembrie”

Bucharest 050711, Romania

{radu, andrei.avram, vasile, maria}@racai.ro

{vergi, elena, valentin.badea}@racai.ro

## Abstract

The paper presents an open-domain Question Answering system for Romanian, answering COVID-19 related questions. The QA system pipeline involves automatic question processing, automatic query generation, web searching for the top 10 most relevant documents and answer extraction using a fine-tuned BERT model for Extractive QA, trained on a COVID-19 data set that we have manually created. The paper will present the QA system and its integration with the Romanian language technologies portal RELATE, the COVID-19 data set and different evaluations of the QA performance.

**Keywords:** BERT fine-tuning, open-domain QA, Romanian, TEPROLIN, COVID-19.

## 1 Introduction

According to [Zhu et al. \(2021: 1\)](#), open-domain Question Answering (QA) has the ability “to answer a given question without any specified context”, by searching for the relevant documents on the web and extracting the relevant answer from one (or more) of the retrieved documents. In contrast, Machine Reading Comprehension “aims to enable machines to read and comprehend specified context passage(s) for answering a given question” which entails that, given a question and one (or more) passage(s) of text that (can) contain the answer, the QA system is able to identify it in the given text piece. The “open-domain” designation of a QA system also pertains to the ability of the system to answer factoid questions (factual questions) from any domain, according to [Lewis et al. \(2020: 1\)](#).

The QA system that is presented in this paper is “open-domain” from both points of view: it only takes the input question and automatically searches for the relevant documents on the web but, for the answer selection, it employs a fine-tuned

BERT model for Extractive QA that, using the input question together with the snippet that the web search engine produces for each relevant document, highlights the answer to the input question. Although we present an instance of this system for the COVID-19 domain, given other targeted data sets, the exact same pipeline can be applied to answer questions from those domains.

The QA system was developed in the European project [Enrich4All<sup>1</sup>](#), a project aiming at a Digital Single Market strategy, which is linked with lowering language barriers for online services and public administration procedures. The architecture of the QA system enables it to answer administrative questions about a certain topic (e.g. COVID-19, construction permits, etc.) that citizens may have for public authorities, by automatically searching for relevant documents on the public authority web site. Being available 24/7, it has the potential to reduce the administrative burden for public authorities.

In what follows, we present related approaches to open-domain QA in [Section 2](#), followed by a description of the COVID-19 data set in [Section 3](#). [Section 4](#) details the fine-tuning of different Romanian BERT models to COVID-19 Extractive QA, while [Section 5](#) describes the architecture and the underpinnings of the open-domain QA system. We end the paper with [Section 6](#) devoted to the evaluation of the QA system and [Section 7](#) presenting concluding remarks and future work plans.

## 2 Related work

Open-domain QA (ODQA) aims at answering questions from large open-domain corpora (e.g., Wikipedia). Recent success in this field mainly comes from fine-tuning and improving the pre-trained LMs, like ELMo ([Peters et al., 2018](#)) and

<sup>1</sup><https://www.enrich4all.eu/>



BERT (Devlin et al., 2018).

Wang et al. (2019) proposed a multi-passage BERT model to globally normalize answer scores across all passages of the same question, enabling the QA model to find more precise answers utilizing more text passages. Splitting articles into passages with the length of 100 words improved performance by 4% and trained on the OpenSQuAD data set, the model gained 21.4% EM and 21.5% F1 over all non-BERT models, and 5.8% EM (exact match) and 6.5% F1 over BERT-based models.

Yang et al. (2019) integrated BERT with the open-source Anserini information retrieval toolkit. They showed that combining a BERT-based reader with passage retrieval using the Anserini IR toolkit yields towards an improvement in question answering directly from a Wikipedia corpus. During training, passages corresponding to the same question are taken as independent training instances. The authors report that fine-tuning pre-trained BERT with SQuAD is sufficient to achieve high accuracy in identifying answer spans.

Lee et al. (2019) showed that it is sub-optimal to incorporate a standalone IR system in an OpenQA system, therefore they developed and they develop an OpenRetrieval Question Answering system (ORQA) system that treats the document retrieval from the information source as a latent variable and trains the whole system only from question-answer string pairs based on BERT. The system was evaluated on open versions of five QA data sets and outperformed BM25 model by up to 19 points in exact match.

Karpukhin et al. (2020) used BERT pre-trained model (Devlin et al., 2018) and a dual-encoder architecture (Bromley et al., 1993) in order to develop a training scheme that uses a relatively small number of question and passage pairs. The authors demonstrated that by fine-tuning the question and passage encoders on existing question-passage pairs the system outperformed models, such as TF-IDF or BM25 and also that applying a reader model to the retrieved passages leads to comparable or better results on multiple QA data sets in the open-retrieval setting. Furthermore, the study showed that in the context of open-domain question answering, a higher retrieval precision translates to a higher end-to-end QA accuracy.

Guu et al. (2020) used contextualized word representations to predict a span as answer. The authors showed the effectiveness of Retrieval-Augmented

Language Model pre-training (REALM) by fine-tuning on the task of ODQA. The system outperformed previous methods by a significant margin (4-16% absolute accuracy), and also provided qualitative benefits such as interpretability and modularity.

Yamada et al. (2021) introduced Binary Passage Retriever (BPR), a memory-efficient neural retrieval model that integrates a learning-to-hash technique into a Dense Passage Retriever (DPR) (Karpukhin et al., 2020). BPR has two main objectives: to generate efficient candidates based on binary codes and re-ranking based on continuous vectors. When compared with DPR, BPR reduced the memory cost from 65GB to 2GB without a loss of accuracy.

### 3 The COVID-19 data set

The COVID-19 data sets we designed are a small corpus and a question-answer data set. The targeted sources were official websites of Romanian institutions involved in managing the COVID-19 pandemic, like The Ministry of Health, Bucharest Public Health Directorate, The National Information Platform on Vaccination against COVID-19, The Ministry of Foreign Affairs, as well as of the European Union. We also harvested the website of a non-profit organization initiative, in partnership with the Romanian Government through the Romanian Digitization Authority, that developed an ample platform with different sections dedicated to COVID-19 official news and recommendations. News websites were avoided due to the volatile character of the continuously changing pandemic situation, but a reliable source of information was the website of a major private medical clinic, that provided detailed medical articles on important subjects of immediate interest for the readers and patients, like immunity, the emergent treating protocols, or the new variants of the virus.

Both the corpus and the question-answer data set were manually collected and revised. Data was checked for grammatical correctness and missing diacritics were introduced. The corpus is structured in 55 UTF-8 documents and contains 147,297 words.

The question-answer data set comprises 185 entries made up of a label (see the list of labels below), a question and its answer. The questions have been multiplied manually: rephrasing techniques have been used (such as active-passive constructions,

personal-impersonal ones, constructions with or without (epistemic and/or deontic) modality, synonyms, antonyms, hypernyms, etc.), always making sure the question’s meaning is not altered, so that the existing answer could be appropriate for each of the resulted questions. All these diverse ways of expressing the same question are meant to serve as a train base for the BERT model so that it can recognize alternative ways of inquiring about a certain topic.

Each entry in the question-answer data set was associated with one of the labels: covid-spread, covid-symptoms, covid-treatment, covid-vaccination, covid-logistics, covid-passport, covid-testing and covid-others.

To use the BERT model as a QA system (see Section 4), we had to manually mark the relevant answer to the question in the provided answer paragraph. An example of such an entry in the question-answer data set is given below: three different formulations of the same question, all marked with “Q:”, synonyms for a word or phrase enumerated within the question, using square brackets and slashes, and finally, the enlarged answer comes last, marked with “A:”, in which the more to-the-point answer is marked using square brackets as well:

L: covid-spread

Q: Vremea caldă [**previne/ne ferește de/ne protejează de**] infectarea cu Coronavirus? (“Does the warm weather [**prevent the/keep us safe from the/protect us against**] infection with the Coronavirus?”)

Q: Vara putem să [**facem/ne îmbolnăvim de**] COVID-19? (“Can we [**catch/get sick with**] COVID-19 in the summer?”)

Q: Dispare covidul [**pe vreme caldă/vara/la soare/la temperaturi mari**]? (“Does Covid vanish [**in warm weather/in the summer/in the sun/at high temperatures**]?”)

A: Datele existente arată că [**infecția poate fi dobândită în toate zonele climatice, inclusiv în cele calde**]. (“Existing data shows that [**the infection can be acquired in all climates, including the warm ones**].”)

Starting from the example above we can generate  $3 \cdot 2 \cdot 4 = 24$  different, but semantically equivalent formulations of the question “Does the warm weather protect us from the Coronavirus?” to which the answer is highlighted in the answer para-

graph: “Existing data shows that [**the infection can be acquired in all climates, including the warm ones**].” Having question-answer data points annotated in this way, we were able to automatically generate a SQuAD 2.0 data set (Rajpurkar et al., 2018) on which we fine-tuned a Romanian BERT model for Extractive QA, as described in the next section.

The COVID-19 SQuAD 2.0 data set<sup>2</sup> contains 1,388 question-answer data points, after automatically expanding all possible question formulations as described above. Since each question entry has multiple formulations, in order to be fair to the Extractive learning model, we randomly set aside 10% of a question alternative formulations for the test set (if there were less than 10 alternative formulations for a question, we kept a single formulation for the test set). This selection procedure gave us 180 question-answer data points in the development set and 1,208 question-answer data points in the training set, which represents a 13%-87% split of the data.

#### 4 Fine-tuning BERT for COVID-19 Extractive QA

To create the QA model, we employed the standard BERT fine-tuning procedure described in (Kenton and Toutanova, 2019) that consists of putting two feed-forward layers on top of the contextualized embeddings to predict the start and the end of an answer. On a more granular level, this operation is equivalent to taking the dot product between either a start vector  $S$  or an end vector  $E$  and each of the contextualized embedding  $T_i$  produced by the BERT model, and then applying the softmax function over the results:

$$P(start_i) = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}} \quad (1)$$

$$P(end_i) = \frac{e^{E \cdot T_i}}{\sum_j e^{E \cdot T_j}} \quad (2)$$

where  $i$  is the index of the contextualized embedding. Then we select as answer to a question the span from  $i$  to  $j$  that maximizes the  $S \cdot T_i + E \cdot T_j$ , and that satisfies  $j > i$  and  $j - i < \xi$ , where  $\xi \in \mathbb{N}$  is a tunable hyperparameter that controls the maximum number of tokens admitted in a span.

<sup>2</sup><https://github.com/racai-ai/e4a-covid-qa/tree/master/data>

Model	Exact %	F1 %
BERT-base-ro-uncased	71.33	<b>77.25</b>
BERT-base-ro-cased	<b>73.33</b>	76.75
RoBERT-small	58.00	61.64
RoBERT-medium	59.33	63.06
RoBERT-large	61.00	64.12
Distil-BERT-base-ro	51.33	70.39
Distil-RoBERT-base	55.33	61.72
DistilMulti-BERT-base-ro	51.33	70.39

Table 1: Results of the Romanian BERT models on our QA task

We fine-tuned eight Romanian BERT models on the question-answer data set introduced previously:

- **BERT-base-ro-uncased** and **BERT-base-ro-cased**: the cased and uncased versions of the first Romanian BERT (Dumitrescu et al., 2020).
- **RoBERT-small**, **RoBERT-medium** and **RoBERT-large**: the second iteration of Romanian BERTs introduced in (Masala et al., 2020). In comparison with the first BERT models, the authors introduce a large and a small version, trained on different corpora.
- **Distil-BERT-base-ro**, **Distil-BERT-base-ro** and **DistilMulti-BERT-base-ro**: distilled version of Romanian BERTs (Avram et al., 2021). The first variant was obtained by distilling the knowledge of BERT-base-ro-cased, the second of RoBERT-base and the last of both BERT-base-ro-cased and RoBERT-base.

We trained each model for 5 epochs, using a batch size of 8, a learning rate of  $3 \cdot 10^{-5}$ , a weight decay of  $10^{-3}$  and a maximum span  $\xi = 30$  of tokens. We used the train-test split of the data set that was described in Section 3.

The results are depicted in Table 1 where we outline the capabilities of the tested models to find the exact answer (i.e. Exact %), and the overlap percentage between the predicted and the true spans (i.e. F1 %). The highest exact score of 73.33% was obtained by BERT-base-ro-cased and the highest F1-score by BERT-base-ro-uncased with 77.25%. As it can be observed, there is a significant difference in performance between the RoBERT and BERT-base-ro variants. While the size of the test data is small (180 question-answer pairs), which could emphasize a disproportionate difference due

to selection bias, another, more likely explanation could hold: the BERT-base-ro vocabulary contains 50K word pieces while RoBERT vocabulary contains 38K word pieces, which puts BERT-base-ro in a better position to cover COVID-19 vocabulary, thus making the fine-tuning process more successful.

## 5 The open-domain QA pipeline

The trainable, open-domain QA system<sup>3</sup> executes the following operations, in sequence, for an input question:

- **Question processing**: the question string is run through the TEPROLIN Romanian text processing web service<sup>4</sup> (Ion, 2018) to obtain tokenization, lemmatization and dependency parsing annotation.
- **Query generation**: the question is analyzed to see which words are useful to form a query for the web search engine. Our web search engine of choice is Microsoft’s Bing search engine<sup>5</sup> because it is one of the few that offers API-based querying and offers access to the search hits via a JSON object containing text snippets and URLs of the relevant documents. Subsection 5.1 details how the query is formed from the processed version of the input question.
- **Answer mining**: entailing web search results re-ranking and answer highlighting, for each hit of the Bing web search engine (out of the total 10 that we ask for), we call the previously described, fine-tuned QA BERT model with the input question and the Bing-found text snippet (see the red rectangle in Figure 1) as the question context and get back a highlight of the answer (within the snippet) that BERT model thinks is a right fit for the input question. The highlighting comes with a confidence measure, topping at 1 for certainty and going towards 0 when the confidence level drops. If the hit has rank  $r$ ,  $0 \leq r < 10$  provided by Bing and BERT’s model confidence in highlighting the correct answer is  $c$ ,

<sup>3</sup><https://github.com/racai-ai/e4a-covid-qa>

<sup>4</sup><https://relate.racai.ro/index.php?path=teprolin/complete>

<sup>5</sup><https://www.microsoft.com/en-us/bing/apis/bing-web-search-api>

$0 \leq c \leq 1$ , then the combined confidence of the answer is

$$q = c \cdot \frac{10 - r}{10} \quad (3)$$

### 5.1 Query generation

The query generation algorithm takes the processed input question and produces a list of query terms for the Bing search engine. The question has been processed with the TEPROLIN text processing web service and we have, thus, lemmas, parts of speech and dependency information for each token in the input question.

We have experimented with three query generation procedures:

- **The baseline algorithm:** just take the input question as it is and feed it to the Bing search engine.
- **The content word selection algorithm:** take all nouns, numerals, verbs, adjectives and adverbs, in the order they appear in the question and form a maximally matching disjunctive query, e.g. for three terms  $t_1, t_2, t_3$  the query is " $t_1 t_2 t_3$ ". Very frequent Romanian verbs are not included in the query, such as "a avea" (to have), "a fi" (to be), "a exista" (to exist), "a face" (to do), etc.
- **No diacritics content word selection algorithm:** the content word selection algorithm query from which we automatically remove the Romanian diacritics. We see that Romanian pages are written with or (more likely) without the proper diacritics and thus, we have to accommodate query terms with and without diacritics.

While Bing works surprisingly well with the baseline query algorithm (the input question), by empirical experimentation we find that results from the union of the content word selection algorithm and its "no diacritics" version provide a better ranking for the most relevant documents. For instance, for the question from Figure 1 ("Do I need the COVID certificate to be allowed in the mall?"), the generated query is "nevoie certificatul verde intrarea mall" which produces 2 relevant documents on the first page, while the default query with just the input question yields a single relevant document on the first page. Furthermore, Bing does not

filter out Romanian functional words, e.g. "pentru" (for), and considers them to be relevant (easily seen because these are in bold in the returned snippets).

### 5.2 Integration with RELATE

RELATE (Păiș et al., 2020; Păiș, 2020) is a modular platform allowing access through a web based interface to multiple natural language processing applications for Romanian language. It follows, in a simpler way, the European Language Grid<sup>6</sup> philosophy of integrating components based on a micro-services architecture. In this context, the developed QA system was first exposed as a JSON REST API. This allowed it to be easily integrated in the RELATE platform and thus it became accessible through the platform's web front-end.

The QA interface<sup>7</sup> allows the user to enter a question, select the desired model<sup>8</sup> and then pass the input data to the system. Results are displayed in the form of text snippets extracted from various Internet sources. The actual answer is highlighted in the text snippet and the user is given the opportunity to access the source web site associated with the snippet. Finally, the user can return to the previous page in order to ask a new question. The output of the QA interface is presented in Figure 2.

## 6 Evaluation

To fairly evaluate our QA system, we have developed a new test set, containing 65 COVID-19 related questions, that were created to be different from the ones in the data set presented in Section 3. Thus, each of the authors of this paper independently recorded the most 10 interesting questions, from a personal point of view, mimicking real usage of the system. The intent behind this decision was to evaluate our QA system in real-world scenario where users may ask all sorts of questions, using Romanian diacritics or not, about a very rapidly evolving subject such as COVID-19. We have incrementally developed our COVID-19 data set in a time frame of about 6 months, in which time some understudied or evolving aspects of COVID-19 (e.g. the duration of the vaccine-induced immunity, the different vaccines efficiencies or the number of days the infected persons are quarantined) have

<sup>6</sup><https://www.european-language-grid.eu/>

<sup>7</sup><https://relate.racai.ro/index.php?path=qa/demo>

<sup>8</sup>Currently, a single model is available.



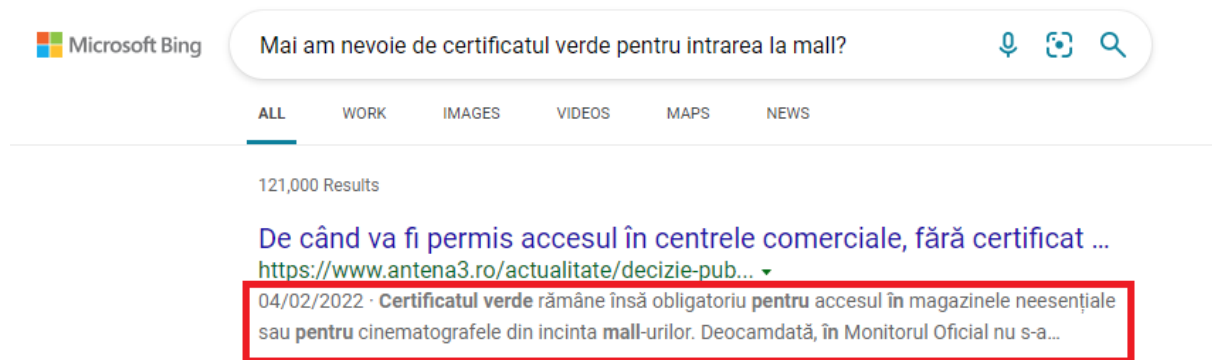


Figure 1: Bing search results

Question Answering

Intrebare noua

**Question:** *Mai am nevoie de certificatul verde pentru intrarea la mall?*

**Answers:**

Certificatul verde rămâne însă obligatoriu pentru accesul în magazinele neesențiale sau pentru cinematografele din incinta mall-urilor. Deocamdată, România și pentru a evita carantina și nu s-a publicat nici ordonanța care aduce modificări formularului de localizare a pasagerilor. [SURSA](#)

Guvernul schimbă regulile. În ce condiții se poate intra la mall, chiar fără certificat verde, și ce documente trebuie complete la intrarea în țară Liber p sfârșit voie și cei fără certificat verde. Însă doar în anumite condiții. [SURSA](#)

Figure 2: Question answers

changed, as results of ongoing studies were published that shed new light on these aspects.

We have found out that, for most of the questions, Bing retrieves more than one document containing relevant information for the answer, even if not spot on. This is because many questions in our newly developed test set are open-ended and opinions about the correct answer vary. There were questions with contradictory answers that were being marked as correct since it is not the duty of (this) QA system to infer the correct answer to a input question, but merely to present likely options as possible answers. One such example is the question “Ajută vitamina D la prevenirea COVID?” (“Does vitamin D help preventing COVID infections?”) for which we find that either “Lipsa vitaminei D crește riscul de infecție, inclusiv cu SARS-Cov-2.” (“The lack of vitamin D causes a rise in the risk of a SARS-Cov-2 infection.”) or “... dovezile existente nu susțin eficacitatea vitaminei D pentru tratarea virusului Covid-19.” (“... existing evidence does not support the efficiency of vitamin D in treating COVID-19”).

Consequently, our 65-question test set contains,

for each question, a list of the URLs of the documents containing the relevant and recent (if necessary) answer. For each URL, we retain the Bing-extracted text snippet in which we manually highlight one or more likely answers. For 10 questions out of 65, our automatic query generation algorithm did not find any suitable documents. We can thus evaluate the effectiveness of the content word selection query generation algorithm at  $1 - \frac{10}{65} \approx 85\%$ .

Table 2 presents the results of the QA system, on the test set presented above, using the baseline query generation algorithm (the input question itself) vs. the content word (CW) selection query generation algorithm. We compute the following:

- **Mean Reciprocal Rank (MRR)** of the returned documents: if multiple documents contain relevant information, we choose the one with the highest ranking to contribute to the MRR. The text snippets retrieved by Bing are re-ranked using Equation 3, which provides the order in which the user sees the results.
- **Exact answer matching:** percentage of BERT highlighted answers that exactly match

Query gen.	MRR	Exact %	F1 %
Baseline query gen.	0.4056	33.85	65.07
CW query gen.	0.5337	50.77	76.69

Table 2: Results of the QA system on the new test set

a human highlighted answer in the test set. Because our QA system only highlights a single answer in the returned text snippet, if the question has multiple answers that are annotated as correct, we test each annotated answer for an exact match.

- **F1 overlap matching:** if an exact match does not exist between the BERT answer and any of the annotated answers, we find the annotated answer that has the highest overlap with the BERT answer and compute the F1 measure of the overlapped characters.

Table 2 shows convincingly that the content word selection query generation algorithm is much better than using the input question as the Bing query. A MRR of more than 0.5 shows that the user sees the snippet containing a likely answer in the top two results returned by the QA system. It is also encouraging that the F1 overlap score for the BERT answer highlight algorithm is on par with the F1 obtained when training on the data set presented in Section 3 (see Table 1), even if the newly developed test set contains more recent questions than the ones in that data set.

## 7 Conclusions

We have presented an open-domain QA system that uses a fine-tuned BERT model to highlight probable answers to the input question in the Bing-returned text snippets. With a MRR bigger than 0.5, we have the guarantee that the user sees relevant content in the first two results returned by the QA system, with more relevant content following, as Bing finds useful information in more than one document. Furthermore, a character overlap F1 of almost 77% between the correct answer and the BERT supplied answer will steer user’s attention effectively towards the correct answer.

Comparing the data sets sizes, we see that our COVID-19 data set is two orders of magnitude smaller than the data sets presented in Rajpurkar et al. (2018). The exact match and F1 scores on the SQuAD 1.1 data set are 78.6% and 85.8% respectively, suggesting that we have room to improve

in these areas, provided that we can grow our data set significantly. But even with the current performance, the QA system is useful as it is.

The open-domain QA pipeline is trainable in the sense that, given a data set similar to the one presented in Section 3 but in a different domain, one can fine-tune the chosen BERT model to answer questions in the domain of the new data set. Relying on a web search engine such as Bing, indexing billions of documents on a regular basis, the part of answer retrieving is assured, irrespective of the chosen domain.

In a different use case, the QA system can be adapted to work on e.g. public institution web sites, by having the public institution web sites indexed using either a local search engine or local Bing/Google indexing and using these specialized search results instead of the web-wide search results. Coupled with a data set of questions on the specific topic of interest (e.g. “tax payments”, “public transportation”, “resident parking”, etc.), the QA pipeline can work to answer targeted questions from citizens.

The next steps in the development of this QA system are:

- To test it with other languages by using the *eTranslation* online machine translation service provided by the European Commission. We could automatically translate the input question into Romanian, run the QA pipeline and then translate the output of the QA system into the input question language.
- To develop an answer mining algorithm that is better than Bing’s algorithm for mining the text snippet that is most relevant to the query. We would parse the URL of the returned document and select the text snippet that contains the relevant answer ourselves, instead of using the provided text snippet.

## Acknowledgments

The Action 2020-EU-IA-0088 (“Enrich4All”) has received funding from the European Union’s Connecting Europe Facility 2014-2020 – CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278547.

## References

Andrei-Marius Avram, Darius Catrina, Dumitru-Clementin Cercel, Mihai Dascălu, Traian Rebedea,

- Vasile Păiș, and Dan Tufiș. 2021. Distilling the knowledge of romanian bert's using multiple teachers. *arXiv preprint arXiv:2112.12650*.
- Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. *Advances in neural information processing systems*, 6.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of romanian bert. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.
- Radu Ion. 2018. TEPROLIN: An Extensible, Online Text Preprocessing Platform for Romanian. In *Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2018)*, Iași, România.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.
- Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2020. Question and answer test-train overlap in open-domain question answering datasets.
- Mihai Masala, Stefan Ruseti, and Mihai Dascalu. 2020. Robert—a romanian bert model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6626–6637.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:1802.05365*.
- Vasile Păiș. 2020. Multiple annotation pipelines inside the relate platform. In *The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*, pages 65–75.
- Vasile Păiș, Radu Ion, and Dan Tufiș. 2020. A processing platform relating data and tools for Romanian language. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 81–88, Marseille, France. European Language Resources Association.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad.
- Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage bert: A globally normalized bert model for open-domain question answering. *arXiv preprint arXiv:1908.08167*.
- Ikuya Yamada, Akari Asai, and Hannaneh Hajishirzi. 2021. Efficient passage retrieval with hashing for open-domain question answering. *arXiv preprint arXiv:2106.00882*.
- Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718*.
- Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering.



# Zero-shot Event Causality Identification with Question Answering

**Daria Liakhovets**

AIT Austrian Institute of Technology  
daria.liakhovets@ait.ac.at

**Sven Schlarb**

AIT Austrian Institute of Technology  
sven.schlarb@ait.ac.at

## Abstract

Extraction of event causality and especially implicit causality from text data is a challenging task. Causality is often treated as a specific relation type and can be considered as a part of relation extraction or relation classification task. Many causality identification-related tasks are designed to select the most plausible alternative of a set of possible causes and consider multiple-choice classification settings.

Since there are powerful Question Answering (QA) systems pretrained on large text corpora, we investigated a zero-shot QA-based approach for event causality extraction using a Wikipedia-based dataset containing event descriptions (articles) and annotated causes. We aimed to evaluate to what extent reading comprehension ability of the QA-pipeline can be used for event-related causality extraction from plain text without any additional training. Some evaluation challenges and limitations of the data were discussed. We compared the performance of a two-step pipeline consisting of passage retrieval and extractive QA with QA-only pipeline on event-associated articles and mixed ones. Our systems achieved average cosine semantic similarity scores of 44 – 45% in different settings.

**Keywords:** event causality identification, question answering, semantic similarity search.

## 1 Introduction

The aim of the work was to exploit the reading comprehension of pre-trained Question Answering (QA) models to address zero-shot event causality extraction from text. Since implicit causality can be expressed in various, potentially infinite number of

ways, and causality expressions can be distributed throughout sentences, identification of event causality remains a challenging task.

Many related data resources are designed for binary statement classification, multiple-choice QA, or relation classification. For our experiments we used a semantic similarity search-based dataset obtained from annotated Wikipedia articles. The dataset was designed for event-related causality extraction from plain text. However, the data had some limitations discussed in the related section.

We compared a two-step extraction pipeline consisting of relevant text passage retrieval based on semantic similarity search and cause candidate retrieval based on QA. The experiments were performed in two different settings: related documents and mixed documents subsets.

The paper is structured as follows: Section 2 overviews related work on causality identification, including some question-driven approaches. Section 3 describes our data, experiments and evaluation metrics, and Section 4 presents the results.

## 2 Related work

### 2.1 Causality identification: resources and approaches

Resources, approaches, and problems in causal relation identification in NLP are discussed by (Han and Wang, 2021). The authors distinguish causal relation classification and causal relation extraction and the classification level (word-, sentence- or passage-level). Causality is often treated as a specific type of entity relations. Some datasets combine event causality and temporal relations, e.g., (Caselli and Vossen, 2017). There are some domain-specific resources, e.g., (Kyriakaki et al., 2019), (Mariko et al., 2020). Others, e.g., (Huang et al., 2019), (Ponti et al., 2020), are designed for commonsense multiple-choice causal QA. There

are also knowledge bases containing causal relations or lexical markers.

Causality expressions can be explicit (e.g., “because”) or implicit, the latter are more common but more difficult to recognize. Open class lexical markers, AltLexes (Prasad et al., 2008), are somewhere in the middle due to their linguistic variety (Hidey and McKeown, 2016).

Since existing labelled event causality detection datasets are limited in size, data augmentation techniques used, such as synonym substitution (Staliūnaitė et al., 2021) or external causal knowledge (Dalal et al., 2021). (Zuo et al., 2020) suggested a data augmentation framework based on lexical and causal commonsense knowledge. (Ruan et al., 2019) used WHY-type question-answer pairs from QA datasets and Question-Statement Conversion for training set expansion.

(Han and Wang, 2021) summarize methods for causal relation identification. While unsupervised methods are mainly based on predefined rules and patterns, supervised methods use feature engineering, global optimization, or deep learning approaches on labelled data. Despite the achieved good performance in many causal relation identification tasks, extracting implicit causal knowledge from the free text is still an unsolved task.

(Doan et al., 2019) used dependency parsing on lemmatized POS-tagged tweets to extract cause-effect relations for several health-related effects (e.g., “headache”). (Kyriakaki et al., 2019) used transfer learning to detect causal sentences in commonsense datasets and in BioCausal data and experimented with the BIGRUATT layer. (Kadowaki et al., 2019) investigated ensemble approaches based on individual judgements of three annotators and exploiting background context knowledge for binary classification of statement pairs. (Mariko et al., 2020) fine-tuned BERT for binary sentence classification in financial news. (Liang et al., 2022) proposed a novel model that exploits the advantages of both feature engineering and neural model-based approaches. (Zhao et al., 2021) proposed a document-level context-based graph inference mechanism to identify event causality.

## 2.2 Question-driven approaches

Event causality identification can be considered as a part of automated story generation. (Castricato et al., 2021) proposed a novel approach that reconstructs the story backwards by iteratively generating “why“-questions to find the preceding event from the given one. (Zhou et al., 2021) used QA to identify nested causality in traffic accident data.

Zero-shot methods aim to overcome the limitations of predefined relation set-based approaches towards extracting new unseen types of relations or facts. (Levy et al., 2017) used QA to perform zero-shot relation extraction by associating natural-language questions with each relation type and demonstrated the generalization ability of the approach on unseen relation types. (Goodwin et al., 2020) applied multi-task fine-tuning for zero-shot conditional summarization that selects the most salient points based on a question or a topic of interest. (Chakravarti et al., 2020) addressed a zero-shot industrial QA task introducing the model GAAMA with improved attention mechanisms. (Zhou et al., 2021) proposed a novel method for automatic transfer of explanatory knowledge in zero-shot science QA.

## 3 Experimental setup

### 3.1 Data

To address event-related causality identification from free text, we obtained a dataset from the Wikipedia *List of protests in the 21st century*<sup>1</sup>. The dataset language was English. We extracted human-annotated “caused by” attributes from “infobox” sections (Figure 1).

Since extractive methods require annotations to appear in text, we looked for annotated causes in text. Some annotations were matched exactly in the related article, others had to be searched for by their paraphrased appearances, e.g., “*authoritarianism*” could be found as “*authoritarian rule*”.

We created two dataset versions: using fuzzy string-matching functions from `thefuzz`<sup>2</sup> package and using semantic similarity search with `Sentence Transformers`<sup>3</sup> introduced by (Reimers and Gurevych, 2019). While the first

<sup>1</sup>[https://en.wikipedia.org/wiki/List\\_of\\_protests\\_in\\_the\\_21st\\_century](https://en.wikipedia.org/wiki/List_of_protests_in_the_21st_century)

<sup>2</sup><https://github.com/seatgeek/thefuzz>

<sup>3</sup><https://www.sbert.net/>

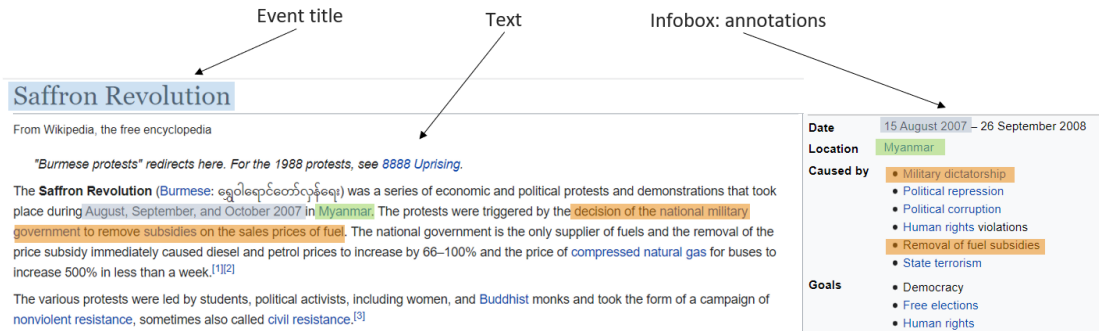


Figure 1: Wikipedia-article with an infobox-section.

approach is based purely on token similarity, the second one uses embeddings produced by the all-mpnet-base-v2<sup>4</sup> model to compute the cosine similarity of two sequences. Thus, it can capture the semantic content, even if it is expressed with different words. As many annotated causes appear as paraphrased expressions, we used the second version of the dataset for our experiments. The minimum threshold of cosine similarity score was set to 0.70 to obtain a subset with better appearances of original annotations. The final subset contained 905 annotated causes linked to 297 unique articles; 245 causes were matched exactly (score 1.00), and 660 causes with similar phrases.

The data has the following limitations:

- **Objectivity:** authors of the Wikipedia articles may be biased. One may argue whether such annotations should be used as ground truth labels.
- **Completeness:** causal reasons may appear in the text without being annotated and therefore cannot be evaluated reliably.
- **Unlinked and inconsistently structured annotations:** firstly, annotations are not linked to their appearance in the text. For reliable evaluation, approximatively matched causes should be confirmed manually. Secondly, authors use different separators and list styles. Splitting the annotations into single cause items may break sentences into parts unevaluable for causality.

### 3.2 Question-driven cause candidate extraction

We used a question-driven two-step extraction approach to identify the cause candidates for an event of interest. To extract causality of a specific event, we constructed a question using the event title – in our case the Wikipedia article title – to complete the following simple question template:

*What caused <EVENT\_TITLE>?*

We split articles into smaller passages, with a maximum of 200 WordPiece (Schuster and Nakajima, 2012) tokens, retaining the text structure, i.e., sentences and paragraphs. We exploited embeddings from multi-qa-mpnet-base-dot-v1<sup>5</sup>, a model designed for semantic search to compute the dot similarity score of the question and passages and extract relevant ones (Figure 2).

In the next step we used xlm-roberta-large-squad2<sup>6</sup>, a model designed for extractive QA, to retrieve answers from three most relevant passages (Figure 3). Since one article usually had multiple annotations, we retrieved several answer candidates from each passage and then selected two best ranked answers more than the number of annotations. Answer candidates were selected based on their probability of being an answer for the asked questions, which was calculated by the QA model.

Once several cause candidates for the article had been extracted, we had to match them with the annotations. We computed pairwise cosine similarity scores based on all-mpnet-base-v2

<sup>4</sup> <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

<sup>5</sup> <https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1>

<sup>6</sup> <https://huggingface.co/deepset/xlm-roberta-large-squad2>

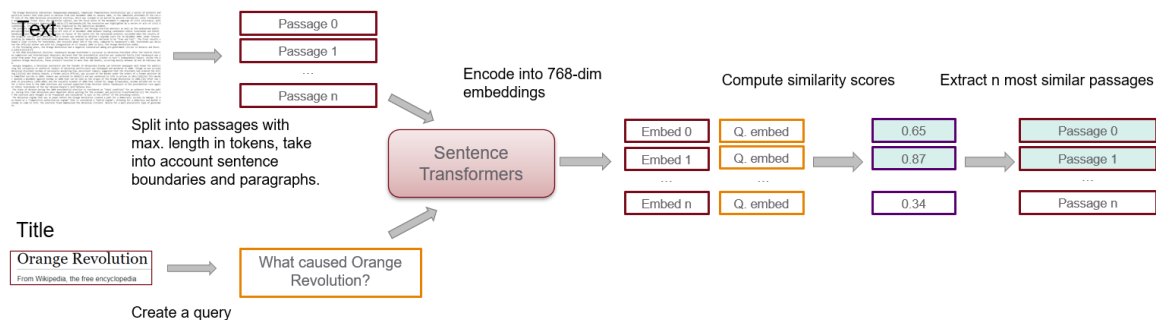


Figure 3: Step 1: Passage retrieval using semantic similarity.

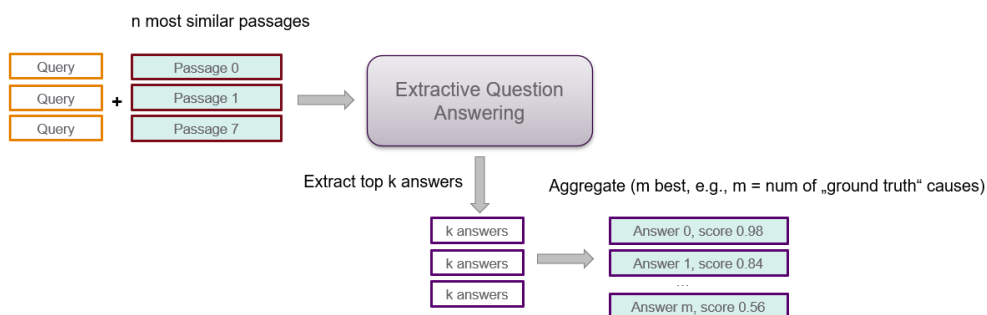


Figure 2: Step 2: Cause candidate extraction using QA.

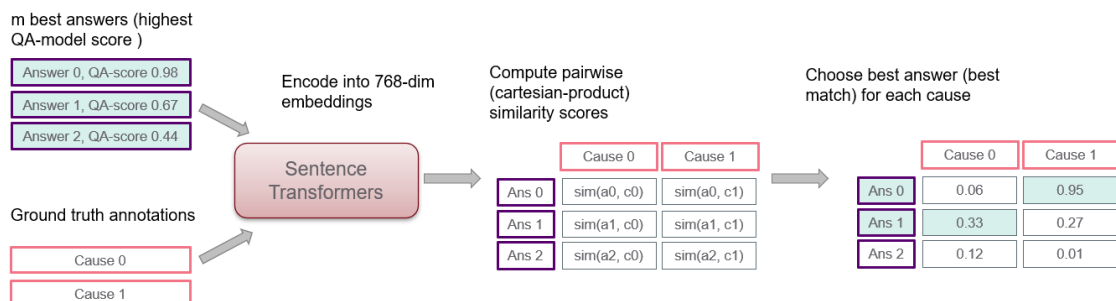


Figure 4: Matching annotations with extracted answers.

embeddings. For each annotation, the best match was selected and then used for the evaluation (Figure 4).

We compared the two-step extraction with the QA-only approach which has no passage retrieval step and just retrieves answer candidates from each text passage.

We experimented with two settings: extracting cause candidates only from a related article and from mixed documents, which is more realistic. In the second case, for each article we created a subset

of 10 documents: the article itself and 9 random articles.

### 3.3 Evaluation metrics

To evaluate the retrieved answer candidates, we used the semantic similarity score (cosine similarity) computed based on `all-mpnet-base-v2` embeddings during best answer matching, F1-score, and exact match (EM). We removed punctuation and stop-words and compared two lowercased sets of tokens to obtain

the F1-score. For EM, we compared two lowercased phrases without punctuation.

The F1-score is based on the lexical overlap of two token sequences, and EM just indicates whether the sequences are identical or not. Since more than 70% of entries in our data cannot be found exactly in the related articles, the semantic similarity score is more useful for evaluation. One could also use cross-encoder model-based scoring, as proposed by (Risch et al., 2021). For measuring lexical overlap, ROUGE metric (Lin, 2004) can be useful.

#### 4 Results and discussion

Metric	Rel.: 2-step	Rel.: QA-only
Cos. similarity, avg.	0.4451	0.4588
F1-score, avg.	0.1516	0.1666
Exact match, n	7	9
No answer, n	0	0

Table 3: Evaluation results on related articles.

Metric	Mixed: 2-step	Mixed: QA-only
Cos. similarity, avg.	0.4397	0.4386
F1-score, avg.	0.1489	0.1513
Exact match, n	7	8
No answer, n	3	0

Table 2: Evaluation results on mixed articles.

The evaluation results are summarized in Table 1 (related articles) and Table 2 (mixed articles).

The number of exact matches is very low (< 1%) in all cases. There is no significant difference between the two approaches, judging by the metrics. However, QA-only is more time-intensive because it processes all text chunks.

The QA-only approach provided two and one more exact matches than passage retrieval + QA in related document- and mixed document-settings, respectively, as well as slightly higher F1-scores. In the mixed setting QA-only was able to find candidates for all annotations while the two-step approach missed candidates for three causes, i.e., some salient text passages were ignored during

passage retrieval. This issue can be addressed by increasing the number of passages and/ or improving the quality of passage ranking techniques. However, we still think that the passage extraction step can have advantages when dealing with large text collections. Further experiments are needed to prove this.

The results could be improved by additional domain-specific model training and increasing the number of retrieved passages and answer candidates. Generative summarization could be a better choice than using only extractive methods.

Table 3 contains some examples of extracted cause candidates. The top half refers to the dataset: the “True cause” column contains original annotations, “Best match” presents the most similar phrase found in the article, and “Matching score” shows their similarity score. In the bottom half, “Best answer” contains the best candidate for the “True cause” and the appropriate “Answer

#	Example 1	Example 2	Example 3
<b>True cause</b>	Mexican Drug War	2017 wealth tax repeal	religious nationalism
<b>Best match</b>	Mexican Drug War	to reinstate a wealth tax	nationalism
<b>Matching score</b>	1.00	0.76	0.85
<b>Best answer</b>	Mexican Drug War,	Their principal concern was tax justice.	mobs attacking Muslims.
<b>QA score</b>	0.02	0.22	0.79
<b>Answer matching score</b>	0.97	0.44	0.41

Table 1: Examples of results. Top half: True cause, Best match, Matching score refer to the dataset. Bottom half: Best answer, QA-score, Answer matching score refer to extracted causes.

matching score”. “QA score” presents scores computed by the QA model.

The first example demonstrates a large gap between the low probability of being an answer to the asked question and the high score of matching with the ground truth annotation. In a real-world

application, relying on the QA score, this answer would be low-ranked. The second example can be considered satisfactory by human judgement. Although the best answer conveys the main idea of the true annotation, its answer matching score is relatively low, as well as its QA score. The third example illustrates a cause candidate scored highly by the QA model but having a relatively low answer matching score. These examples demonstrate the need to define a sufficient level of similarity, because even similarity scores under 0.5 may still indicate adequate matches.

## 5 Conclusions and future work

In this work, we conducted experiments to evaluate the zero-shot event causality identification with semantic search-based passage retrieval and QA on a dataset obtained from Wikipedia. We compared the two-step and the QA-only approaches on related and mixed documents and demonstrated their similar performance in the experimental settings. While the two-step approach could not find any candidates for a few ground truth annotations in the mixed document setting, QA-only was able to find candidates in all cases. QA-only also performed slightly better on related documents, however, it required more computational time. Further experiments are necessary to identify whether the passage retrieval step bring other advantages when processing large document collections. Our systems achieved average cosine semantic similarity scores of 44 – 45% in different settings.

We think that the reading comprehension of QA models can be used to address the challenge of event causality extraction. In the future work, both passage and answer retrieval can be improved by using models with domain-specific knowledge, as well as increasing the number of retrieved passages and candidate. Using other or multiple question templates could help to retrieve more various cause candidates.

## References

Caselli, T. & Vossen, P., 2017. The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction. *Proceedings of the Events and Stories in the News Workshop*, p. 77–86.

Castricato, L., Frazier, S., Balloch, J. & Riedl, M., 2021. Tell Me A Story Like I'm Five: Story

Generation via Question Answering. *Proceedings of the 3rd Workshop on Narrative Understanding*.

Chakravarti, R. et al., 2020. Towards building a Robust Industry-scale Question Answering System. *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*.

Dalal, D., Arcan, M. & Buitelaar, P., 2021. Enhancing Multiple-Choice Question Answering with Causal Knowledge. *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*.

Doan, S. et al., 2019. Extracting health-related causality from twitter messages using natural language processing. *BMC Med Inform Decis Mak* 19.

Goodwin, T. R., Savery, M. E. & Demner-Fushman, D., 2020. Towards Zero-Shot Conditional Summarization with Adaptive Multi-Task Fine-Tuning. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Han, M. & Wang, Y., 2021. A Survey on the Identification of Causal Relation in Texts. *2021 IEEE International Conference on Progress in Informatics and Computing (PIC)*, pp. 1-7.

Hidey, C. & McKeown, K., 2016. Identifying Causal Relations Using Parallel Wikipedia Articles. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 1424–1433.

Huang, L., Bras, R. L., Bhagavatula, C. & Choi, Y., 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. *EMNLP'2019*.

Kadowaki, K. et al., 2019. Event Causality Recognition Exploiting Multiple Annotators' Judgments and Background Knowledge. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Kyriakaki, M., I. A., Ametllé, J. G. i. & Saudabayev, A., 2019. Transfer Learning for Causal Sentence Detection. *BioNLP 2019 workshop*.

Levy, O., Seo, M., Choi, E. & Zettlemoyer, L., 2017. Zero-Shot Relation Extraction via Reading Comprehension. *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*.

Liang, S. et al., 2022. A multi-level neural network for implicit causality detection in web texts. *Neurocomputing, Volume 481*, pp. 121-132.

Lin, C.-Y., 2004. ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, p. 74–81.

Mariko, D. et al., 2020. The Financial Document Causality Detection Shared Task (FinCausal 2020). *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*.

Ponti, E. M. et al., 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Prasad, R. et al., 2008. The Penn Discourse TreeBank 2.0. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Reimers, N. & Gurevych, I., 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, January, pp. 3973-3983.

Risch, J., Möller, T., Gutsch, J. & Pietsch, M., 2021. Semantic Answer Similarity for Evaluating Question Answering Models. *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, November, p. 149–157.

Ruan, H. et al., 2019. Using WHY-type Question-Answer Pairs to Improve Implicit Causal Relation Recognition. *2019 International Conference on Asian Language Processing (IALP)*.

Schuster, M., & Nakajima, K., 2012. Japanese and Korean Voice Search. *International Conference on Acoustics, Speech and Signal Processing*, p. 5149-5152.

Staliūnaitė, I., Gorinski, P. J. & Iacobacci, I., 2021. Improving Commonsense Causal Reasoning by Adversarial Training and Data Augmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhao, K. et al., 2021. Document-level event causality identification via graph inference mechanism. *Information Sciences* 561(3).

Zhou, G. et al., 2021. Nested Causality Extraction on Traffic Accident Texts as Question Answering. *NLPCC 2021: Natural Language Processing and Chinese Computing*.

Zhou, Z., Valentino, M., Landers, D. & Freitas, A., 2021. Encoding Explanatory Knowledge for Zero-shot Science Question Answering.

*Proceedings of the 14th International Conference on Computational Semantics (IWCS)*.

Zuo, X., Chen, Y., Liu, K. & Zhao, J., 2020. KnowDis: Knowledge Enhanced Data Augmentation for Event Causality Detection via Distant Supervision. *Proceedings of the 28th International Conference on Computational Linguistics*.



# Ontology of Visual Objects

Svetla Koeva

Institute for Bulgarian Language, Bulgarian Academy of Sciences  
svetla@ddcl.bas.bg

## Abstract

The focus of the paper is the **Ontology of Visual Objects** based on WordNet noun hierarchies. In particular, we present a methodology for bidirectional ontology engineering, which integrates the pre-existing knowledge resources and the selection of visual objects within the images representing particular thematic domains. The Ontology of Visual Objects organizes concepts labeled by corresponding classes (dominant classes, classes that are attributes to dominant classes, and classes that serve only as parents to dominant classes), relations between concepts and axioms defining the properties of the relations. The Ontology contains 851 classes (706 dominant and attribute classes), 15 relations and a number of axioms built upon them. The definition of relations between dominant and attribute classes and formulations of axioms based on the properties of the relations offers a reliable means for automatic object or image classification and description.

## 1 Introduction

The recent trends in Computer vision are directed towards the robust combination of deep learning techniques and image processing methods to solve problems, such as image and video understanding, robot vision and processing of multimodal and multilingual content. Despite this, much effort is still directed to specific domain knowledge or even to specific object instance recognition, and the significant progress in the field as a whole does not mean that particular tasks have been solved satisfactorily.

The concept of Cognitive vision (Vernon, 2021) was introduced quite a long time ago (Auer et al., 2005): “A cognitive vision system can achieve the four levels of generic computer vision functionality of detection, localization, recognition, and understanding. It can engage in purposive goal-directed behaviour, adapting to unforeseen changes of the

visual environment, and it can anticipate the occurrence of objects or events” (Vernon, 2006).

Such understanding of Cognitive vision systems involves the application of ontology-based representations in modern Computer vision systems in order to add real world relations between static objects and video feed (Xie et al., 2020; Chaisiriprasert et al., 2021). Ontology-based applications might be powerful tools for diverse Computer vision tasks: application of semantics according to the function of an object (Agostini et al., 2015), ontology-based object recognition in robotics (Riazuelo et al., 2015), and so on.

The focus of the paper is the **Ontology of Visual Objects**.<sup>1</sup> In particular, we present a methodology for bidirectional ontology engineering, which integrates the pre-existing knowledge resource (WordNet) and the selection of visual objects within the images representing particular thematic domains.

We show how the presented Ontology benefits from WordNet (Miller et al., 1990; Fellbaum, 1999): providing ontological representation of visual objects based on WordNet noun hierarchies, and building interconnectivity of classes by means of the WordNet. On the other hand, we present how the Ontology of Visual Objects builds on the WordNet by adding new concepts corresponding to concrete objects, and formulating new relations that express the objects’ function, purpose, location, etc.

We begin with a brief overview of the current state in the art in Section 2. In Section 3 we present the principles of Ontology-based image annotation. Section 4 is dedicated to the main components of an ontology and the description of the Ontology of Visual Objects. Finally, evaluation (section 5), conclusions and future directions of our work (section 6) are presented.

---

<sup>1</sup><https://doi.org/10.57771/a0w5-8480>

## 2 Related Work

In this section, we briefly present some of the most prominent knowledge representation resources, the image datasets, which involve (in different ways) ontologies in the process of their building, and the few existing examples of ontologies specially dedicated to image descriptions.

### 2.1 Ontology-based Semantic Resources

The taxonomic organization of nouns in **WordNet** allows for using more abstract and fine-grained categories when describing objects. WordNet<sup>2</sup> is a semantic network, whose nodes host synonyms denoting different concepts, and whose arcs, connecting the nodes, encode different types of relations (semantic: genus-kind, part-whole, etc.; extralinguistic: membership in a thematic domain; interlanguage: translation equivalents).

The idea for organizing the lexicon of a given language into a (lexico-)semantic network was first executed in the Princeton WordNet (Miller et al., 1990). Some of the fundamental ideas on which the WordNet is based encompass: a) the use of a semantic network which embraces taxonomies, meronomies and non-hierarchical relations with clearly defined properties, which allow for quick and easy automatic processing; b) a different organization of the lexicon in comparison with the traditional dictionaries where words are ordered alphabetically and the links among semantically related words (such as between sister hyponyms, between a whole and its parts, etc.) are not explicitly presented (Miller, 1986).

WordNet is connected to a generic ontology based on **DOLCE**.<sup>3</sup> A set of heuristics for mapping all WordNet nouns, verbs and adjectives to the ontology were developed, which also allows to represent predicates in a uniform and interoperable way, regardless of the way they are expressed in the text and in which language (Laparra et al., 2012). Together with the ontology, the WordNet mappings provide powerful basis for semantic processing of text in different domains.

Some ontologies have been developed on top of the existing resources. The **YAGO** ontology<sup>4</sup> is a large knowledge base with general knowledge about people, cities, countries, movies, and organizations (Suchanek et al., 2007). YAGO contains

both entities (such as *movies*, *people*, *cities*, *countries*, etc.) and relations between these entities (who played in which movie, which city is located in which country, etc.). The entities are arranged in classes: *Elvis Presley* belongs to the class of *people*, *Paris* belongs to the class of *cities*, and so on, which in their turn are arranged in a taxonomy: the class of *cities* is a subclass of the class of *populated places*, etc. YAGO combines Wikidata – the largest general-purpose knowledge base on the Semantic Web and schema.org (plus BioSchemas) – a standard ontology of classes and relations.

**BabelNet**<sup>5</sup> (as WordNet) combines features of multilingual encyclopaedic dictionary (with its wide lexicographic and encyclopaedic coverage of terms), and of semantic network or ontology, which links concepts and named entities in a very large network of semantic relations (about 20 million entries as of 2021) (Navigli et al., 2021). BabelNet brings together heterogeneous resources, such as WordNet, Wikipedia, OmegaWiki, Wikidata, Wiktionary, GeoNames, Open Multilingual WordNet and many others, and aims at providing as complete picture as possible of lexical and semantic knowledge available in many languages. BabelNet represents each meaning based on the WordNet notion of a synset. Analogously to WordNet, BabelNet can be viewed as a graph where synsets are nodes and edges are semantic relations between them.

### 2.2 Ontology-supported Image Datasets

There are several datasets that have been widely used as a benchmark for object detection, semantic segmentation and classification tasks. Only a few of them use ontologies or ontology-like resources for object classification.

Thousands of images, hundreds of thousands of polygon annotations and sequence frames with at least one tagged object are all included in the **LabelMe** dataset<sup>6</sup> (Russell et al., 2008). This collection is being created by users who can upload images, add categories and annotate images with these categories. However, depending on how each annotator chooses to use the annotation protocol, this choice can lead to some degree of inconsistency. By using the WordNet noun synsets, categories are expanded, inconsistent editing is avoided and user-provided descriptions are unified.

<sup>2</sup><http://wordnet.princeton.edu/>

<sup>3</sup><http://www.loa.istc.cnr.it/dolce/overview.html>

<sup>4</sup><https://yago-knowledge.org/>

<sup>5</sup><https://babelnet.org/>

<sup>6</sup><http://labelme.csail.mit.edu>

One of the collections that sets standards in the increase of datasets sizes is **ImageNet**.<sup>7</sup> A dataset with roughly 50 million full-resolution images that have been accurately labelled has been set as a target (Deng et al., 2009). The WordNet noun hierarchies are used for image collection and labelling. ImageNet comprises 14,197,122 annotated images that are arranged according to the semantic hierarchy of WordNet and employs 21,841 synsets for focused image search (as of August 2014) (Rusakovsky et al., 2015).

More than 328,000 images with carefully annotated object instances (2.5 million) can be found in the **COCO (Microsoft Common Objects in Context)** dataset<sup>8</sup> (Lin et al., 2014). Since 2014, the dataset has undergone a number of updates and covers object detection, segmentation, keypoint detection and captioning. The different parts of the dataset are annotated with bounding boxes (for object detection) and per instance segmentation masks with 80 object categories; natural language descriptions of the images; keypoints (17 possible key points, such as *left eye*, *nose*); per pixel segmentation masks with 91 stuff categories, such as *grass*, *wall*; full scene segmentation, with 80 thing categories (such as *person*, *bicycle*, *elephant*); dense pose – each labelled person is annotated with a mapping between image pixels and a template 3D model.

WordNet is typically utilized in current practice to generate text queries for building search-based image collections. Some of the datasets were developed using shallow ontologies (Griffin et al., 2007), and overall, the potential power of the ontological structure is not completely exploited.

### 2.3 Existing Ontologies of Visual Objects

The **LSCOM** ontology consists of 1,000 concepts and approximately 450 of them were used for the manual annotation of 80 hours of news video (Naphade et al., 2006). The taxonomy design organized concepts into six categories on a top level: *objects*, *activities/events*, *scenes/locations*, *people*, *graphics*, and *program* categories. These categories were further refined, such as by subdividing objects into *buildings*, *ground vehicles*, *flying objects*, etc.

**Photo Tagging Ontology** covering 100 concepts was issued with the ImageCLEF annotation task (Xioufis et al., 2011). The ontology restricts si-

multaneous assignment of some concepts (disjoint classes) and defines that one concept postulates the presence of other concepts. The purpose of the ontology is to allow integration of semantic knowledge in the algorithms for image annotations.

A **Visual Concept Ontology** organizes visual concepts (objects or abstract notions that are typically depicted in photos) (Botorek et al., 2014). For the construction of Visual Concept Ontology over 400 “significant” noun synsets (that have at least 300 hyponyms) were extracted from WordNet; then synsets with a very “general” meaning, such as *entity* or *thing*, were removed. This results in 14 top-level ontology classes, which are divided further into 90 more specific classes. On top of these, a final high-level generalization was performed, producing 4 super-classes: *nature*, *person*, *object* and *abstract concepts*. Semantically similar synsets are merged into a common class and additional links are established between semantically related synsets, such as *roof* and *house*. In other words, the ontology simplifies and flattens the WordNet hierarchy, removing concepts not relevant to the visual domain and adding semantic connections between interrelated WordNet subtrees. Relations are of two basic types – class-to-class and class-to-individual.

It has been demonstrated that combining ontology knowledge with image recognition technologies can increase recognition precision, enhance high-level semantic recognition capability, decrease the need for a large number of training samples and improve the scalability of the image recognition systems (Ding et al., 2019).

In conclusion, it can be stated that the ontological representation of knowledge is not fully exploited in Computer vision: neither in the process of creating annotated datasets, nor in the implementation of algorithms and models for the recognition and classification of objects and images.

## 3 Ontology-based Image Annotation

The **Ontology of Visual Objects** was developed to serve for the annotation of the image objects in the Multilingual Image Corpus,<sup>9</sup> which provides pixel-level annotations, thus offering data to train models specialised in object detection, segmentation and classification in these domains (Koeva, 2021; Koeva et al., 2022).

Different ways of incorporating semantics to describe an image are discussed (Tousch et al., 2012).

<sup>7</sup><https://www.image-net.org>

<sup>8</sup><https://cocodataset.org>

<sup>9</sup>[doi.org/10.57771/p2n7-f015](https://doi.org/10.57771/p2n7-f015)

One possible level incorporates the relations between concrete and abstract objects, for example, a *crying person* vs. the notion of *pain*, which might be a subjective conclusion based on the knowledge of the semantic context. The other level describes generic vs. specific objects (individual instances), i.e., a *bridge* vs. *Golden gate bridge*. In our approach, we concentrate on visual (concrete) objects; however, specific instances of an object can be further related with it, and further inferences to abstract notions might be drawn as well.

We defined the following criteria for the development of the **Ontology of Visual Objects**:

- The specificity or generality of the concept (we include only specific concepts at a certain level of granularity: more concrete comparing to classes that are usually used in image datasets, for example *taxi* and *sedan* instead of a *car*, but not too concrete, in order for the annotators to be able to choose among the classes without employing specific knowledge for different thematic domains, (for example *sedan*, but not *Bentley* or *Dacia*).

- High frequency of occurrence of words denoting visual objects in everyday life and of respective objects depicted in images. The everyday use is based on the inclusion of the words in the so-called common vocabulary, which is evidenced by the Age of acquisition list of words (Brysbaert and Biemiller, 2017). The assumption is that words that are mastered at an early age belong to the basic vocabulary. The frequency of encounters of objects in the images is observed empirically, based on the collected over 750,000 images, of which about 21,000 were selected for annotation in the Multilingual Image Corpus. For example, although the object *baby rattle* is expected to meet frequently along with dominant objects, such as a *baby* and a *stroller*, empirical observations in images have shown a low frequency of encounters, and this visual object is not included in the Ontology.

- Coverage in ontologies (concepts already encoded within the WordNet and through WordNet in other ontologies).

- Covering gaps in existing ontologies, for example, some objects we observed in the collected images (such as *handball player*, *pole vaulter*, etc. have not been included in the Princeton WordNet so far).

The proposed **Ontology of Visual Objects** includes concepts that are characteristic for the thematic domains of **Sport**, **Transport**, **Arts**, and

**Security**. The Multilingual Image Corpus contains 130 smaller datasets pertaining to different subdomains, each of which can be classified to one of the four main ones, for example, **Chess** and **Pole vaulting** are subdomains of **Sport**, while **Sedan** and **Double-decker** – to **Transport**, and so on. The choice of thematic domains and subdomains is motivated by two main factors:

- (1) The images should contain objects that could be automatically recognized and labelled with upper-level classes (for example, *man* and *car*), which then could be sub-classified as *chess player*, *pole vaulter*, *sedan* and *taxi*;

- (2) There should be a sufficient number of appropriate images available to illustrate objects from the selected thematic subdomains.

Ontologies are classified into three basic types: *top ontologies*, which contain a restricted set of general classes and are not related to a particular thematic domain; *top-domain ontologies*, which include essential classes that represent a particular thematic domain; and *domain ontologies*, which contain classes that comprehensively describe a particular thematic domain (Tan and Lambrix, 2009). From the point of view of this classification, the proposed ontology can be classified as a set of several domain ontologies.

The **Ontology of Visual Objects** provides options for extracting relationships between annotated objects, between diverse datasets with different levels of granularity of object classes, or between appropriate sets of images illustrating different thematic domains. Last but not least, the use of the Ontology of Visual Objects allows the expansion of the dataset depending on the specific needs of scientific or commercial projects.

The annotators' tasks were to create new polygons or approve or modify the automatic segmentation for objects in the images, and then classify the objects according to the specified Ontology's classes. The annotation adheres to the following conventions:

- An object displayed within an image is annotated if it represents an instance of a concept included in the Ontology.
- All objects from the selected dominant class and attribute classes related with it are annotated (for example, the *tennis player* and the related objects *racket* and *tennis ball*; *chess player* and the related objects *chessman*, *chess board* and *clock*).



The following are some advantages of utilizing an ontology for object classification:

- Selection of mutually exclusive classes.
- Build-in interconnectivity of classes by means of formal relations.
- Easy extension of the proposed ontology with more concepts corresponding to visual objects.

#### 4 Ontology of Visual Objects

It was pointed out that different knowledge representations share the following minimal set of components (Corcho et al., 2006): **concepts**, which represent sets or classes of entities in a thematic domain; **relations** between concepts; **instances**, which represent the actual entities (individuals); and **axioms**, which represent facts that are always true in the topic area of the ontology. We accepted the following definition (Bozsak et al., 2002): An ontology is a structure

$$O := (C, \leq_C, R, \leq_R)$$

consisting of (i) two disjoint sets  $C$  and  $R$  called concept identifiers and relation identifiers respectively, (ii) a partial order  $\leq_C$  on  $C$  called concept hierarchy or taxonomy, (iii) a function  $\sigma : R \rightarrow C \times C$  called signature and (iv) a partial order  $\leq_R$  on  $R$  called relation hierarchy.

The **Ontology of Visual Objects** organizes concepts (represented by dominant classes, classes that are attributes to dominant classes and classes that serve only as parents to dominant classes), relations between concepts and axioms.

##### 4.1 Classes

**Classes** correspond to (WordNet) concepts that can be represented by visual objects. Among the classes, we made a differentiation between dominant classes and attribute (contextual) classes.

Each thematic domain is represented by several **dominant classes**, which show the main “players” within this domain differentiated by their type or their function. For example, the dominant classes for the domain Security are: **policeman, soldier, fireman**, etc., altogether 15 dominant classes. For the definition of the dominant classes, we use the WordNet sister **hyponyms** at a certain level (the lowest level allowing classification without specific knowledge for the domain). So far, the selected

dominant classes for all thematic domains in focus are 137.

For each dominant class a parent class is selected from the WordNet noun hierarchies and this procedure is repeated consecutively up to the final class that represents a visual object. For example, classes like *basketball player, acrobat, football player*, etc. are **hyponyms** of *athlete* ‘a person trained to compete in sports’. *Athlete* in its turn is a **hyponym** of *contestant* ‘a person who participates in competitions’ which is a hyponym of *person*. However, the **hyponym** of *person* is *organism*, an abstract notion, which is not included in the ontology. As a result of this approach, thousands of annotations will be assigned to objects representing a small number of classes, while the annotations with more general classes will be inherited automatically. The WordNet hierarchical trees are very detailed, that is way only hypernyms, which are visual objects are selected with only one abstract notion on the top. For example, *jersey* is a *shirt*, which, in turn, is a *clothing*. From the hierarchy the node *garment* (an article of clothing) between *shirt* and *clothing* is excluded.

The Ontology design organized the 851 concepts into 11 categories on the top level, such as *person, animal, furniture, equipment* and so on (approximately half of the Ontology classes are contained in WordNet, 485 out of 851 classes)).

Following the strategy for category selection of the ImageNet, we applied the rule for no overlapping between the dominant classes and their attributes: “for any synsets  $i$  and  $j$ ,  $i$  is not an ancestor of  $j$ ” (Deng et al., 2009). Mutually exclusive classes are also defined for other well-known datasets, for example for the COCO thing and stuff classes (Caesar et al., 2018). As pointed out, the mutual inclusion might lead to some inconsistencies. An example was given with the PASCAL Context (Mottaghi et al., 2014) classes *bridge* and *footbridge*, which are in a parent-child relation (Caesar et al., 2018). The parent term can replace the child term in some context, but not vice versa; thus: if two images are annotated as *bridge* and *footbridge* respectively, it will not be known whether the parent concept can refer also to the child concept or not.

**Attributes** in the ontology are classes related with the dominant ones. The type of the dominant class and the type of attribute class determine the type of the relation between them, which expresses the specificity of property attribution: **wears, uses,**



Figure 1: Attribute classes in the Ontology

**has part**, etc. For example, the attribute classes for *cricketer* are *cricket bat*, *cricket ball*, *cricket helmet*, *wicket* and *referee*, while for *climber* – *climbing helmet*, *chalk bag*, *claiming backpack*; the attribute classes for *chess player* are *chessman* and *chessboard*, and for the *figure skater* – *skate* and *leotard* (Figure 1), and so on.

For the definition of attribute classes, we use some WordNet relations, such as meronymy. In most of the cases, such relations are not overtly established in WordNet and they are additionally defined in the Ontology.

#### 4.2 Relations

The Ontology not only specifies the visual concepts, but also defines the relationship between concepts. Thus the **relations** used in the Ontology are relations between classes. The **is-a** relation is inherited from WordNet, where nouns build hierarchical structures based on the relations of **hypernymy** and **hyponymy**, assuming that WordNet contains representation for both members of the relation. When it comes to new concepts (not presented in WordNet), they are connected to the proper parent concept in WordNet.

Depending on their properties, the relations do or do not project hierarchical structures. Hierarchical relations (relations of inclusion) are of three basic types – taxonomic (classificatory, which associate an entity of a particular type with an entity of a more generic type), meronomic (expressing the relation of the whole to its parts) and proportional series (expressing proportions between values in a given series) (D. A. Cruse, 1996). Taxonomic relations are inverse and transitive (**is-a**) and meronomic relations are also inverse and could be transitive (**has part**). Non-hierarchical relations are inverse and non-transitive (most of the relations between dominant classes and their attribute classes),

Relation	Reverse R	Number
has hyponym	is hyponym of	827
wears	is worn by	241
has part	is part of	210
uses	is used by	119
is next to		34
plays with	is a devise for	23
is on	is a surface for	22
drives	is driven by	18
plays	is played by	17
is in	is around	15
operates	is operated by	14
propel	is propelled by	12
plays at	is where to play	10
creates	is created by	9
rides	is ridden by	9

Table 1: Types of relations and number of their occurrences

and symmetric, irreflexive and non-transitive (**is next to**).

Relations between dominant and attribute classes are not hierarchical. For the linking of attribute classes, we use one WordNet relation – **has part** and 13 relations that are not overtly established in WordNet and are additionally created for the Ontology, for example, (**wears**, **is next to** and **plays with**). Altogether, 15 relations are used in the Ontology, with 827 instances of the **is a** relation; 241 instances of the **wears** relation, 210 instances of the **has part** relation, and so on. Table 1 shows the relations included in the Ontology of Visual Objects, their properties and number of occurrences.

#### 4.3 Axioms

**Axioms** serve to model sentences that are always true (Gruber, 1995) and they can be used to infer new knowledge.

An axiom system for an ontology is a pair  $(AI, \alpha)$  where (i)  $AI$  is a set whose elements are called axiom identifiers and (ii)  $\alpha$  is a mapping. The elements of  $A := \alpha(AI)$  are called axioms (Cimiano and Handschuh, 2003).

Axioms are assertions that are driven by the properties of the relations. In the **Ontology of Visual Objects** the axioms are:

If  $X$  is a hypernym of  $Y$ , then  $Y$  is a hyponym of  $X$ .

If  $X$  is a hypernym of  $Y$ , and  $Y$  is a hypernym of  $Z$ , then  $X$  is also a hypernym of  $Z$ .

If  $X$  is a holonym of  $Y$ , then  $Y$  is a meronym of  $X$ .

If  $X$  is a holonym of  $Y$ , and  $Y$  is a holonym of  $Z$ , then  $X$  is also a holonym of  $Z$ .

If  $X$  plays  $Y$ , then  $Y$  is played by  $X$ .

If  $X$  wears  $Y$ , then  $Y$  is worn by  $X$ .

If  $X$  uses  $Y$ , then  $Y$  is used by  $X$ .

If  $X$  plays at  $Y$ , then  $Y$  is a place where  $X$  plays.

If  $X$  plays with  $Y$ , then  $Y$  is a device with which  $X$  plays.

If  $X$  is on  $Y$ , then  $Y$  is a surface on which  $X$  is.

If  $X$  rides  $Y$ , then  $Y$  is ridden by  $X$ .

If  $X$  propel  $Y$ , then  $Y$  is propelled by  $X$ .

If  $X$  drives  $Y$ , then  $Y$  is driven by  $X$ .

If  $X$  creates  $Y$ , then  $Y$  is created by  $X$ .

If  $X$  is in  $Y$ , then  $Y$  is around  $X$ .

If  $X$  is next to  $Y$ , then  $Y$  is next to  $X$ .

The set of non-hierarchical relations, which hold among target concepts, also holds among higher concepts, for example if a *soccer player* is next to a *referee*, then a *person* is next to a *person*.

#### 4.4 Ontology format

The concepts are represented by the respective WordNet ILI (Inter-Lingual-Index) number or an Ontology index (if the concepts are not represented in WordNet) and a unique label: either the most representative literal (synonym) from the WordNet synsets or a term picked as a more adequate to refer to the concept. The differentiation between dominant, attribute and only hypernym classes is explicitly stated. The relations between classes are also explicitly stated. In case of reverse relations, only the direct relation is encoded, and in case of symmetric relations only one record of the relation is encoded. The Ontology is defined in a JSON format. For example:

```
{
  "HYPERNYM_ID": "eng-30-09761310-n",
```

```
  "HYPERNYM_LEMMA": "accordionist",
  "RELATION": "IS A",
  "HYPERNYM_ID": "eng-30-10340312-n",
  "HYPERNYM_LEMMA": "musician"
},
```

The Ontology is intended to be language-independent but the concepts are attached manually with labels in English and Bulgarian. All Ontology classes (used as annotation labels) have been presented in 25 languages: English (Princeton WordNet), Bulgarian, Albanian, Basque, Catalan, Croatian, Danish, Dutch, German, Greek, Finnish, French, Galician, Icelandic, Italian, Lithuanian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovene, Spanish, Swedish. In providing translation equivalents to Ontology classes, priority is given to WordNet, employing openly available wordnets from the Extended Open Multilingual Wordnet project or official distribution webpages of particular wordnets. The synonyms of Ontology classes, the definitions of the concepts and some usage examples (if available) were extracted from the synsets in different languages. Where WordNet translations are not available, some additional sources of translations are employed: BabelNet and Machine translation (Koeva, 2021; Koeva et al., 2022).

## 5 Evaluation

A number of studies aimed at ontologies' evaluation are known (Hloman and Stacey, 2014; Vrandečić, 2009; Raad and Cruz, 2015; Walisadeera et al., 2016; Khalilian, 2019; Wilson et al., 2021). On their basis several criteria for the evaluation of ontologies can be defined directed to confirm the ontology quality and correctness:

- Accuracy states if the definitions of classes are correct.
- Completeness measures if the domain of interest is appropriately covered.
- Conciseness states that the ontology does not include any unnecessary or useless definitions or explicit redundancies between definitions of terms do not exist.
- Adaptability measures if the ontology offers the conceptual foundation for a range of anticipated tasks.



- Clarity measures how effectively the ontology communicates the intended meaning of the defined concepts.
- Computational efficiency measures the ability of the used tools to work with the ontology.
- Consistency describes that the ontology does not include or allow for any contradictions.

We can define our approach for the evaluation as a corpus-based approach (Raad and Cruz, 2015). Instead of comparing an ontology with the content of a text corpus that covers significantly a given domain, we use the image annotation process to evaluate the Ontology of Visual Objects. At the beginning, we have identified 1,037 classes grouped in ten thematic domains: Sport, Medicine, Arts, Education, Food, Transport, Clothing, Security, Indoors, and Nature. For four of them (Sport, Transport, Arts and Security) an evaluation of the Ontology classes is performed during the annotation: whether a class is a visual object or not; whether all depicted objects in selected images can be described with the Ontology classes; and whether new classes can be added if necessary.

For the definition of classes we rely on the definition of concepts in WordNet; the definition of new classes is provided by means of finding their correct place within the WordNet taxonomy by linking them with already defined concepts. Finally, we made some evaluation tests for all selected classes with other sources providing lists with concrete objects, such as concreteness ratings (Brysbaert and Biemiller, 2017), word acquisition ratings (in our case of nouns) (Kuperman et al., 2012) and picture dictionaries (Parnwell, 2008).

## 6 Conclusion and Future Work

To improve object annotation and classification, several approaches based on ontologies have been proposed. However, image classification and annotation remain a challenging problem and one of the reasons is possible overlapping of selected classes. The use of a specially designed ontology improves the speed of object annotation as well as the accuracy of object classification.

Our contributions consist of the following:

(1) Definition of an Ontology of Visual Objects, whose classes are sufficient to annotate objects in 130 thematic subdomains related in four general domains;

(2) Introduction of attribute classes, which, in general, are related to the location, function and context of objects in focus (the dominant classes);

(3) Definition of relations between dominant and attribute classes and formulations of axioms based on the properties of the relations. This offers a reliable means for automatic object or image description, automatic assignment of image captions or classification of images and objects.

Using the **Ontology of Visual Objects** ensures the selection of mutually exclusive classes, built-in interconnectivity of classes via formal relations, and the ability to easily extend the proposed ontology with more concepts corresponding to visual objects.

Applying semantics can improve not only the performance of object recognition but also the performance and quality of individual tasks required for object recognition, such as image segmentation. Furthermore, the Ontology can be used to reduce the gap between human image comprehension and machine image interpretation, allowing for better automation in training neural networks (Bhandari and Kulikajevs, 2018).

A possible application of the Ontology of Visual Objects includes further use of the relations to compile bigger training datasets (for example, utilizing higher level concepts) or to construct contexts in which a particular object may or may not appear. The Ontology of Visual Objects provides options for extracting: relationships between annotated objects, diverse datasets with different levels of granularity of object classes and appropriate sets of images illustrating different thematic domains.

The ontological organization of object classes provides data for learning associations between objects in images, for identifying relations between objects and for aligning objects and relations with text fragments. Last but not least, using the Ontology of Visual Objects enables the dataset to be expanded based on the particular needs.

## Acknowledgments

The Multilingual Image Corpus (MIC21) project was supported by the European Language Grid project through its open call for pilot projects. The European Language Grid project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 825627 (ELG).

## References

- Alejandro Agostini, Mohamad Javad Aein, Sandor Szedmak, Eren Erdal Aksoy, Justus Piater, and Florentin Würgüter. 2015. [Using structural bootstrapping for object substitution in robotic executions of human-like manipulation tasks](#). In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6479–6486. IEEE.
- Peter Auer, Isabelle Bloch, Hilary Buxton, Patrick Courtney, Sven Dickinson, Bob Fisher, Goesta Granlund, Walter Kropatsch, Giorgio Metta, Bernd Neumann, Axel Pinz, Giulio Sandini, Gerald Sommer, David Vernon, Aude Billard, Pia Boettcher, Henrik Christensen, Andrew Crookell, Christof Eberst, Wolfgang Forstner and Vaclav Hlava and Ales Leonardis, Hans-Hellmut Nagel, Heinrich Niemann, Fiora Pirri, Bernt Schiele, John Tsotsos, Markus Vincze, Horst Bischof, Heinrich Bulthof and Tony Cohn and James Crowley and Jan-Olof Eklund and John Gilby and Josef Kittler, Jim Little, Bernhard Nebel, Lucas Paletta, Gerhard Sagerer, Rebecca Simpson, and Monique Thonnat. 2005. *CA Research Roadmap of Cognitive Vision*. ECVision: The European Research Network for Cognitive Computer Vision Systems.
- Sandeepak Bhandari and Audrius Kulikajevas. 2018. [Ontology Based Image Recognition: A Review](#). In *Proceedings of the International Conference on Information Technologies*, pages 13–18.
- Jan Botorek, Petra Budíková, and Pavel Zezula. 2014. [Visual Concept Ontology for Image Annotations](#). *CoRR*, abs/1412.6082.
- Erol Bozsak, Marc Ehrig, Siegfried Handschuh, Andreas Hotho, Alexander Maedche, Boris Motik, Daniel Oberle, Christoph Schmitz, Steffen Staab, Ljiljana Stojanovic, et al. 2002. [KAON — towards a large scale Semantic Web](#). In *International Conference on Electronic Commerce and Web Technologies*, pages 304–313. Springer.
- M. Brysbaert and A. Biemiller. 2017. [Test-based age-of-acquisition norms for 44 thousand English word meanings](#). *Behavior research methods*, 49(5):1520–1520.
- Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. [COCO-stuff: Thing and stuff classes in context](#). In *Conference on Computer Vision and Pattern Recognition*, pages 1209–1218.
- Parkpoom Chaisiriprasert, Karn Yongsiriwit, Matthew N. Dailey, and Chutiporn Anutariya. 2021. [Ontology-based Framework for Cooperative Learning of 3D Object Recognition](#). *Applied Sciences*, 11(17).
- Philipp Cimiano and Siegfried Handschuh. 2003. [Ontology-based Linguistic Annotation](#). In *Proceedings of the ACL 2003 Workshop on Linguistic Annotation: Getting the Model Right*, pages 14–21, Sapporo, Japan. Association for Computational Linguistics.
- Óscar Corcho, Mariano Fernández-López, and Asunción Gómez-Pérez. 2006. [Ontological Engineering: Principles, Methods, Tools and Languages](#). In Coral Calero, Francisco Ruiz, and Mario Piattini, editors, *Ontologies for Software Engineering and Software Technology*, pages 1–48. Springer.
- D. A. Cruse. 1996. *Lexical Semantics*. Cambridge University Press, Cambridge.
- Jia Deng, Wei Dong, Socher Richard, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. [ImageNet: A large-scale hierarchical image database](#). *2009 IEEE Conference on Computer Vision and Pattern Recognition*, page 248–255.
- Zheyuan Ding, Li Yao, Bin Liu, and Junfeng Wu. 2019. [Review of the Application of Ontology in the Field of Image Object Recognition](#). In *Proceedings of the 11th International Conference on Computer Modeling and Simulation, ICCMS 2019, North Rockhampton, QLD, Australia, January 16-19, 2019*, pages 142–146. ACM.
- Christiane Fellbaum, editor. 1999. *WordNet: an Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Greg Griffin, Alex Holub, and Pietro Perona. 2007. [Caltech-256 object category dataset](#). In *Technical Report 7694*, page 1–20, California Institute of Technology.
- Thomas R. Gruber. 1995. [Toward principles for the design of ontologies used for knowledge sharing](#). *Int. J. Hum.-Comput. Stud.*, 43:907–928.
- Hlomani Hlomani and Deborah Stacey. 2014. [Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey](#). *Semantic Web Journal*, 1(5):1–11.
- Saiede Khalilian. 2019. [A Survey on Ontology Evaluation Methods](#). *Quarterly Knowledge and Information Management Journal*, 6(2):25–34.
- Svetla Koeva. 2021. [Multilingual Image Corpus: Annotation Protocol](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 701–707, Held Online. INCOMA Ltd.
- Svetla Koeva, Ivelina Stoyanova, and Jordan Kravev. 2022. [Multilingual Image Corpus – Towards a Multimodal and Multilingual Dataset](#). In *Proceedings of the Language Resources and Evaluation Conference*, pages 1509–1518, Marseille, France. European Language Resources Association.
- Victor Kuperman, H. Stadthagen-Gonzalez, and Marc Brysbaert. 2012. [Age-of-acquisition ratings for 30,000 English words](#). *Behavior Research Methods*, 44:978–990.

- Egoitz Laparra, German Rigau, and Piek Vossen. 2012. [Mapping WordNet to the Kyoto ontology](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2584–2589, Istanbul, Turkey. European Language Resources Association (ELRA).
- Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. [Microsoft COCO: Common Objects in Context](#). In *European Conference on Computer Vision (ECCV)*, pages 740–755, Zürich.
- George Miller. 1986. [Dictionaries in the mind](#). *Language and Cognitive Processes*, 1:171–185.
- George Miller, R. Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. [Introduction to WordNet: An on-line lexical database](#). *International Journal of Lexicography*, 3:235–244.
- Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Namgyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. 2014. [The role of context for object detection and semantic segmentation in the wild](#). In *Conference on Computer Vision and Pattern Recognition*, pages 891–898.
- Milind R. Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston H. Hsu, Lyndon S. Kennedy, Alexander G. Hauptmann, and Jon Curtis. 2006. [Large-scale Concept Ontology for Multimedia](#). *IEEE Multim.*, 13(3):86–91.
- Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. [Ten Years of BabelNet: A Survey](#). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4559–4567. International Joint Conferences on Artificial Intelligence Organization. Survey Track.
- E. C. Parnwell. 2008. *The New Oxford Picture Dictionary*. Oxford University Press, New York, Oxford.
- Joe Raad and Christophe Cruz. 2015. [A survey on ontology evaluation methods](#). In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*.
- Luis Riazuelo, Moritz Tenorth, Daniel Di Marco, Marta Salas, Dorian Gálvez-López, Lorenz Mösenlechner, Lars Kunze, Michael Beetz, Juan D Tardós, Luis Montano, et al. 2015. [RoboEarth semantic mapping: A cloud enabled knowledge-based approach](#). *IEEE Transactions on Automation Science and Engineering*, 12(2):432–443.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. [ImageNet large scale visual recognition challenge](#). *International Journal of Computer Vision*, 116:157–173.
- Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2008. [LabelMe: a database and web-based tool for image annotation](#). *International Journal of Computer Vision*, 77:157–173.
- Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. [Yago: A Core of Semantic Knowledge](#). In *Proceedings of the 16th International Conference on World Wide Web, WWW '07*, pages 697–706, New York, NY, USA. ACM.
- He Tan and Patrick Lambrix. 2009. [Selecting an Ontology for Biomedical Text Mining](#). In *Proceedings of the BioNLP 2009 Workshop*, pages 55–62, Boulder, Colorado. Association for Computational Linguistics.
- Anne-Marie Tousch, Stéphane Herbin, and Jean-Yves Audibert. 2012. [Semantic hierarchies for image annotation: A survey](#). *Pattern Recognit.*, 45(1):333–345.
- David Vernon. 2006. [The Space of Cognitive Vision](#). In *Cognitive Vision Systems, Sampling the Spectrum of Approaches [based on a Dagstuhl seminar]*, pages 7–24.
- David Vernon. 2021. *Cognitive Vision*, pages 164–167. Springer International Publishing, Cham.
- Denny Vrandečić. 2009. [Ontology evaluation](#). In *Handbook on ontologies*, pages 293–313. Springer.
- Anusha Indika Walisadeera, Athula Ginige, and Gihan Nilendra Wikramanayake. 2016. [Ontology evaluation approaches: a case study from agriculture domain](#). In *International Conference on Computational Science and Its Applications*, pages 318–333. Springer.
- RSI Wilson, JS Goonetillake, WA Indika, and Athula Ginige. 2021. [Analysis of Ontology Quality Dimensions, Criteria and Metrics](#). In *International Conference on Computational Science and Its Applications*, pages 320–337. Springer.
- Xiao Xie, Xiran Zhou, Jingzhong Li, and Weijiang Dai. 2020. [An Ontology-based Framework for Complex Urban Object Recognition through Integrating Visual Features and Interpretable Semantics](#). *Complexity*, 44:1–15.
- Eleftherios Spyromitros Xioufis, Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis P. Vlahavas. 2011. [MLKD's Participation at the CLEF 2011 Photo Annotation and Concept-based Retrieval Tasks](#). In *CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands*.

# Sense-Annotated Corpus for Russian

**Alexander Kirillovich**  
Higher School of Economics  
Kazan Federal University  
Moscow & Kazan, Russia  
alik.kirillovich@gmail.com

**Natalia Loukachevitch**  
Lomonosov Moscow State University  
Institute for System Programming of RAS  
Moscow, Russia  
louk\_nat@mail.ru

**Maksim Kulaev**  
Higher School of Economics  
Moscow, Russia  
kulaevma@yandex.ru

**Angelina Bolshina**  
Moscow State University  
Moscow, Russia  
angelina\_ku@mail.ru

**Dmitry Ilvovsky**  
Higher School of Economics  
Moscow, Russia  
dilv\_ru@yahoo.com

## Abstract

We present a sense-annotated corpus for Russian. The resource was obtained by manually annotating texts from the OpenCorpora corpus, an open corpus for the Russian language, by senses of Russian wordnet RuWordNet. The annotation was used as a test collection for comparing unsupervised (Personalized Pagerank) and pseudo-labeling methods for Russian word sense disambiguation.

**Keywords:** corpus linguistics, word sense disambiguation, wordnet, Russian

## 1 Introduction

The task of automatic word sense disambiguation is the central task of automatic semantic analysis of texts and consists in choosing the correct word sense in the context of its use. The best results in this task have been achieved through the use of machine learning methods, which are based on preliminary manual annotation of a text corpus by lexical senses.

Most existing text collections for word sense disambiguation are annotated using sense inventory of WordNet-like resources (Miller et al., 1990; Petrolito and Bond, 2014; Pasini et al., 2021). In this paper we consider a new corpus annotated word senses for Russian, which uses the word sense inventory of Russian wordnet - RuWordNet (Loukachevitch et al., 2016). We also test some baseline methods using the created corpus such as the most frequent sense (MFS), unsupervised personalized pagerank method (Agirre and Soroa, 2009; Agirre et al., 2018), and pseudolabeling based on so-called monosemous relative approach (Martinez et al., 2008; Bolshina and Loukachevitch, 2020a).

## 2 Related work

### 2.1 WSD methods

The best results for automatic methods for word sense disambiguation are achieved by supervised methods (Bevilacqua et al., 2021; Pasini et al., 2021). The training of such methods requires manual sense annotation of a large text corpus, which is a laborious work. Large semantically annotated corpora are available mostly for English (Pasini et al., 2021).

There can be two main approaches to reduce data labeling costs. The first approach is based on automatic annotation of data using some additional resources, so-called automatic pseudolabeling. Pseudo-labeling methods can be based on different techniques of annotation such as parallel text collections (Taghipour and Ng, 2015), monosemous related words (so called monosemous relatives) (Martinez et al., 2008) and others. Such automatically annotated data are then used for training supervised methods.

The second group of methods are unsupervised methods, which do not require any labelled dataset for disambiguation. Such methods usually use manual dictionaries or thesauri (such as wordnets), their inventories of senses and corresponding information (word sense definitions, relations between words and senses) to disambiguate words (Navigli and Lapata, 2009; Moro et al., 2014; Agirre and Soroa, 2009). They are the most useful ones in case of dealing with low-resource data or modelling of some link-based dependencies.

The main assumption for unsupervised WSD is that semantically-related senses are presented in similar contexts. In this case a method of disambiguation should include a semantic similarity



metric. In graph-based techniques an analogue of such metric may be a link between entities in a graph. Therefore, it is possible to calculate semantic similarity based on the length of the shortest path between nodes.

One of the most known unsupervised method applied for word sense disambiguation is PageRank method (Agirre and Soroa, 2009; Duque et al., 2018), which was initially proposed for calculating authoritative Internet pages and based on page links (Page et al., 1999). In word sense disambiguation, PageRank is applied to graph-based semantic resources such as WordNet.

## 2.2 Word Sense Disambiguation in Russian

For Russian, in (Loukachevitch and Chuiko, 2007) the authors studied the all-word disambiguation task on the basis of the RuThes thesaurus (Loukachevitch et al., 2018) - resource for natural language processing of Russian texts. They experimented with various parameters (types of the thesaurus paths, window size, etc). The work (Kobritsov et al., 2005) describes developed disambiguation filters to provide semantic annotation for the Russian National Corpus. The semantic annotation was based on the taxonomy of lexical and semantic facets. In (Mitrofanova and Lyashevskaya, 2009) statistical word sense disambiguation methods for several Russian nouns were described. Alexeyevsky and Temchenko (Alexeyevsky and Temchenko, 2016) tested a number of algorithms based on parsing of monolingual dictionaries.

In (Bolshina and Loukachevitch, 2020a) the authors study an approach to automatic semantic annotation of a text corpus based on so called "monosemous relatives" technique, which exploits monosemous related words. The proposed approach involves not only monosemous synonyms, hyponyms or hypernyms as usual, but also "far" relatives located up to four relations from the initial sense according to Russian wordnet RuWordNet (Loukachevitch et al., 2016). Gathered related words are then filtered according to corpus-based vector similarity to synsets corresponding senses of the target word. In such a way, the approach allows adapting to specific genre-specific or domain collections (Bolshina and Loukachevitch, 2020b).

In (Panchenko et al., 2018) the authors describe the results of the first shared task on word sense induction (WSI) for the Russian language. The par-

ticipants were asked to group contexts of a given word in accordance with its senses that were not provided beforehand. For the task, new evaluation datasets based on sense inventories with different sense granularity were created. The contexts in the datasets were sampled from texts of Wikipedia, the academic corpus of Russian, and an explanatory dictionary of Russian. In the Russian SuperGLUE benchmark (Shavrina et al., 2020) the datasets from RUSSE-2018 were transformed into the Word-in-Context task, which is a binary classification task: given two sentences containing the same polysemous word, the task is to determine, whether the word is used in the same sense in both sentences, or not.

Thus we see that some research has been done for word sense disambiguation in Russian. But by this time there is no text corpus annotated with word senses. The above-mention annotation in the Russian National Corpus is based on general semantic categories, not specific word senses.

## 3 Sense-annotated collection

For creating a sense-annotated collection, we use texts collected in the OpenCorpora project<sup>1</sup>. The OpenCorpora corpus gathered Russian texts and develop several layers of annotation for the open use of these data by researchers (CC BY-SA license) (Bocharov et al., 2011). Currently, the Opencorpora corpus has a subcorpus with morphological annotation annotated by crowdsourcing. The morphological corpus was used for developing one of the most known Russian morphological analyzers PyMorphy2 (Korobov, 2015). But the OpenCorpora does not contain texts with word sense annotation.

### 3.1 RuWordNet

For word sense annotation, we use sense inventory of Russian lexical-semantics resource RuWordNet<sup>2</sup> (Loukachevitch et al., 2016; Nikishina et al., 2022). RuWordNet is a resource similar to WordNet (Miller et al., 1990). It was semi-automatically created from other Russian resource - RuThes thesaurus (Loukachevitch et al., 2018). As other WordNet-like resources, RuWordNet consists of synsets, connected with semantic relations. Current RuWordNet version includes more than 133 thousand Russian words and expressions of three parts

<sup>1</sup><http://opencorpora.org/>

<sup>2</sup>[ruwordnet.ru](http://ruwordnet.ru)

Entity type	Count
Synset	59,905
Lexical entry	133,468
Word	71,365
Multiword expression	62,103
Sense	154,111
Synset relation	254,007
hypernym / hyponym	74,736
instance hypernym / hyponym	5,803
part holonym / meronym	3,450
antonym	922
entailment	1,033
cause	568
domain topic	38,608
POS synonym	44,898
Link to inter-lingual index	23,162
Definition	20,054

Table 1: RuWordNet statistics.

of speech: nouns, verbs and adjectives. RuWordNet contains more than 15 thousand ambiguous Russian words presented in more than 20 thousand synsets. Tables 1 presents detailed RuWordNet statistics.

### 3.2 Manual sense annotation

For sense annotation, texts of average length were selected from the OpenCorpora corpus, beginning from texts containing several sentences. The texts were subdivided into sentences, lemmatized, matched with RuWordNet lexical entries, and transformed into the text format covering maximal information, useful for selecting an appropriate word sense in context. The created format presents the following items in structure:

- sentence,
- list of words in a column,
- each word is associated with a lemma and a part of speech,
- list of senses for each word found in RuWordNet,
- each sense is provided with the synset name, synonyms and hypernyms, presenting several levels up along the RuWordNet hierarchy.

Main statistics of the annotated corpus is presented in Table 2.

Metric	Num
Documents	807
Sentences	6,751
Lemmas	109,893
Annotated lemmas	46,320
Lexical entries	17,126
Annotated lexical entries	10,683
RWN synsets	8,619

Table 2: Description of the collection.

## 4 Evaluation of WSD methods on the collection

We experimented with two approaches for Russian word sense disambiguation: unsupervised PageRank method and automatic pseudo-labeling based on 'monosemous relatives'.

### 4.1 Applying PageRank for Russian word sense disambiguation

The assumption is that it is possible to solve WSD task for Russian as well as for English using PageRank. However, a WordNet-like database should be used to correctly repeat all steps. RuWordNet enables us to apply it because its structure is close to the structure of original WordNet.

The main idea of PageRank is to calculate the relative importance of a node (rank) in the graph  $G$ . It may be calculated using a number of directed links incoming a considered node. Besides, the strength of the link from  $i$  to  $j$  depends on the rank of node  $i$ : the more important node  $i$  is, the more strength its votes will have. Alternatively, PageRank can also be viewed as the result of a random walk process, where the final rank of node  $i$  represents the probability of a random walk over the graph ending on node  $i$ , at a sufficiently large time.

The calculation of the PageRank vector  $Pr$  for  $N$  nodes of graph  $G$  is equivalent to resolving the following equation:

$$Pr = cM \cdot Pr + (1 - c) \cdot v$$

where  $M$  is  $N \times N$  transition probability matrix,  $M_{ij} = \frac{1}{d_i}$ ,  $d_i$  is the number of outbound links of node  $i$ .  $V$  is a  $N \times 1$  vector whose elements are  $\frac{1}{N}$  and  $c$  is the so called damping factor, a scalar value between 0 and 1. The first term of the sum represents the above-described voting scheme. The second term correspond to the probability of a surfer

Procedure	Train	Test
Random	63.9	63.6
Most frequent sense	85.7	71.1
Pseudo-labelling	73.6	74.1
Basic PPR	-	67.4
PPR with a subset of relations	-	71.1
(previous) & not incl. target word	-	73.7
(previous) & hyperparameter optimization (damping_factor=0.95, n_iter=30)	73.7	74.2
(previous) & sliding window optimization (w=5)	74.2	74.3
(previous) & collocations	75.0	75.4

Table 3: Precision of considered methods.

randomly jumping to any node, e.g. without following any paths on the graph. The second term in the equation can be seen as a smoothing factor that makes any graph fulfill the property of being aperiodic and irreducible. It allows avoiding deadlocks and loops in the graph, thereby guaranteeing that PageRank calculation converges to a unique stationary distribution (Page et al., 1999).

In the traditional PageRank formulation the vector  $v$  assigns equal probabilities to all nodes in the graph in case of random jumps. However, the vector  $v$  can be modified to be non-uniform. For example, stronger probabilities can be assigned to certain kinds of nodes - creating so called Personalized PageRank (PPR) method (Haveliwala, 2003).

In (Agirre and Soroa, 2009), the authors applied the PPR algorithm to word sense disambiguation based on WordNet (Miller, 1995) and showed that the results are better than for other graph-based algorithms.

To apply the PPR algorithm, several steps should be performed:

1. Determine types of relations between synsets of WordNet-like resource to be used. Some relations may be weak and may add noise to this graph. It is proposed to save the following relations: part meronym, part holonym, instance hyponym, instance hypernym, hyponym, hypernym.
2. Convert this resource to a graph.
  - (a) Each sense corresponds to a node,
  - (b) Each selected relation corresponds to an edge.

3. Decide whether a target word will be included in this context graph while solving disambiguation or not. The main benefit of the first variant is that it is more computationally effective. However, it leads to a problem of importance increase of related senses in the context (Agirre and Soroa, 2009). In the second variant, for each target word  $W_i$ , initial probability mass is concentrated in the senses of the words surrounding  $W_i$ , but not in the senses of the target word itself, so that context words increase its relative importance in the graph (Agirre and Soroa, 2009).
4. Determine a sliding context window, i.e. a number of words before and after a target one to be considered as a context.
5. Set PPR hyperparameters – number of iterations and damping factor (probability of random jumps).

Changes in each of these steps lead to different realisations of this method. Then, a resulting algorithm is the following:

1. For each TEXT in COLLECTION:
  - (a) For each TARGET\_WORD in TEXT:
    - i. Take CONTEXT\_WORDS using WINDOW.
    - ii. Insert CONTEXT\_WORDS in a graph – create a directed link from them to their possible senses.
    - iii. Declare PPR method and assign initial probability mass to nodes of CONTEXT\_WORDS .



- iv. Fit PPR on this graph.
- v. Take all possible senses of TARGET\_WORD and their final probabilities.
- vi. Choose a sense with a maximum probability.

It can be seen from Table 1 that RuWordNet contains a large number of multiword expressions (collocations). For each collocation, senses of word components (sense\_id) are described. For example, component senses of phrase "отвратительный на вид" (disgusting looking) are described as follows:

- <sense name="отвратительный" id="118920-A-145306" synset\_id="118920-A"/>
- <sense name="вид" id="107545-N-134500" synset\_id="107545-N"/>

Therefore the PPR algorithm may be modified using collocations from the RuWordNet knowledge base. Collocations can be inserted in a graph, they also may be considered as an additional information for disambiguation. There are two ways of introducing collocations into the algorithm implementation:

1. Take a sense for target word from an expression if it is a component of such expression in the given text.
2. Use tokens of collocations contained in the context to resolve disambiguation of other words.

The first method is simpler because it does not require to consider context while resolving disambiguation.

This method was implemented for both original and personalized ways. Moreover, hyperparameters were optimized and some of previously mentioned improvements were introduced. Results will be presented in the appropriate section.

## 4.2 Pseudo-labeling method

Automatic pseudo-labeling method is based on the monosemous relative technique. The related monosemous words or expressions can be located on the distance up to 4 RuWordNet relations from

the initial sense (Bolshina and Loukachevitch, 2020a). For example, a single-sense co-hyponym can serve as a monosemous relative (2 relations).

We suppose that contexts of monosemous relatives can be appropriate for the target sense and we can use for training disambiguation models. Any monosemous relative in fact can be quite different in context of usage from the target sense, therefore additional check and selection of monosemous relatives are needed. The monosemous relatives of the target words are additionally scored in accordance to the cosine similarity between word2vec vector of the relative and averaged vector of so-called *synset nest*.

The synset nest represents a set of words (or phrases) most closely related to a particular sense of the target word, specifically target word synonyms and all the words from directly related synsets within two steps from the target word (Bolshina and Loukachevitch, 2020a). A fragment of the nest for the Russian word *taksa* (“dachshund”) is as follows: *hunting dog, hunting dog, doggie, four-legged friend, dog, dog, terrier, dog, greyhound dog...* (translated from Russian).

The word2vec vectors can be calculated on different text collections, which allows tuning of relative selection on the specific genre of texts (Bolshina and Loukachevitch, 2020b). The pseudolabeling includes the following steps:

- selection of monosemous related words for each sense of ambiguous word in RuWordNet at the distance up to 4 relations from the sense synset,
- scoring monosemous relatives according to word2vec similarity to the synset nests for each word sense calculated on a selected text corpus,
- extraction of monosemous relatives’ contexts for training a supervised model training taken in proportion to similarity scores between monosemous relatives and synset nest.

In the current study word2vec training and context extraction was implemented on a Russian news corpus (2 million documents). For each sense, 200 contexts originating from different monosemous relatives were extracted. For context representation, the ELMO model<sup>3</sup> was used. Logistic regression

<sup>3</sup><https://rusvectors.org/ru/models/>

model was trained for disambiguation of each ambiguous word on the automatically annotated word sense contexts.

### 4.3 Results

The approaches described in this article were implemented on the created corpus. Moreover, different settings and hyperparameters were tried. Precision was calculated as a performance measure of disambiguation methods. It was measured in two different ways: including one-sense words and not. This should be considered because a human annotator might indicate that there is no correct sense for this word (in the context, of course) in our knowledge base.

Some simple methods were considered as baselines. They include: the most frequent sense method and the random method. The sense annotated collection was randomly split on train and test sets (it makes sense only for a limited number of methods) to exclude over-fitting. Final results are presented in Table 3.

It can be seen that the most frequent sense method demonstrates the best performance on the training set and nearly the worst one on the test set. And it is notable that the unsupervised PPR method outperforms the supervised pseudo-labeling approach only when preliminary parameter setting and optimisation were conducted.

## 5 Conclusion

We presented a sense-annotated corpus for Russian. The total size of the corpus is 109,893 lemmas, out of which 46,320 ones are manually annotated by 8,619 RuWordNet synsets.

The obtained corpus was used as a test collection for evaluating two word-sense disambiguation methods: personalized PageRank and pseudo-labelling. The precision of PPR is 75.4% and the precision of pseudo-labelling is 74.1%.

Our future work will be undertaken in two directions: (1) Firstly, we are going to use the corpus not only as test data, but also as a training collection for supervised methods. (2) Secondly, we are going to further develop the corpus itself, including annotating multi-word expressions and publishing the corpus in the Linguistic Linked Open Data cloud.

The corpus has been published on GitHub: <https://github.com/LLOD-Ru/OpenCorpora-RuWordNet>.

## Acknowledgments

This work is supported by the Russian Science Foundation, grant no. 19-71-10056. The work of Natalia Loukachevitch in manual and automatic word sense annotation is supported by a grant for research centers in the field of artificial intelligence (agreement identifier 000000D730321P5Q0002 dated November 2, 2021 No. 70-2021-00142 with ISP RAS).

## References

- Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2018. The risk of sub-optimal use of open source nlp software: Ukb is inadvertently state-of-the-art in knowledge-based wsd. *ACL 2018*, page 29.
- Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41.
- Daniil Alexeyevsky and Anastasiya V Temchenko. 2016. Word sense disambiguation in monolingual dictionaries for building russian wordnet. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 9–14.
- Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.
- Victor Bocharov, Svetlana Bichineva, Dmitry Granovsky, Natalia Ostapuk, and Maria Stepanova. 2011. Quality assurance tools in the opencorpora project. *Computational linguistics and intellectual technologies*.
- Angelina Bolshina and Natalia Loukachevitch. 2020a. All-words word sense disambiguation for russian using automatically generated text collection. *Cybernetics and Information Technologies*, 20(4):90–107.
- Angelina Bolshina and Natalia Loukachevitch. 2020b. Comparison of genres in word sense disambiguation using automatically generated text collections. In *Fourth International Conference Computational Linguistics in Bulgaria*, page 155.
- Andres Duque, Mark Stevenson, Juan Martinez-Romo, and Lourdes Araujo. 2018. Co-occurrence graphs for word sense disambiguation in the biomedical domain. *Artificial intelligence in medicine*, 87:9–19.
- Taher H Haveliwala. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796.

- Bors Kobritsov, Olga Lyashevskaya, and Olga Shemanaeva. 2005. Disambiguation of lexico-semantic ambiguity in news and newspaper-magazine texts: surface filters and statistical evaluation. *Proceedings of the Contest "Internet Mathematics"*.
- Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In *International conference on analysis of images, social networks and texts*, pages 320–332. Springer.
- Natalia Loukachevitch and Daria Chuiko. 2007. Thesaurus-based word sense disambiguation [avtomaticheskoe razreshenie leksicheskoy mnogoznachnosti na baze tezaurusnykh znaniy]. *Proceedings of the Contest "Internet Mathematics"*, pages 108–117.
- Natalia Loukachevitch, German Lashevich, and Boris V Dobrov. 2018. Comparing two thesaurus representations for russian. In *Proceedings of the 9th Global Wordnet Conference*, pages 34–43.
- Natalia V Loukachevitch, German Lashevich, Anastasia A Gerasimova, Vladimir V Ivanov, and Boris V Dobrov. 2016. Creating russian wordnet by conversion. In *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*, pages 405–415.
- David Martinez, O Lopez de Lacalle, and Eneko Agirre. 2008. On the use of automatically acquired examples for all-nouns word sense disambiguation. *Journal of Artificial Intelligence Research*, 33:79–107.
- George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to Wordnet: an on-line lexical database. *International journal of lexicography*, 3(4):235–244.
- Olga Mitrofanova and Olga Lyashevskaya. 2009. Disambiguation of taxonomy markers in context: Russian nouns. In *17th Nordic Conference of Computational Linguistics NODALIDA—2009, Odense, Denmark*.
- Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.
- Roberto Navigli and Mirella Lapata. 2009. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE transactions on pattern analysis and machine intelligence*, 32(4):678–692.
- Irina Nikishina, Mikhail Tikhomirov, Varvara Logacheva, Yuriy Nazarov, Alexander Panchenko, and Natalia Loukachevitch. 2022. Taxonomy enrichment with text and graph vector representations. *Semantic Web*, 13(33):441–475.
- Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.
- Alexander Panchenko, Anastasiya Lopukhina, Dmitry Ustalov, Konstantin Lopukhin, Nikolay Arefyev, Alexey Leontyev, and Natalia Loukachevitch. 2018. Russe’2018: a shared task on word sense induction for the russian language. *arXiv preprint arXiv:1803.05795*.
- Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Xl-wsd: An extra-large and cross-lingual evaluation framework for word sense disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press.
- Tommaso Petrolito and Francis Bond. 2014. A survey of wordnet annotated corpora. In *Proceedings of the Seventh Global WordNet Conference*, pages 236–245.
- Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. Russiansuperglue: A russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726.
- Kaveh Taghipour and Hwee Tou Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 338–344.

# A Romanian Treebank Annotated with Verbal Multiword Expressions

**Verginica Barbu Mititelu**

Romanian Academy

RACAI

vergi@racai.ro

**Mihaela Cristescu**

University of Bucharest

mihaela.ionescu@litere.unibuc.ro

**Maria Mitrofan**

Romanian Academy

RACAI

maria@racai.ro

**Bianca-Mădălina Zgreabă**

Utrecht University

madalinazgreaban0@gmail.com

**Elena-Andreea Bărbulescu**

University of Bucharest

adabarbulescu7@gmail.com

## Abstract

In this paper we present a new version of the Romanian journalistic treebank annotated with verbal multiword expressions of four types: idioms, light verb constructions, reflexive verbs and inherently adpositional verbs, the last type being recently added to the corpus. These types have been defined and characterized in a multilingual setting (the PARSEME guidelines for annotating verbal multiword expressions). We present the annotation methodologies and offer quantitative data about the expressions occurring in the corpus. We discuss the characteristics of these expressions, with special reference to the difficulties they raise for the automatic processing of Romanian text, as well as for human usage. Special attention is paid to the challenges in the annotation of the inherently adpositional verbs. The corpus is freely available in two formats (CUPT and RDF), as well as queryable using a SPARQL endpoint.

**Keywords:** multiword expressions, Romanian, inherently adpositional verbs, idioms, light verb constructions.

## 1 Introduction

Language resources of the type electronic corpora annotated with syntactic information (most of the times on top of lexical and morphological annotations), i.e. treebanks, are now quite common for languages and even dialects. If a decade ago the number of treebanks for various languages was rather scarce, now we can find many such resources, though still of a modest size. The situation has greatly improved given the existence of two major multilingual initiatives: Universal Dependencies<sup>1</sup> (UD) (Nivre et al., 2016; de Marneffe et al., 2021) and PARSEME Cost Action (Savary et al., 2015), two open community efforts, active in improving and enhancing their results. UD is an ini-

tiative created with the aim of offering the instruments for a cross-lingual description at the morphologic and syntactic levels. Seventeen universal parts of speech (e.g., NOUN, VERB, AUX, PRON, ADJ, etc.) and a set of morphological features (e.g., Number, Gender, Tense, etc.) are used for the morphologic level, and 37 universal relations (e.g., `nsubj` for the nominal subject, `csubj` for the clausal subject, `obj` for the nominal direct object, `ccomp` for the clausal direct object, etc.) are defined for the syntactic description. These morphologic instruments are considered enough for the description of any language, while the inventory of syntactic relations is admittedly universal, but subtypes of the 37 universal relations are accepted for a more specific syntactic analysis: e.g., `nsubj:pass` for the nominal subject in passive constructions for the languages that do have passive; 26 such subtypes have been defined so far, which are specific to one or more languages. In its last release (May 2022), UD boasts 228 treebanks for 130 languages, all freely available.

The existence of treebanks for various languages released through UD has offered the premise for the development of automatic tools (Straka et al., 2016) that can be trained on these treebanks and further used to annotate new corpora. This paved the way to initiatives such as PARSEME, in which new corpora, collected according to certain requirements (such as text genre, size, license, etc.), were automatically morphosyntactically annotated with such tools and further enriched with a new level of annotation, i.e. semantic: verbal multiword expressions (VMWEs) were manually annotated following the same guidelines for all languages, that identify universal, quasi-universal and language-specific VMWE types. Within PARSEME, treebanks for 26 languages were annotated and one of them is for Romanian.

There are already several treebanks for Roma-

<sup>1</sup>[universaldependencies.org](http://universaldependencies.org)

nian freely available: within UD, there is the Romanian Reference Treebank (RRT) (Barbu Mititelu, 2018) (containing sentences from various text genres), Romanian Non-Standard treebank (Colhon et al., 2017) (containing sentences from old texts or from folklore), the medical treebank SiMoNERo (Barbu Mititelu and Mitrofan, 2020) (which has an extra annotation level: medical entities of the types anatomical parts, chemicals, disorders and procedures) and the treebank of the Aromanian dialect of Romanian ArT (Barbu Mititelu et al., 2021). There is also another treebank, unavailable in UD, LegalNERO (Păiș et al., 2021), which has a further level containing gold annotations for five entity classes: organizations, locations, persons, time expressions and legal resources mentioned in legal documents.

In this paper we present a new version of the Romanian treebank, whose annotation started in PARSEME and which has recently been enriched with a new type of verbal expressions, i.e. inherently adpositional verbs. We call this corpus *PARSEME-Ro*. We will first outline the context of development of this corpus, namely the PARSEME shared tasks (Section 2), then present some idiosyncrasies displayed by the verbal expressions occurring in the corpus (Section 3). We describe the corpus itself: its levels of annotation (Section 4) and problems raised by annotating the new type of verbal expressions. Some general statistics about the corpus and statistics about the VMWE types in the corpus are given in Section 6, before concluding the paper.

## 2 Context of development

PARSEME is an international and multilingual community aiming at identifying MWEs in running texts. Although so far the interest has manifested only for verbal MWEs in a concerted way, MWEs of other morphological classes will also be approached in a multilingual perspective. The PARSEME shared tasks editions 1.0 (Savary et al., 2017), 1.1 (Ramisch et al., 2018) and 1.2 (Ramisch et al., 2020) focused on the identification of VMWEs because of their challenging features: complex structure, discontinuity, variability, ambiguity (Savary et al., 2017). The main aim of this initiative is to eventually automatically recognize VMWEs in corpora. The annotation guidelines are unified across languages and have been enhanced from edition 1.0 (Savary et al., 2017) to edition 1.1 (Ramisch et al., 2018).

Based on the experience gathered in the annotation for edition 1.0, as well as on the types of VMWEs identified in the corpora, starting with edition 1.1 of PARSEME, the following types of VMWEs have been annotated (Savary et al., 2018):

- *Universal categories*, that are valid for all languages participating in the task:
  - **Light verb constructions** (LVCs) with two subcategories:
    - \* **LVC.full**, in which the verb is semantically totally bleached: EN *to give a lecture*, RO *a lua o decizie* (to make a decision), *a face parte* (to be part);
    - \* **LVC.cause**, in which the verb adds a causative meaning to the noun: EN *to give a headache*, RO *a da bătăi de cap* (to give a bad time), *a pune capăt* (to pun an end);
  - **Verbal idioms** (VIDs), which have at least two lexicalized components including a head verb and at least one of its dependents and is characterised by a high degree of semantic non-compositionality: EN *to go bananas*, RO *a trage pe sfoară* (to double-cross), *a o lua la goană* (to start running);
- *Quasi-universal categories*, valid only for some languages:
  - **Inherently reflexive verbs** (IRVs), in which the reflexive clitic either always co-occurs with a given verb or changes its meaning or subcategorization frame: EN *to help oneself*, RO *a se gândi* (to think), *a se face* (to become);
  - **Verb-particle constructions** (VPC), which are made up of a verb and a particle: EN *to do in*, *to eat up*; this type is not applicable to Romanian;
  - **Multi-verb constructions** (MVC), which are made up of two verbs: EN *to let go*, *to make do*; neither is this type applicable to Romanian;
- *Language-specific categories*, valid only for the language for which they are defined, unless other languages claim them as well: so far, only one such type has been defined, namely **inherently clitic verbs** for Italian: it consists of a verb and one or more non-reflexive clitics



that represent the pronominalization of one or more complements: IT *infischiarsene* (not to worry about);

- *Experimental category*, annotated in the post-annotation step: **Inherently adpositional verbs** (IAVs), made up of a verb and a preposition: EN *to rely on*, RO *a conta pe* (to rely on).

For each language, a team of linguists was trained to apply the guidelines<sup>2</sup> for identifying VMWEs in a corpus and classifying them into one of the existing categories. Simultaneously, quality of the annotation was acquired by applying semi-automatic methods for ensuring full coverage of the VMWEs in the corpus, as well as for their consistent classification.

This is the context in which the creation of PARSEME-Ro took place, alongside corpora annotated with VMWEs for other languages.

The three editions of the PARSEME Cost Action (1.0, 1.1, 1.2) covered 18, 20, and 14 languages, respectively, from several language families: Romance languages (French, Italian, Portuguese, Romanian, Spanish), Balto-Slavic languages (Bulgarian, Czech, Croatian, Lithuanian, Polish, Slovene), Germanic languages (German, English, Swedish, Yiddish), and other languages (Arabic, Greek, Basque, Farsi, Maltese, Hebrew, Hindi, Hungarian, Turkish, Chinese, Irish).

All the annotated corpora from the editions 1.0<sup>3</sup>, 1.1<sup>4</sup> and 1.2<sup>5</sup> are available for download, under the Creative Common license.

### 3 Characteristics of VMWEs Contributing to their (Automatic) Processing Difficulty

MWEs are defined as “idiosyncratic interpretations that cross word boundaries (or spaces)” (Sag et al., 2002). They are “lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity” (Baldwin and Kim, 2010).

The identification of VMWEs in texts is a well-known challenge for NLP applications, because of

<sup>2</sup><https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=home>

<sup>3</sup><https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2282>

<sup>4</sup><https://gitlab.com/parseme/sharedtask-data/tree/master/1.1>

<sup>5</sup><https://gitlab.com/parseme/sharedtask-data/-/tree/master/1.2>

their special characteristics, including discontinuity, overlaps, non-compositionality, heterogeneity, and syntactic variability. They are problematic not only for machines, but also for humans: on the one hand, for students learning a second language and, on the other, for native speakers who are exposed to rarer expressions.

One key characteristic of a VMWE is for it to be idiomatic. This property refers to “markedness or deviation from the basic properties of the component lexemes, and applies at the lexical, syntactic, semantic, pragmatic, and/or statistical levels” (Baldwin and Kim, 2010).

*Lexical idiomaticity* is displayed when one or more components of a VMWE are not *part of the conventional lexicon or are not used outside the respective VMWEs*: for Romanian, this is the case of the boldfaced words in the VMWEs *a-și aduce aminte* (‘to remember’) or *a avea habar* (‘to have a clue’). Various Romanian VMWEs conserve lexical or semantic archaisms as the boldfaced words in the following expressions show: *aduce aminte* (‘remind’), *nu da în brânci* (‘have a soft job’), *da ortul popii* (‘die’), *nu avea habar* (‘have a clue’), *veni de hac* (‘bear down’), *băga de seamă* (‘notice’), *nu da inima ghes* (be reluctant to), *scoate la iveală* (‘reveal’), *lua la rost* (‘chide’), etc.

On the *morphological* level, there are VMWEs that display 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. or with respect to possible derived forms: e.g. RO *a nu privi cu ochi buni* (not watch with eyes good, ‘to regard with disfavour’) is always used with the negative marker *nu* ‘not’. In addition, there are VMWEs that contain obsolete inflectional and derived forms, such as *a bate câmpii* (beat the fields ‘to beat around the bush’) (in which *câmpii*<sup>6</sup> is an old plural form of the word *câmp*, whose current plural form is *câmpuri*), *a pune pe roate* (pun on wheels ‘to get on its feet’) (in which *roate* is an old plural form of the word *roată*, whose current plural form is *roți*), *a lua cu binișorul* (take with wellness\_DIMINUTIVE ‘to let down easily’) (the diminutive noun *binișorul* is not currently used outside expressions) (Căpățână, 2007).

*Syntactic idiomaticity* arises when the syntax of the VMWE is not derived directly from that of its components. The syntactic level of description

<sup>6</sup>The form *câmpii* is the definite one for *câmpii*.



would include any restrictions on the word order of the VMWE components and of the possible dependents. For example, in the RO VMWE *a da ortul popii* (give coin-the to priest-the ‘to die’) the object *ortul* always precedes the indirect object *popii*, though Romanian allows for any order of the direct and indirect object in case of their co-occurrence (though with different pragmatic salience in each case).

*Semantic idiomaticity* means non-compositionality of the expression, i.e. the meaning of a MWE is not explicitly derivable from the semantics of its parts. VMWEs displaying semantic idiomaticity have frequently components with metaphoric (*a lua taurul de carne* take the bull of horns ‘to take the bull by the horns’), hyperbolic (*a crăpa de frig* crack by cold ‘to be very cold’) or metonymic (*a nu ridica un deget* not lift a finger ‘not to lift a finger’) meaning in addition to their literal meaning. Semantic idiomaticity may imply either the fact that the expression’s meaning is given rather by one of the components (see the descriptions for LVCs in Section 2) or the fact that the global sense of the expressions has nothing to do with the senses of the components: e.g. the words making up the VID *a tăia frunză la câini* (cut leaf for dogs ‘to dally’) have no semantic relation to the sense of the expression.

*Pragmatic idiomaticity* occurs when a VMWE is associated with a fixed set of situations or a particular context of use: see the case of greetings that are specific to the different parts of the day: e.g. EN *good morning*, RO *noapte bună* (night good ‘good night’), etc.

*Statistical idiomaticity* is triggered by the high frequency a particular combination of words occurs with: e.g. EN *black and white* is semantically equivalent to RO *alb-negru* (white-black), in spite of the lack of the conjunction and of the reversed order of the two components.

All these characteristics of expressions may prevent their correct automatic interpretation, but also their understanding in inter-human communication, needless to say their grammatically correct and semantically adequate usage by second language learners. These justify the necessity for (computational) linguists’ focusing more on phraseology. The insufficient attention paid to them leads to inconsistent terminology, inconsistent treatment of such units in lexicography, partial descriptions in

grammars and dictionaries and little focus on it in textbooks, though, admittedly, expressions are a touchstone of language command.

#### 4 Annotation Levels

The PARSEME corpus for Romanian (PARSEME-Ro) is journalistic and was automatically tokenized, part-of-speech tagged, lemmatized and syntactically parsed using UDPipe (Straka et al., 2016) trained on RRT. In a first step (consisting of all three annotation phases pertaining to the participation in the three editions of the shared tasks), the annotation of the different types of VMWEs was manual: the annotators identified and classified the VMWEs belonging to the LVC.full, LVC.cause, VID and IRV types. In the first edition four annotators were involved, in the second one there were three, and two participated to the last edition. Each annotator had a portion of the morpho-syntactically processed corpus to annotate: using the FLAT platform<sup>7</sup> (Savary et al., 2017), their task was to read the text, to spot a potential VMWE and, using the decision tree and the battery of tests from the PARSEME guidelines, to decide if the respective string was indeed a VMWE and specify its type. Only for a small portion of the data (2500 sentences) was the annotation double, so as to measure the agreement between annotators (Savary et al., 2017; Ramisch et al., 2018).

In a second, recent, step, IAVs were annotated in PARSEME-Ro. This time, the annotation was automatic, followed by manual validation and correction, in two phases. Starting from the list of 1,725 adpositional verbs created by Geană (2013), all their occurrences in the corpus were identified and annotated as IAVs. This was done automatically by using a Python script that performed a global search of the IAVs tokens within the corpus text. This search was enhanced to include a span window in order to capture situations where other tokens were interleaved with the actual IAV in the corpus text. In cases where several matches were found for one of the tokens of the IAV (this applies mostly to prepositions) the principle of the minimum distance length between the tokens was used. Finally, based on these matches, the corpus tokens found to correspond to an IAV were automatically annotated. Then the first phase of the manual validation and correction step followed: two annota-

<sup>7</sup>[github.com/proycon/flat](https://github.com/proycon/flat), [flat.science.ru.nl](http://flat.science.ru.nl)

tors, students in linguistics, were presented with all automatically annotated instances and, using an annotation platform, they could modify the annotations in the sense of deleting expressions or adjusting their size (i.e. adding or removing parts), using the BRAT tool (Stenetorp et al., 2012), integrated in the RELATE platform (Păiș et al., 2020).

Several sources of errors could be identified in the automatic annotation of IAVs:

- homonyms that had been erroneously part-of-speech-tagged as verbs: adjectives with participle origin (*scutite de la plată* ‘exempted from payment’), nouns zero-derived from participles (*în trecut la* ‘in the past at’), etc.;
- ambiguity: the structure verb + adposition is ambiguous between an IAV and a mere combination with a different meaning from that specific to the IAV construction: the combination *a se lovi de* (REFL.CL hit of ‘to bump into’) is an IAV in a sentence like *Copilul s-a lovit de perete*. (Child-the REFL.CL-has hit of wall ‘The child bumped into the wall.’), but not in the sentence *Copilul s-a lovit de dimineață* (Child-the REFL.CL-has hit of morning ‘The child got hit in the morning.’), where the same preposition introduces a time adverbial. A particular example of this type is represented by constructions that are structurally similar to prepositionally marked direct objects: e.g. *a lăsa pe* (leave on ‘to bend on’) (as in *Ion s-a lăsat pe spate*. ‘John leaned back.’) as opposed to *lăsa pe cineva* (leave/let someone): as in *lăsând-o orfană pe micuța Ornella* (leaving-CL3SgFem orphan PE little Ornella ‘leaving little Ornella orphan’), where PE is a marker of the direct object;
- overgeneration: the presence of the adposition in the context of the verb, although syntactically belonging to a phrase without direct dependence on the verb, is misinterpreted as being part of an IAV: *a lua două șunci de porc* (take two hams of pork): here, *de* is a preposition linking the noun *pork* to its nominal head *șunci*, not to the verb;
- the combination verb – adposition is already part of another VMWE: *a da în judecată* (give in trial ‘to sue’) is already classified as VID, thus no IAV is annotated in this case;

#	Total IAVS	correctly annotated	
		#	%
AUTO annot.	4,686	3,128	66.75
annot. 1	3,462	3,085	89.11
annot. 2	3,519	3,185	90.5
both annots	-	2,981	
<b>gold IAVs</b>	-	<b>3,311</b>	

Table 1: General statistics of the IAV annotation process

- using the span window to match IAVs that have interleaved tokens has made the algorithm match false-positives to a high degree (34% of all automatically annotated IAVs).

Consistency of annotation was ensured differently for each step: for the annotation in the context of the shared tasks, we used the consistency checking tools made available by the organizers (Savary et al., 2018), helping to spot the skipped occurrences of VMWEs, as well as inconsistent type assignment of the same VMWE.

For the step involving the annotation of IAVs, we envisaged a second phase of the validation and correction step: all cases of agreement between the two student annotators were considered correct decisions (see the 2,981 cases marked as “both annots” in Table 1). All cases of disagreement between them were further checked by two linguists experienced in the PARSEME annotation. Table 1 shows that two thirds of the automatically annotated IAV are actually correct IAVs and that the decision to automatically annotate them was a time saving one. They represent 94.47% of the IAVs that should have been annotated, i.e. of the cases called “gold IAVs” in the table and which are the result of the experienced annotators’ validation and correction of the two student annotators’ validations. Each individual initial manual validation covers almost 90% of all correct cases: see the last column of lines two and three in Table 1.

## 5 Defining and refining the class of IAVs annotated in PARSEME-Ro

PARSEME guidelines 1.2 define an IAV as follows: “It consists of a verb or VMWE and an idiomatic selected preposition or postposition that is either always required or, if absent, changes the meaning of the verb or VMWE significantly.”<sup>8</sup> Their annotation is done after the annotation of other VMWEs,

<sup>8</sup>[https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=050\\_](https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=050_)

because (i) adpositional verbs occurring within other VMWEs should not be annotated as IAVs, and (ii) VMWEs can also be adpositional, just like verbs. The annotation of IAVs in the PARSEME-Ro corpus aimed at marking only the adpositional verbs for the time being, so as to serve as an exercise that would reveal the challenges this type of constructions raises.

PARSEME guidelines offer only one test for IAVs, which is meant to show the semantic difference between the verb occurring with the adposition and its use without it: if, in response to a declarative statement containing the potential IAV, a question cannot be asked about the circumstances of the verbal event using the verb, but not the adposition, then the combination verb – adposition is annotated as IAV.

Geană (2013: p. 46) defines adpositional verbs as constructions in which the verb is “capable of getting a prepositional complement”, where the complement is defined as an obligatory valence of the verb, irrespective of its semantics. This means that in the class of adpositional verbs we can have examples in which the adposition is part of an adverbial, e.g. a place adverbial: *I live in London*. Geană (2013: p. 118) further distinguishes between adpositional verbs using as criterion the type of adposition, namely:

- merely **functional adpositions**: their sole role is to case-mark the nominal which is a thematic argument of the verb: e.g. *Noi ne bazăm pe ajutorul vostru*. (En. “We count on your help”) – the adposition *pe* imposes the accusative case on the noun *ajutorul*;
- **semi-lexical adpositions** (Corver and van Riemsdijk, 2013): in the case of verbs requiring a semantic argument, the adposition carries the specific semantic content and, at the same time, case-marks the nominal with that role: e.g. *Ne plimbăm pe alee*. (En. We walk on the alley.) – the adposition has a locative meaning and imposes the accusative case to the noun *alee*.

Testing the two types of examples against the PARSEME criterion, we notice that in the case of functional adpositions the test holds, as one cannot ask about the circumstances of the verbal event using the verb only, not also the adposition: \**Când*

Cross-lingual\_tests/070\_Inherently\_adpositional\_verbs\_\_LB\_IAV\_RB\_

no. of sentences	56,664
no. of tokens	1,014,908
no. of words	806,540
no. of verbal lemmas	61,323
no. of unique verbal lemma	3,815

Table 2: General statistics of the PARSEME-Ro corpus

*ne bazăm noi?* (En. “When do we count?”) is not a grammatically complete question in Romanian. However, in the case of semi-lexical adpositions, the test does not hold: asking a question like *Când vă plimbați?* (En. When do you walk?) is grammatically complete.

Given these remarks on the types of IAVs annotated in PARSEME-Ro, we consider that the annotated data will need some further refinement: adpositional verbs will need to be further classified into two subtypes: IAV.functional and IAV.semi-lexical. The existence of subclasses inside a class is not new for PARSEME: see the two subtypes of LVCs, namely LVC.full and LVC.cuase (Section 2). However, continuing the PARSEME custom of testing classes and subclasses against data in more languages before coining them officially, the next step we envisage is collaborating with teams working on IAVs for other languages, so as to share findings.

## 6 Corpus Statistics

General information about the corpus size is available in Table 2, whereas information about the VMWEs types and their frequency in the corpus is provided in Table 3, which shows that the majority (2 thirds) of the VMWEs in the corpus are reflexive verbs or adpositional ones. Such distribution of the types in the corpus should not be taken as general in the language, but should be interpreted with respect to the corpus texts genre, as well as their characteristics inherent to their source: being issues of the same daily newspaper, written by the same journalists, on more or less similar topics.

The most frequent (usually ten<sup>9</sup>) verbs in each type of VMWEs are enumerated below, and, between brackets, their frequency with the respective type of VMWEs; for verbs that are among the 20 most frequent ones in the corpus, we also indicate between brackets the relative frequency with which they are used in that type of VMWEs:

<sup>9</sup>We give less than 10 verbs when they have more than 1 occurrence.

Type	Number
IRV	3.826
LVC.cause	182
LVC.full	516
VID	1.644
IAV	3.311
<b>TOTAL</b>	<b>9479</b>

Table 3: Number of VMWEs of each type

- IRV: *desfășura* (unfold) (303, i.e. 47% of all its occurrences in the corpus), *afla* (find) (294, i.e. 42% of all its occurrences in the corpus), *adresa* (address) (203), *putea* (can) (190, i.e. 8% of all its occurrences in the corpus), *prezenta* (present) (117, i.e. 19% of all its occurrences in the corpus), *derula* (unreel) (93), *încheia* (finish) (91), *naște* (give birth) (87), *număra* (count) (63), *deplasa* (travel) (61);
- LVC.cause: *pune* (put) (179), *da* (give) (6);
- LVC.full: *avea* (have) (192, i.e. 7% of all its occurrences in the corpus), *face* (make, do) (173, i.e. 17% of all its occurrences in the corpus), *lua* (take) (108), *da* (give) (26), *aduce* (bring) (10), *pune* (put) (7);
- VID: *avea* (have) (804, i.e. 31% of all its occurrences in the corpus), *pune* (put) (108), *lua* (take) (102), *da* (give) (85), *fi* (be) (76, i.e. 9% of all its occurrences in the corpus), *intra* (enter) (65), *ține* (hold) (51), *trimite* (send) (50), *face* (make, do) (43, i.e. 4% of all its occurrences in the corpus), *sta* (stay) (41);
- IAV: *beneficia* (benefit) (185), *participa* (participate) (149), *desfășura* (unfold) (130, i.e. 20% of all its occurrences in the corpus), *intra* (enter) (120), *ajunge* (reach) (116), *pune* (put) (100), *trece* (pass) (98), *duce* (take to) (81), *lua* (take) (63), *ridica* (lift) (59).

We notice that verbs may tend to occur in one type of VMWEs, but there are many exceptions, with the verb *pune* (put) occurring with LVC.cause, LVC.full, VID and IAV expressions, and the verb *lua* (take) occurring with three types: LVC.full, VID and IAV. There are others occurring only with LVC.full and VID: *avea* (have), *face* (make, do), *da* (give). All are verbs with rich polisemy, sometimes even bleached in frozen combinations.

## 7 Conclusions

The new version of the PARSEME-Ro corpus comes with a new type of VMWEs: IAV. Such expressions are widely spread in the corpus: they represent a third of the total number of VMWEs occurring therein. This makes them an important phenomenon to be made explicit in a corpus. A comparative analysis of the cases when the same combination verb + adposition is either annotated as an IAV or not will be carried out, coupled with grammatical and semantic characteristics of the context, to better understand what the specific contexts for IAVs are.

So far, only verbal IAVs have been annotated in PARSEME-Ro, while VMWEs IAVs (IAVMWEs) are left for further investigations. Prior to this, we consider that the status of IAVs needs to get clarified, as our analysis of such expressions has shown that the type could be further classified into two subtypes: IAV.functional and IAV.semi-lexical.

The new version of PARSEME-Ro will be made fully and freely available in the first annual release within PARSEME, scheduled for mid 2022, in a format that will be agreed upon within the community. It is also available on our website of language resources in Linked Data format<sup>10</sup> and can be queried using the SPARQL endpoint<sup>11</sup>.

## Acknowledgments

Part of the work reported here has been carried out within Action 2020-EU-IA-0088 which has received funding from the European Union's Connecting Europe Facility 2014-2020 - CEF Telecom, under Grant Agreement No. INEA/CEF/ICT/A2020/2278547.

## References

- Timothy Baldwin and Su Nam Kim. 2010. *Handbook of Natural Language Processing, 2nd edition*, chapter Multiword Expressions. CRC Press, Boca Raton, FL, USA.
- Verginica Barbu Mititelu. 2018. Modern Syntactic Analysis of Romanian. In Daniela Butnaru Marius-Radu Clim Veronica Olariu Ofelia Ichim, Luminița Botoșineanu, editor, *Clasic și modern în cercetarea filologică românească actuală*, pages 67–78. Publishing House of "Alexandru Ioan Cuza" University.
- <sup>10</sup><https://www.racai.ro/p/llod/index.html>
- <sup>11</sup><https://relate.racai.ro/datasets/dataset.html>



- Verginica Barbu Mititelu, Mihaela Cristescu, and Manuela Nevaci. 2021. Un instrument modern de studiu al dialectului aromân: corpus adnotat morfosintactic. In Ioan-Mircea Farcaș Manuela Nevaci, Irina Floarea, editor, *Ex Oriente lux. In honorem Nicolae Saramandu*, pages 143–162. Edizioni dell’Orso, Alessandria.
- Verginica Barbu Mititelu and Maria Mitrofan. 2020. The Romanian Medical Treebank - SiMoNERo. In *Proceedings of the 15th International Conference “Linguistic Resources and Tools for Natural Language Processing”*, pages 7–16.
- Mihaela Colhon, Cătălina Mărănduc, and Cătălin Mititelu. 2017. Multiform Balanced Dependency Treebank for Romanian. In *Proceedings of Knowledge Resources for the Socio-Economic Sciences and Humanities, (KnowRSH)*, pages 9–18, Varna, Bulgaria.
- Norbert Corver and Henk van Riemsdijk, editors. 2013. *Semi-lexical Categories: The Function of Content Words and the Content of Function Words*. De Gruyter Mouton.
- Cecilia Căpățână. 2007. *Elemente de frazeologie*. Editura Universitaria Craiova.
- Ionuț Geană. 2013. *Construcții verbale prepoziționale în limba română*. Editura Universității din București.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. *Universal Dependencies*. *Computational Linguistics*, 47(2):255–308.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. *Universal Dependencies v1: a multilingual treebank collection*. In *Proc. of LREC*, pages 1659–1666, Portorož, Slovenia.
- V. Păis, M. Mitrofan, V. Gasan, C.L. and Coneschi, and A. Ianov. 2021. Named Entity Recognition in the Romanian Legal Domain. In *Proceedings of the Natural Legal Language Processing Workshop*, pages 9–18.
- Vasile Păis, Radu Ion, and Dan Tufiș. 2020. A processing platform relating data and tools for romanian language. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 81–88.
- Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archana Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoia Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 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, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archana Bhatia, Uxoia Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. *Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions*. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.
- Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. *Multiword expressions: A pain in the neck for NLP*. In *Proceedings of CICLing 2002*, pages 1–15.
- Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze. 2018. *PARSEME multilingual corpus of verbal multiword expressions*. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press, Berlin.
- Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. *The PARSEME shared task on automatic identification of verbal multiword expressions*. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.
- Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørðal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Matthieu Constant, Petya Osenova, and Federico Sangati. 2015. *PARSEME – PARSing and Multiword Expressions within a European multilingual network*. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland.
- Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. 2012. *Brat: a web-based tool for nlp-assisted text*

annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UD-Pipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, page 4290–4297, Portorož, Slovenia.



# A Parallel English - Serbian - Bulgarian - Macedonian Lexicon of Named Entities

Aleksandar Petrovski

Faculty of Informatics

International Slavic University

Marshal Tito 77 Sv. Nikole, North Macedonia

a.petrovski.sise@gmail.com, aleksandar.petrovski@msu.edu.mk

## Abstract

This paper describes the creation of a parallel multilingual lexicon of named entities from English to three South Slavic languages: Serbian, Bulgarian and Macedonian, with Wikipedia as a source. The basics of the proposed methodology are well known. This methodology provides a cheap opportunity to build multilingual lexicons, without having expertise in target languages.

Wikipedia's database dump can be freely downloaded in SQL and XML formats. The method presented here has been used to build a Python application that extracts the English – Serbian – Bulgarian – Macedonian parallel titles from Wikipedia and classifies them using the English Wikipedia category system. The extracted named entity sets have been classified into five classes: PERSON, ORGANIZATION, LOCATION, PRODUCT, and MISC (miscellaneous). It has been achieved using Wikipedia metadata. The quality of classification has been checked manually on 1,000 randomly chosen named entities. The following are the results obtained: 97% for precision and 90% for recall.

**Keywords:** parallel lexicons, named entities, Wikipedia

## 1 Introduction

Wikipedia is a free online encyclopedia, made and maintained as an open coordinated effort venture by a network of volunteer editors, utilizing a wiki – based editing system. Hosted and supported by the Wikimedia Foundation, since its start in 2001, the site has grown in both popularity and size. At the time of writing this paper (March 2022), Wikipedia contained over 58 million articles in 323 languages; its English version has over 6 million articles. The

richness of information and texts continuously makes it an object of special research interest among the NLP (Natural Language Processing) community. By attracting approximately 6 billion visitors per month (Statista, 2021), it is the largest and most popular general reference work on the World Wide Web.

The term named entity (NE) refers to expressions describing objects, like persons, locations, and organizations. It was first introduced to the NLP community at the end of the 20th century. Named entities are often denoted by proper names. They can be abstract or have a physical existence. Some other expressions, describing money, percentage, time, and date might also be considered as named entities. Examples of named entities include *United States*, *Paris*, *Google*, *Mercedes Benz*, *Microsoft Windows*, or anything else that can be named.

The role of named entities has become more and more important in NLP. Their information is crucial in information extraction. As recent systems mostly rely on machine learning techniques, their performance is based on the size and quality of given training data. This data is expensive and cumbersome to create because experts usually annotate corpora manually to achieve high-quality data. As a result, these data sets often lack coverage, are not up to date, and are not available in many languages. To overcome this problem, semi – automatic methods for resource construction from other available sources were deployed.

Even though Wikipedia isn't made and maintained by linguists, metadata about articles, for instance, translations, disambiguations, or categorizations are accessible. Its structural features, size, and multilingual availability give a reasonable base to derive specialized resources, like multilingual lexicons (Bøhn and Nørvag,

2010). Researchers have found that around 74% of Wikipedia pages describe named entities (Nothman et al., 2008), a clear indication of Wikipedia’s high coverage for named entities. Each Wikipedia article associated with a named entity is identified with its title, which is itself a named entity. That is a perfect opportunity to build parallel lexicons of named entities between them.

## 2 Related work

Building multilingual lexicons from Wikipedia has been a subject of research for more than 10 years. Schönhofen et al. (Schönhofen et al., 2007) exploited Wikipedia hyperlinkage for query term disambiguation. Tyers and Pienaar (Tyers and Pienaar, 2008) described a simple, fast, and computationally inexpensive method for extracting bilingual dictionary entries from Wikipedia (using the interwiki link system) and assessed the performance of this method with respect to four language pairs. Yu and Tsujii (Yu and Tsujii, 2009) proposed a method using the interlanguage link in Wikipedia to build an English-Chinese lexicon. Knopp (Knopp, 2010) showed how to use the Wikipedia category system to classify named entities. Bøhn and Nørvåg (Bøhn and Nørvåg, 2010) described how to use Wikipedia contents to automatically generate a lexicon of named entities and synonyms that are all referring to the same entity. Halek et al. (Hálek et al., 2011) attempted to improve machine translation from English of named entities by using Wikipedia. In (Ivanova, 2012), the author evaluated a bilingual bidirectional English-Russian dictionary created from Wikipedia article titles. Higashinaka et al. (Higashinaka et al., 2012) aimed to create a lexicon of 200 extended named entity (ENE) types, which could enable fine-grained information extraction. Oussalah and Mohamed (Oussalah and Mohamed, 2014) demonstrated how to use info-boxes in order to identify and extract named entities from Wikipedia.

## 3 English, Serbian, Bulgarian, and Macedonian Wikipedias

The English Wikipedia is the English language edition of the Wikipedia online encyclopedia. English is the first language in which Wikipedia

was written. It was started on 15 January 2001 (Wikimedia Foundation, 2001b), but versions of Wikipedia in other languages were quickly developed. There are three Wikipedias in concerned South Slavic languages among these versions: Serbian, Bulgarian, and Macedonian. They are all written in the Cyrillic alphabet, although there are few articles in Serbian Wikipedia written in Latin. The Serbian Wikipedia (Wikimedia Foundation, 2003c) was initiated in February 2003, the Macedonian (Wikimedia Foundation, 2003b) in September 2003, and the Bulgarian (Wikimedia Foundation, 2003a) in December 2003.

A list of all Wikipedias is published regularly on the Internet, along with several parameters for each language (Wikimedia Foundation, 2001a). Four parameters are considered: number of articles, the total number of pages (articles, user pages, images, talk pages, project pages, categories, and templates), number of active users (registered users who performed at least one change in the last thirty days), and depth (a rough indicator of the collaborative quality of Wikipedia, which shows how often articles are updated).

As shown in Table 1, as of 01 April 2022, the English Wikipedia contains 6,476,873 articles. It is by far the largest edition on Wikipedia. The Serbian Wikipedia contains 657,062 articles, the Bulgarian 280,535, and the Macedonian 126,265. According to the number of articles, the Serbian Wikipedia is the 21st largest edition of Wikipedia, the Bulgarian is 39th, and the Macedonian is 65th. The low value of the depth parameter for the Bulgarian Wikipedia is noticeable. It does not refer to low academic quality, which cannot be computed, but to Wikipedia quality, i.e. the depth of collaborativeness.

## 4 Method

Wikipedia’s database dump can be freely downloaded in SQL and XML formats (Wikimedia Foundation, 2001c). The method presented here has been used to build a Python application that extracts the English – Serbian – Bulgarian – Macedonian parallel titles from Wikipedia and classifies them using the English Wikipedia category system.

The flowchart presented in Figure 1 shows

Parameter	en	sr	bg	mk
Number of articles	6,476,873	657,062	280,535	126,265
Total number of pages	55,506,698	3,959,090	625,235	516,913
Number of active users	127,578	868	756	258
Depth	1,111	156	27	88

Table 1: Parameters of the English, Serbian, Bulgarian, and Macedonian Wikipedias.

the process used for building the lexicon.

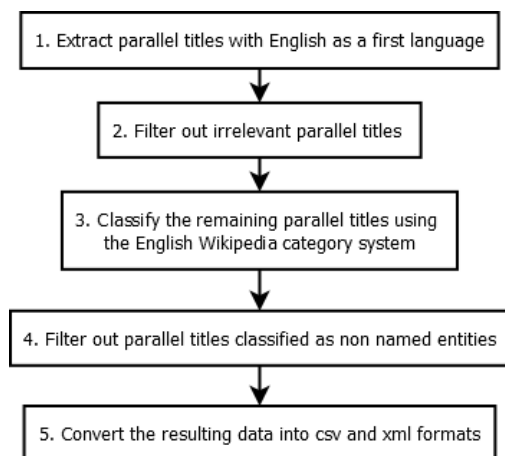


Figure 1: The process flowchart

1. *Extract parallel titles with English as a first language* – For building multilingual lexicons, two tables from the database are necessary: table of pages and table of interlanguage links. The page table is the "core of the wiki". It contains titles and other essential metadata for different Wikipedia namespaces. The interlanguage links table contains links between pages in different languages. Using these two tables, it is an easy programming task to create huge multilingual dictionaries without having any language expertise. In case of building multilingual lexicons with more than one language besides English, a new entry is created if there is a match between English and at least one of the other languages.

2. *Filter out irrelevant parallel titles* – The extracted parallel titles from the previous step contain a lot of noise. This step solves this issue. First, the program removes all the titles that don't belong to the main, template, or category namespaces. Second, there are titles containing some words or word stems that increase the noise and should be filtered out. The page table contains many entries that could not be a part of any lexicon, like usernames, nicknames, template names, etc. There are also titles,

containing exclusively digits or blanks, which should be removed, too.

3. *Classify the remaining parallel titles using the English Wikipedia category system* – To classify the extracted named entities, one additional table from the database is required: a table of category links. The task of classifying named entities using category links is more complex. Wikipedia articles are generally members of categories. A category may have subcategories, each subcategory its subcategories, etc. The problem is that the graph could be cyclic, which may cause the program to go into an endless loop. Various authors propose different classes for named entities. Here, there are five classes: PERSON, ORGANIZATION, LOCATION, PRODUCT, and MISC (MISCELLANEOUS). Each named entity belongs to at least one of these classes. The classes comprise:

ORGANIZATION – political organizations, companies, schools, rock bands, sports teams;

PERSON – humans, gods, saints, fictional characters;

LOCATION – geographical terms, fictional places, cosmic terms;

PRODUCT – industrial products, software products, weapons, artworks, documents, concepts, standards, laws, formats, anthems, algorithms, journals, coats of arms, platforms, websites;

MISC – events, languages, peoples, tribes, alliances, orders, scientific discoveries, theories, titles, currencies, holidays, dynasties, positions, projects, historical periods, battles, competitions, alliances, deceases, programs, set of locations, awards, musical genres, missions, artistic directions, sets of organizations, networks.

4. *Filter out parallel titles classified as non-named entities* – Most Wikipedia titles are named entities, but not all of them. For example, certain natural terms – like biological species and substances which are very common

on Wikipedia are not included in the lexicon.

5. *Convert the resulting data into CSV and XML formats* – The lexicon comes in two formats: CSV and XML. An example of a lexicon entry in XML format is shown in Figure 2. The first four element names of each entry are *en*, *sr*, *bg*, and *mk* for English, Serbian, Bulgarian, and Macedonian, respectively. The text content of the elements is a translation of *Sofia* in respected languages. The fifth element contains the class, or classes, the entry belongs to. In this case, it is a LOCATION.

```
<entry>
  <en>Sofia</en>
  <sr>Софија</sr>
  <bg>София</bg>
  <mk>Софија</mk>
  <classes>
    <class>LOCATION</class>
  </classes>
</entry>
```

Figure 2: A lexicon entry in XML format

## 5 The lexicon

### 5.1 Statistics

The method presented in previous chapter has been used to build a Python application which extracts title sets independently on the languages. This program was applied to the Wikipedia database to extract the English - Serbian - Bulgarian - Macedonian sets of named entities. The result of the extraction after the first two steps from Figure 1 was 586,355 entries for English, 374,691 for Serbian, 258,940 for Bulgarian, and 149,633 for Macedonian. There are few titles in all South Slavic Wikipedias that are written in the original language (mostly English). In addition to that, there are few titles in Serbian Wikipedia in the Latin alphabet. The transliteration from Latin to Cyrillic and vice versa in Serbian is relatively straightforward.

Table 2 shows the number of entries per language after filtering out non named entities. The number of named entities in English is equal to the number of entries in the lexicon. The entries' numbers in Serbian, Bulgarian,

Language	Number of titles
English	400,930
Serbian	257,542
Bulgarian	179,854
Macedonian	106,351

Table 2: Number of titles per language

Class	Number
PERSON	81,724
ORGANIZATION	23,127
LOCATION	161,524
PRODUCT	32,951
MISC	107,973
All	407,299

Table 3: Distribution of classes

and Macedonian are lower, which is understandable taking into account the number of articles in these Wikipedias, given in Table 1.

### 5.2 Distribution of classes

The resulting parallel English – Serbian – Bulgarian – Macedonian lexicon consists of 400,930 named entities. Each one belongs to at least one class, some of them to more. The distribution of classes is presented in Table 3.

The total number of classes, 407,299 is slightly higher than the number of entries, since some named entities may belong to more classes. The lexical entry presented in Figure 3 is such an example. *Bulgarian Academy of Sciences* is classified as both ORGANIZATION (the academy as an organization) and LOCATION (the buildings where the organization is located).

The examples of lexicon entries presented in figures 2 and 3 contain titles in all languages considered. But, this is not always the case. For example, as it is presented in Figure 4, the English title *Mark Antony* has translations only in Serbian and Bulgarian. There is no Macedonian translation since there is not such an article in the Macedonian Wikipedia.

### 5.3 Evaluation of classification

To evaluate classification, two common metrics in information retrieval have been used: precision and recall. Precision refers to the percentage of classes that are correct. On the other hand, recall refers to the percentage of

```

<entry>
  <en>Bulgarian Academy of Sciences</en>
  <sr>Бугарска академија наука</sr>
  <bg>Българска академия на науките</bg>
  <mk>Бугарска академија на науките</mk>
  <classes>
    <class>ORGANIZATION</class>
    <class>LOCATION</class>
  </classes>
</entry>

```

Figure 3: A lexicon entry belonging to two classes

```

<entry>
  <en>Mark Antony</en>
  <sr>Марко Антоније</sr>
  <bg>Марк Антоний</bg>
  <mk></mk>
  <classes>
    <class>PERSON</class>
  </classes>
</entry>

```

Figure 4: A lexicon entry with a missing translation in Macedonian

total relevant classes correctly classified by the algorithm.

An alternative to having two measures is the F – measure, which combines precision and recall into a single performance measure. This metric is known as F1 – score, which is simply the harmonic mean of precision and recall.

In order to evaluate the classification, a random sample containing 1,000 entries has been extracted from the lexicon. The entries from the sample have been classified manually and then compared to the classification performed by the algorithm. The results are presented in Table 4.

The precision of classification is between 94% for ORGANIZATION and 99% for PERSON. The recall is slightly lower, from 83% for PRODUCT and MISC to 97% for PERSON. The overall results are 97% for precision and 90% for recall.

The higher values of precision show that the classification algorithm was adjusted to classify the named entities correctly, rather than to

extract more named entities for the lexicon.

## 6 Utilization

Lexicons, like the one presented in this paper, can be used in machine translation (MT). Most statistical MT systems do not deal explicitly with named entities, simply relying on the model of selecting the correct translation, i.e., mistranslating them as generic nouns. It is also possible that, when not identified, named entities may be left out of the output translation, which also has implications for the readability of the text. Because most NEs are rare in texts, statistical MT systems are not capable of producing quality translations for them. Another problem with MT systems is that failure to recognize NEs often harms the morpho – syntactic and lexical context outside of NEs itself. If named entities are not immediately identified, certain morphological features of adjacent and syntactically related words, as well as word order, may be incorrect. However, developers of commercial MT systems often do not pay enough attention to the correct automatic identification of certain types of NE, e.g. names of organizations. This is partly due to the greater complexity of this problem (the set of proper nouns is open and very dynamic), and partly due to lack of time and other development resources. One solution to this problem is using a parallel lexicon of named entities. If the lexicon contains a translation of the named entity, the translation quality will probably be good.

## 7 Conclusion

Using the methodology presented in this paper, a multilingual lexicon of named entities



Class	Precision	Recall	F1-score
PERSON	99%	97%	98%
ORGANIZATION	94%	87%	90%
LOCATION	98%	92%	95%
PRODUCT	96%	83%	89%
MISC	96%	83%	89%
All	97%	90%	93%

Table 4: The results of the classification check

from English to Serbian, Bulgarian, and Macedonian has been created. The named entities have been classified into five classes: PERSON, ORGANIZATION, LOCATION, PRODUCT, and MISC (miscellaneous). The number of lexical entries for these South Slavic languages varies and is dependent on the size of their Wikipedias, from 106,351 for Macedonian to 257,542 for Serbian. The quality of classification has been assessed: 97% for precision, and 90% for recall.

## References

- Christian Bøhn and Kjetil Nørvag. 2010. Extracting named entities and synonyms from wikipedia. *Proceedings of International Conference on Advanced Information Networking and Applications*, pages 1300–1307.
- Ryuichiro Higashinaka, Kugatsu Sadamitsu, Kuniko Saito, Toshiro Makino, and Yoshihiro Matsuo. 2012. Creating an extended named entity dictionary from wikipedia. *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, pages 1163–1178.
- Ondrej Hálek, Rudolf Rosa, Aleš Tamchyna, and Ondrej Bojar. 2011. Named entities from wikipedia for machine translation. In *Proceedings of the Conference on Theory and Practice of Information Technologies*, pages 23–30.
- Angelina Ivanova. 2012. Evaluation of a bilingual dictionary extracted from wikipedia. In *Computer Science*.
- Johannes Knopp. 2010. *Classification of Named Entities in a Large Multilingual Resource Using the Wikipedia Category System*. University of Heidelberg, Master’s thesis, Heidelberg, Germany.
- Joel Nothman, James Curran, and Tara Murphy. 2008. Transforming wikipedia into named entity training data. *Proceedings of the Australian Language Technology Workshop*.
- Mourad Oussalah and Muhidin Mohamed. 2014. Identifying and extracting named entities from wikipedia database using entity infoboxes. *International Journal of Advanced Computer Science and Applications*, 5:164–169.
- Péter Schönhofen, András Benczúr, Istvan Biro, and Károly Csalogány. 2007. Cross-language retrieval with wikipedia. *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007ed Papers*, 5152:72–79.
- Statista. 2021. Worldwide visits to wikipedia.org from january to june 2021. <https://www.statista.com/statistics/1259907/wikipedia-website-traffic/>, Last accessed on 2022-03-31.
- Francis M. Tyers and Jacques A. Pienaar. 2008. Extracting bilingual word pairs from wikipedia. *Proceedings of the SALT MIL Workshop at the Language Resources and Evaluation Conference, LREC2008*.
- Wikimedia Foundation. 2001a. List of wikipedias. [https://meta.wikimedia.org/wiki/List\\_of\\_Wikipedias](https://meta.wikimedia.org/wiki/List_of_Wikipedias), Last accessed on 2022-03-31.
- Wikimedia Foundation. 2001b. Wikipedia, the free encyclopedia. [https://en.wikipedia.org/wiki/Main\\_Page](https://en.wikipedia.org/wiki/Main_Page), Last accessed on 2022-03-31.
- Wikimedia Foundation. 2001c. Wikipedia:database download. [https://en.wikipedia.org/wiki/Wikipedia:Database\\_download](https://en.wikipedia.org/wiki/Wikipedia:Database_download), Last accessed on 2022-03-31.
- Wikimedia Foundation. 2003a. Bulgarian wikipedia. <https://bg.wikipedia.org/wiki>, Last accessed on 2022-03-31.
- Wikimedia Foundation. 2003b. Macedonian wikipedia. <https://mk.wikipedia.org/wiki>, Last accessed on 2022-03-31.
- Wikimedia Foundation. 2003c. Serbian wikipedia. <https://sr.wikipedia.org/wiki>, Last accessed on 2022-03-31.
- Kun Yu and Jun’ichi Tsujii. 2009. Bilingual dictionary extraction from wikipedia. *Machine Translation Summit*, 12.



# Evaluation of Off-the-Shelf Language Identification Tools on Bulgarian Social Media Posts

Silvia Gargova, Irina Temnikova, Ivo Dzumerov, Hristiana Nikolaeva

Big Data for Smart Society Institute (GATE), Sofia, Bulgaria

svgargova@gmail.com, irina.temnikova@gate-ai.eu,

i.dzumerov@gmail.com, hnikolaeva@gmail.com

## Abstract

Automatic Language Identification (LI) is a widely addressed task, but not all users (for example linguists) have the means or interest to develop their own tool or to train the existing ones with their own data. There are several off-the-shelf LI tools, but for some languages, it is unclear which tool is the best for specific types of text. This article presents a comparison of the performance of several off-the-shelf language identification tools on Bulgarian social media data. The LI tools are tested on a multilingual Twitter dataset (composed of 2966 tweets) and an existing Bulgarian Twitter dataset on the topic of fake content detection of 3350 tweets. The article presents the manual annotation procedure of the first dataset, a discussion of the decisions of the two annotators, and the results from testing the 7 off-the-shelf LI tools on both datasets. Our findings show that the tool, which is the easiest for users with no programming skills, achieves the highest F1-Score on Bulgarian social media data, while other tools have very useful functionalities for Bulgarian social media texts.

**Keywords:** language identification, social media, evaluation, off-the-shelf tools, Bulgarian.

## 1 Introduction

Automatic Language Identification (LI) is a well-addressed task, with many existing approaches, tools, and evaluation initiatives (Jauhiainen et al., 2019; Garg et al., 2014). LI solves the problem of those users, who need to detect the language of a large number of texts, and thus cannot perform this task manually, as it will take them a large amount of time and manual efforts. Such users (for example linguists), do not have the knowledge, skills, or interest to develop their own LI tool or to train existing tools with their own data, and thus prefer using an existing off-the-shelf LI tool. As a first step, they are naturally interested to know which

is the best tool for the specific language (e.g. Bulgarian, Romanian, Hindi) and type(s) of text of their interest (e.g. *news articles* or *social media posts*). However, there is no sufficient information about which off-the-shelf LI tools are the best for all specific language/type-of-text combinations. For this reason, we are sharing our findings of the best off-the-shelf LI tools and their functionalities for the specific language and type of text of our interest. By doing this we aim to assist other users or researchers, who need to use such tools for their language identification tasks.

Our **language of interest is Bulgarian**, and the **the type of text - social media posts**, and in this article we are reporting the results of comparing several off-the-shelf LI tools on Bulgarian social media data.

Our work is motivated by the wish to solve the issue of filtering out any non-Bulgarian tweets from social media corpora. Following our task to collect and pre-process Bulgarian social media datasets for detecting fake content, our first observation was that despite using the Twitter API for collecting only posts in Bulgarian, our dataset contained many tweets (see Table 1 for precise numbers) in languages similar to Bulgarian or written in Cyrillic alphabet (for example Macedonian, Serbian, Russian, Kazakh, etc.). We have observed a similar issue when using other dataset collecting methods, such as Facebook's CrowdTangle. Determining the best LI tool for filtering out non-Bulgarian posts was thus a must.

To be able to identify the most appropriate LI tool and motivate our choices, we have to first understand and describe the characteristics of the language (Bulgarian) and type of text (social media posts) of our interest.

The Bulgarian language is part of the South Slavic languages' group within the Indo-European language family. In lexical, phonetic and grammati-

cal terms, Bulgarian has both Slavic and non-Slavic features. It is officially written in Cyrillic alphabet, but in social media and Internet forums people often use several variants of Latin transcription. Bulgarian is the official language of the Republic of Bulgaria. It has a literary form, used in all spheres of public life, and a number of local dialects, some of which are similar to the languages of North Macedonia and Serbia.

Social media texts (including those in Bulgarian) are known for being different from standard texts by being much shorter (e.g. a tweet can contain a maximum of 280 characters), frequently containing orthographic errors, Internet slang, non-dictionary words, emoticons, hashtags, unfinished sentences, and broken or non-standard syntax, and thus being challenging for many Natural Language Processing applications (Farzindar and Inkpen, 2017). In addition to that, social media posts may sometimes contain words and phrases, written in different languages – a phenomenon, known as *code-switching* (Androutopoulos, 2013).

A LI tool, which would be perfect for recognizing Bulgarian social media posts, thus, should:

1. Have the highest possible performance (e.g. an over 98% F1-score);
2. Be able to recognize Bulgarian texts, written both in Cyrillic alphabet and in the various Latin transcriptions (typical for the Bulgarian Internet slang);
3. Be able to handle the above described social media posts' characteristics, including the cases when the post is written in two or more languages.

In order to discover the most appropriate LI tool for correctly identifying the Bulgarian language posts in social media data, we have determined the most frequently used Off-the-Shelf LI tools (OSLI), by examining publications and consulting other researchers. We have then tested them on two datasets - a multilingual (mostly Bulgarian) dataset, collected from Twitter on the topic of Covid-19 with 2979 tweets, manually annotated for language(s), and a Bulgarian language dataset (Shaar et al., 2021), used for fake content detection initiatives, consisting of 3350 tweets.

The article provides the results of the human annotation and of testing the tools, as well as shows which tools achieve the highest F1-scores on the

two datasets, and which have the most useful functionalities for social media posts.

The rest of the article lists the relevant Related work (Section 2), a description of the datasets that we used for testing the tools (Section 3), our Methodology (including human annotation and the tested tools - in Section 4), the annotation and testing Results and some Discussion (Section 5), and finally, the Conclusions (Section 6).

## 2 Related Work

Automatic Language Identification (LI) is a widely addressed task, but it still has some issues which are hard to resolve. Among them (Jauhiainen et al., 2019) are:

- Distinguishing between similar languages or dialects;
- Short and noisy texts;
- Documents, written in more than one language;
- Languages with different orthographies.

All these issues apply to Bulgarian social media posts.

There have been a number of previous works which include Bulgarian among other languages in their LI tasks or datasets, for example (Zampieri et al., 2015; Jauhiainen et al., 2017; Malmasi, 2017; Bergsma et al., 2012; Baldwin and Lui, 2010; Thoma, 2018). Most of them, however, use datasets compiled from types of texts, which are different from social media (e.g. Wikipedia, news articles, Europarl, and the Universal Declaration of Human Rights). Also, most of these works do not apply existing off-the-shelf LI tools to detect Bulgarian, but rather implement their own methods.

The closest works to ours are those of (Abainia et al., 2016), (Bergsma et al., 2012), (Bankov et al., 2017), and (Lui and Baldwin, 2014). Among them, however, there is no work which compares the performance of different off-the-shelf LI (OSLI) tools on Bulgarian social media posts and publishes the results.

Specifically, (Abainia et al., 2016) are similar to us as they use short forum texts, including such written in Bulgarian, but no testing of OS LI tools is performed. (Bergsma et al., 2012) compare LI methods implemented by them with three off-the-shelf tools (TextCat, Google CLD and langID.py)

on a multilingual Twitter dataset containing also Bulgarian. Their methods outperform the OS LI tools, but there are no results reported separately for Bulgarian. (Bankov et al., 2017) also observes that Twitter’s accuracy for Bulgarian language identification is not satisfactory, however, the author does not test any OS LI tools on Bulgarian tweets.

Finally, there are publications on testing various off-the-shelf LI tools on specific languages, but not on Bulgarian. For example, (Lui and Baldwin, 2014) compared 8 OS LI tools on manually annotated tweets in English, Chinese, and Japanese.

While several OS LI tools include Bulgarian, according to our knowledge, there is no other published comparison of off-the-shelf LI tools for this language, especially for social media texts.

### 3 Data Used

We have used two datasets - a randomly selected subset of our own Twitter dataset, and the Bulgarian language dataset, made available for the CLEF2021 CheckThat! Lab, Task 1 (check-worthiness). From now on we refer to this dataset as *CLEF2021 dataset*<sup>1</sup> (Shaar et al., 2021). The large original version of our dataset contains 52810 tweets, from which we extracted 3124 tweets, which were annotated for their language by human annotators. We have removed some non immediately noticeable duplicates and did some additional cleaning (based on our annotators feedback), and obtained 2966 final human-annotated tweets, on which we tested the LI tools. Respectively, the CLEF2021 CheckThat! Lab, Task 1 dataset for Bulgarian originally contained 3350 entries (tweets).

We have decided to compare the results of the same off-the-shelf LI tools on the subset of our dataset with those on the CLEF2021 Bulgarian dataset, as they both had comparable number of tweets and are on the same topic (Covid-19).

Before testing the LI tools on the CLEF2021 CheckThat! Lab, Task 1 Bulgarian dataset, we have merged the Bulgarian versions of its *train* and *dev* datasets into one to have more data. After a quick analysis of the merged CLEF2021 dataset, we noticed two issues: unusually long entries (consisting in many tweets concatenated in one row) and a few tweets in other languages. We separated the long rows into single posts and removed the

<sup>1</sup>[https://gitlab.com/checkthat\\_lab/clef2021-checkthat-lab/-/tree/master/task1/data/subtask-1a-bulgarian](https://gitlab.com/checkthat_lab/clef2021-checkthat-lab/-/tree/master/task1/data/subtask-1a-bulgarian). Last accessed on April 27, 2022.

Languages	Num. of tweets
Bulgarian	2491
Macedonian	248
Russian	214
English	43
Mongolian	38
Uzbek	22

Table 1: Number of tweets in the most frequent languages in our 3124 tweets Covid-19 dataset.

Stats	Covid-19	CLEF2021
Num. tweets	2966	3373
Total words	47628	66502
Mean tweet length	16.06	19.72
Shortest tweet	1	5
Longest tweet	54	108

Table 2: Statistics of the two datasets used for testing the tools.

tweets that are not in Bulgarian, which resulted in 3373 final tweets.

The tweets in our original large dataset<sup>2</sup> were collected via the Twitter API for the period May 2020 - March 2021. The keywords used were “ваксина” (“vaccine”, Sg.) and “ваксини” (“vaccines”, Pl.) in Bulgarian language and using the Cyrillic alphabet. From this large dataset we have selected a smaller random subset from different time intervals. Each tweet from this final dataset (from now on referred to as *Covid-19 dataset*) was manually annotated by two annotators for its language. The annotation methodology is described in detail in Section 4.1.

The first thing that we noticed during manual annotation, was that our dataset contained posts in other languages besides Bulgarian. Table 1 shows the most frequent languages in our dataset for the languages, in which there are more than 10 tweets. The length of the posts in our dataset was quite varied - we had one-word tweets and much longer tweets (differently from the CLEF2021 dataset, which contained only tweets long enough to be considered fact-checkable *claims*). Many of the posts were written in more than one language. There were also 2 posts that contained only emoji.

Table 2 shows the statistics of both datasets. “Covid-19” stands for our Covid-19 dataset, “Num. tweets” indicates the total number of tweets per

<sup>2</sup>This dataset cannot be shared due to specific access restrictions.

dataset, “Total words” - the total number of words in each dataset, “Mean tweet length” is the mean length of the tweet in words, while “Shortest tweet” and “Longest tweet” were the tweets containing the lowest and the highest number of words.

As it can be seen in Table 2, the CLEF2021 dataset contains longer tweets than ours. This can be of advantage to the LI tools.

## 4 Methodology

In order to test the existing off-the-shelf LI tools, we have performed manual language annotation of the 3124-tweets-subset of our dataset (the one containing duplicates), which is described in Subsection 4.1, and selected a number of freely available and functioning LI tools (described in Subsection 4.3). The methodology, which we followed for testing the tools on both datasets is described in Subsection 4.2.

### 4.1 Annotation methodology

The aim of the manual annotation of our subset tweets dataset was to focus on **distinguishing specifically Bulgarian**, rather than correctly annotating all the languages of all the tweets in our dataset. This was motivated first by our aim to find the best LI tool for Bulgarian, but also by the knowledge of languages of our annotators.

We had two annotators, who are professional linguists, native in and specializing in Bulgarian language. They used Google Spreadsheets as an annotation tool, due to its simplicity. The spreadsheet had three columns, containing the tweet ID, the text of the tweet and several language-related categories in a fall-down menu to choose from. Appendix A shows a screenshot of the spreadsheet containing mock examples of annotated tweets.

The annotators were asked to only decide if the tweets are written in Bulgarian (**bg**) OR in Another language (**another**), without distinguishing exactly in which other language. As we are planning in future work to examine the performance of some of the tools in distinguishing the different languages present in multilingual tweets, we also asked the annotators to comment on which tweets are multilingual and whether they contain Bulgarian language or not. As there were some unclear cases, we provided an additional category “Unknown”. The annotation categories are shown in Table 3.

We have considered multilingual also those tweets, which contained hashtags, written in an-

other language (e.g. in English). However, we have asked the annotators to ignore the keyword “Covid-19” (and its versions, e.g. Covid), written in English, as they were too frequent due to the topic of our dataset.

The two annotators received initial training, worked separately, and did several rounds of the annotation process until the annotations and the guidelines were finalized.

As after this process there were still cases in which the annotators disagreed, to facilitate the comparison with the LI tools, we have assigned a third **hyper-annotator**. The hyper-annotator reviewed the cases of disagreement of the two annotators and decided on a final annotation category for each tweet. The hyper-annotator was also a linguist, specialist in Bulgarian language. In order to take the correct final decisions, the hyper-annotator was allowed to have a look at the original tweet in Twitter and check information about the user who posted it, including his/her location and other tweets.

As testing the tools’ performance in identifying multiple languages within the same tweet is beyond the scope of this article, the tweets, annotated as “bg-multilingual” and “bg” categories were merged into “**bg**” and “another-multilingual” and “another” were merged into the category “**another**”. We have also removed the tweets, left annotated as “unknown” by the hyperannotator. This gave us a final number of 2966 manually annotated tweets.

See Section 5.1 for a discussion of the manual annotation results.

### 4.2 Testing Methodology

Our aim was to test only freely accessible LI tools (not paid ones).

During testing, we wanted to check the performance of the LI tools with the tweets as they are (we call these tweets “**raw**”) and with tweets, from which several Twitter-specific elements were removed (we call these tweets “**cleaned**”) and whether there was any change or improvement in performance if the data was cleaned in advance. Our hypothesis was that Twitter-specific elements (e.g. hashtags, URLs, and mentions) would hinder the performance of LI tools.

For this purpose we performed two experiments - one with *raw data* and another with *cleaned data*. For the first experiment we used our dataset as it is (we only deleted duplicates). For the second



Annot. category	Explanation
<b>bg</b>	You are sure that the tweet is written entirely in Bulgarian language.
<b>another</b>	You are sure that the tweet is not in Bulgarian, regardless of whether you know what other language it is written in.
<b>unknown</b>	You are not sure if the tweet is in Bulgarian or in another language, but you have at least a minimal suspicion that it may be written in Bulgarian.
<b>another-multilingual</b>	The tweet is bilingual or multilingual, but you are sure that none of the languages is Bulgarian.
<b>bg-multilingual</b>	You are sure that the tweet is written in Bulgarian + another language.

Table 3: Annotation categories with their explanations.

experiment we removed URLs, emojis, hashtags (both the # sign and the entire word) and mentions, then we checked again and deleted newly appeared duplicates, and only then performed the testing experiment. We repeated the same process with the dataset from CLEF 2021.

In our manually annotated dataset we have two annotated categories - “bg” (Bulgarian) and “another”. To calculate the accuracy, we first transformed our annotations into binary values. If the label is “bg” we assign 1, if the label is “another” we assign 0. Then we converted also the LI tools results into binary values. If the label is “Bulgarian” we assign 1, otherwise we assign 0. If the tool can detect more than one language we use/take only the first predicted label, or the label with the highest confidence score (usually the first one). Finally we use the binary values to make the calculation.

Unfortunately **spaCy** left some tweets without language labels. To compute its accuracy, we removed these tweets from both datasets.

In addition to the accuracy score, we also calculated precision, recall and F1-score. We obtained these scores for both datasets and their raw and cleaned versions.

### 4.3 Tested Language Identification Tools

All the tools that we tested support Bulgarian language and some of the other languages, written in Cyrillic alphabet, such as Russian, Macedonian, Ukrainian, etc. We have chosen these specific LI tools, because they are free (not paid), well known, and because some of them (e.g. Google Sheets’s DETECTLANGUAGE function) are easy to use. While we are targeting readers, who are not interested to train these tools on their own data, we are providing enough technical details also for more technically-oriented users.

Without a doubt, one of the most famous and

widely used tools is the **Google Translate API**. We found 2 libraries – **TextBlob** and **googletrans**. **TextBlob**<sup>3</sup> has a language detection function which uses Google Translate API, but currently they recommend to use instead the official API. The library **googletrans**<sup>4</sup> also implements Google Translate API. It uses the Google Translate Ajax API to make calls. The authors warn that this is an unofficial library and the maximum character limit on a single text is 15k. Also they cannot guarantee that the library will work properly at all times and recommend the use of the official API for more stability. When we first tested **googletrans**, it assigned language labels to part of the tweets. In the following tests it annotated all tweets with the tag “English”. Due to the above mentioned limitations and the paid access we decided not to test Google Translate API further. However we tested another application from Google, which is free and has a language detection function (the Google Sheets DETECTLANGUAGE<sup>5</sup>). We refer to it from now on as **Google Sheets**.

**fastText**<sup>6</sup> (Joulin et al., 2016a,b) is developed by Facebook AI Research. It is a library for text classification and representation, which transforms text into continuous vectors that can be later used on any language-related task. **fastText** recognizes 176 languages and has been trained on data from Wikipedia, Tatoeba and SETimes. There are two models – a full version which is faster and more accurate, and a compressed version. The new line breaks were an issue for this tool and we had to

<sup>3</sup><https://textblob.readthedocs.io/en/dev/index.html>. Last accessed on April 11, 2022.

<sup>4</sup><https://pypi.org/project/googletrans/>. Last accessed on April 11, 2022.

<sup>5</sup><https://support.google.com/docs/answer/3093278?hl=en>. Last accessed on March 10, 2022.

<sup>6</sup><https://fasttext.cc/docs/en/language-identification.html>. Last accessed on April 11, 2022.

remove them in order to use it.

The next tool is **CLD3**<sup>7</sup>. It is a neural network model for language identification which uses character n-grams and calculates the fraction of times each of them appears. CLD3 supports 107 languages. We discovered that this is the only tool (among all of those that we tested), that has the very useful functionality for social media texts to recognize Bulgarian language written in Latin alphabet.

**langdetect**<sup>8</sup> is a direct port of Google’s language-detection library from Java to Python. It supports 55 languages (including Bulgarian and other languages, written in Cyrillic alphabet). The original tool was trained on data from Wikipedia and tested on data from Google News or other news sites. The library *language-detection* uses Naive Bayes for classification. langdetect is fast and has good accuracy. This is the only tool that gave us an error when annotating a tweet which contains only emojis. The output is a list of the top languages that the model has predicted, along with their probabilities. When the probability of the prediction is less than 0.90, it usually adds more labels.

**LangID**<sup>9</sup> (Lui and Baldwin, 2012) is a fast language detection tool. It comes pre-trained on 97 languages and is not sensitive to domain-specific features (e.g HTML/XML markup). The model consists of a single .py file with minimal dependencies and can be deployed as a web service. The training data was collected from 5 different sources – JRC-Acquis, ClueWeb 09, Wikipedia, Reuters RCV2 and Debian i18n. Please, note that its confidence score is not normalised by default.

Another language detection tool is **polyglot**<sup>10</sup>. It depends on the *pycld2* library which in turn depends on the *cld2* library for detecting languages. This tool is suitable for mixed text messages. If the tweet contains phrases from different languages, the detector can find the most probable languages used in the text along with the confidence level. When there is not enough text to make a decision (e.g. a tweet containing only one word), the detector is forced to switch to the best effort strategy. Sometimes even using the best effort strategy, the

<sup>7</sup><https://github.com/google/cld3>. Accessed on April 11, 2022.

<sup>8</sup><https://pypi.org/project/langdetect/>. Last accessed on April 11, 2022.

<sup>9</sup><https://github.com/saffsd/langid.py>. Last accessed on April 10, 2022.

<sup>10</sup><https://polyglot.readthedocs.io/en/latest/index.html>. Last accessed on April 11, 2022.

detection is not reliable and an “Unknown Language” exception is thrown. In cases where the text contains characters that could belong to more than one language, this can be problematic. Polyglot can identify the languages supported by *cld2* (up to 165). One of the problems with this tool was that our dataset contained some amount of short tweets and it wasn’t very confident in its predictions.

The last tool we tested is **spaCy**<sup>11</sup>. It is a library for advanced Natural Language Processing. spaCy comes with pre-trained pipelines for over 60 languages, uses state-of-the-art speed and neural network models and a lot of features for language processing. It’s open-source and easy to deploy. SpaCy has 2 modules with language detection capabilities: *spaCy-langdetect* and *spaCy-cld*. We used *spaCy-cld* for our research. This tool provides the most probable languages (up to 3) for the text. When the tweets are multilingual, these one to three hypotheses sometimes correspond to the various languages, present in the tweet, however we haven’t tested its accuracy in predicting multiple languages within the same tweet. *spaCy-cld* also uses *pycld2* and *cld2*. As both spaCy and polyglot use the same library, the results they gave were very similar. During our tests, we observed something interesting: the tool has left some tweets not tagged.

## 5 Results and Discussion

### 5.1 Results from the Manual Annotation

The two annotators disagreed on 31 out of 3124 tweets, which equals to 99.2% agreement between the annotators. We have additionally obtained a Cohen kappa value of 0.9691 for the Inter-Annotator Agreement (IAA) between the two annotators. The review done by the hyper-annotator has shown that both annotators did a few mistakes (probably from getting tired). Other specific cases in which they disagreed included:

- Very short tweets, composed of words, that exist in several languages (e.g. in Bulgarian and Russian: “настроение...” or “Логично и.....технологично.”, “Лондон”). Translation in English: “mood...”, “Logically and.....technologically.”, “London”.
- Cases due to the lack of extensive knowledge of the annotators in terms of Bulgarian dialects or other close languages (we cannot

<sup>11</sup><https://spacy.io/>. Last accessed on April 11, 2022.



Tools	Raw data				Clean data			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
fastText	<b>0.96</b>	<b>0.99</b>	<b>0.97</b>	<b>0.98</b>	<b>0.97</b>	<b>0.99</b>	<b>0.98</b>	<b>0.98</b>
CLD3	0.93	0.96	0.96	0.96	0.94	0.97	0.97	0.97
langdetect	0.93	0.96	0.96	0.96	0.93	0.96	0.96	0.96
LangID	0.90	0.98	0.90	0.94	0.91	0.97	0.91	0.94
polyglot	0.90	<b>0.99</b>	0.89	0.94	0.91	<b>0.99</b>	0.90	0.94
Google Sheets	<b>0.99</b>	<b>0.99</b>	<b>0.99</b>	<b>0.99</b>	<b>0.99</b>	<b>0.99</b>	<b>1.00</b>	<b>0.99</b>
spaCy*	0.89	0.99	0.87	0.93	0.90	0.99	0.88	0.93

Table 4: Results of the tests performed on our dataset.

Tools	Raw data				Clean data			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
fastText	<b>0.98</b>	<b>1.00</b>	<b>0.98</b>	<b>0.99</b>	<b>0.99</b>	<b>1.00</b>	<b>0.99</b>	<b>0.99</b>
CLD3	<b>0.98</b>	<b>1.00</b>	<b>0.98</b>	<b>0.99</b>	<b>0.98</b>	<b>1.00</b>	0.98	<b>0.99</b>
langdetect	<b>0.98</b>	<b>1.00</b>	<b>0.98</b>	<b>0.99</b>	<b>0.98</b>	<b>1.00</b>	0.98	<b>0.99</b>
LangID	0.93	<b>1.00</b>	0.93	0.96	0.93	<b>1.00</b>	0.93	0.96
polyglot	0.91	<b>1.00</b>	0.91	0.96	0.93	<b>1.00</b>	0.93	0.96
Google Sheets	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
spaCy*	0.91	1.00	0.91	0.95	0.94	1.00	0.94	0.97

Table 5: Results of the tests performed on CLEF 2021 dataset.

share examples due to Twitter’s data sharing restrictions).

- Tweets in English, but transliterated in Cyrillic letters, e.g. “Толд йа соу” (“Told ya so”).

See the following Section 5.2 on how the LI tools dealt with such short and ambiguous tweets.

## 5.2 Results from Testing the LI Tools

The final results can be seen in Tables 4 and 5, where “F1” stands for F1-Score. Undoubtedly, the best performing language identification tool is Google Sheets, which was a surprise for us. The second best performing tool is fastText. However, it is difficult to make a ranking because each tool has its advantages and disadvantages.

One of the first problems that we noticed while executing the code is that fastText gives an error if the text of the tweet is not in one line. We had to remove all the new line symbols before using the tool. The other tools had no problem with that. The next tool that gave us an error was langdetect. We had to remove tweets that only contained emojis or replace the emojis with some text so that the tool can annotate the data. The other tools did not give emoji-caused errors during code execution, but some of them did not annotate such tweets

(spaCy), some labeled them as “unknown” or “undefined” (polyglot and Google Sheets), and some labeled them as if they were normal text (fastText, CLD3 and langID). Therefore, we removed from our dataset 2 tweets that contained only emojis.

Another problem that we encountered is that spaCy did not assign language labels to some of the tweets. We tried to understand why this was happening, but we couldn’t. For our dataset, the tool did not annotate 96 posts (raw data). The length of these tweets varied between 1 and 29 words (average word length - 6.49), most of the unannotated tweets were 6 words long. The number of cleaned unannotated tweets increased to 200, their length was 1-26 words (with an average length of 7.36). Again, most of the unannotated tweets were 1 word long. We checked if all the 1-word-long tweets were not annotated, but it turned out that some of 1-word-long tweets were annotated. For CLEF2021 datasets, the unannotated tweets were fewer - 32 (raw data) and 114 (clean data). Again, we observe an increase in the number of tweets not annotated by spaCy after cleaning the data. We hypothesize that this might be due to the fact that during “cleaning” whole words (hashtags and mentions) were removed. In the raw data, the length of the tweets varied between 5 and 26 words (average word length - 12.69), with most of the unannotated tweets being

9 words long. The length of the cleaned unannotated tweets in CLEF2021 was 4-42 words (with an average length of 11.61). The highest number of unannotated tweets had a length of 7 words (for raw data) and 12 words (for cleaned data). We looked at the text of the unannotated tweets of the raw datasets, but we could not find the reason (for example, they may have contained only hashtags or code-switching), but the texts were very diverse. As spaCy does not label all the data, its results are separated from the results of the other tools in Tables 4 and 5.

All tools, in addition to language, also provide data on accuracy or a confidence score. However, only 2 of the tools output more than one language label - langdetect and spaCy. It is not described in detail on what principle they put these labels, but we noticed that they usually put several labels if they have detected more than one language in the text or arrange the languages according to accuracy. In our dataset we had only one post in Bulgarian–Latin, which was labeled correctly by CLD3. CLD3 is also the tool that provides the most detailed output.

We tested with which languages the tools most often confuse tweets, written in Bulgarian. When making mistakes, the tools most frequently tag Bulgarian tweets as Macedonian (mk) (see Table 6 for the most common mistakes of the tools when tagging Bulgarian tweets). Some of the tools tag Bulgarian (bg) tweets as Russian (ru) or Serbian (sr). These errors may be due to the amount of data in these languages in the datasets used to train them. We assume that when training fastText, the largest amount of data was in Russian. Respectively, the largest amount of data for polyglot and spaCy was probably in Serbian.

Regarding the very short tweets, which the human annotators struggled with (see the end of Section 5.1), surprisingly, the LI tools correctly recognized the language, even if they had access only to the text of the tweet. As the investigation of the hyper-annotator showed that most of these tweets were written in Russian, our hypothesis was that the tools have been pre-trained on much larger amounts of Russian texts. Further investigation of this issue is necessary.

In terms of speed, all the tools did quite fast in labelling all datasets. FastText, CLD3 and polyglot annotated the tweets in less than 5 seconds, and langID annotated data in about 10 seconds. The rest

Tools	Covid-19		CLEF 2021	
	Raw	Clean	Raw	Clean
fastText	ru	ru	ru	ru
CLD3	mk	mk	sr	mk
langdetect	mk	mk	mk	mk
LangID	mk	mk	mk	mk
polyglot	sr	sr	sr	sr
Google Sheets	mk	mk	mk	sr
spaCy	sr	sr	sr	sr

Table 6: The most common mistakes of the LI tools when providing language labels to tweets, written in Bulgarian.

of the tools were slower, but the annotation time remains less than 1 minute. It takes spaCy about 40 seconds to annotate the data, and langdetect about 30.

## 6 Conclusions

In this article we have presented the results from comparing 7 well-known off-the-shelf Language Identification (LI) tools on identifying Bulgarian language posts in two Twitter datasets, composed of around 3000 tweets each. We provided a presentation of each tool along with its useful functionalities and eventual shortcomings. We are confident that this information will be of use to any researchers, who would like to know the performance of off-the-shelf LI tools on Bulgarian social media posts, without training them.

Our results show that the tool which has the highest scores is the **DETECTLANGUAGE()** functionality of Google Sheets. The second best is **fastText**. We have found out that CLD3 has also the functionality to recognize Bulgarian, written with Latin letters, which is useful for social media and Internet forums texts. Testing its performance for this task has still to be done. We have also discovered that **polyglot** and (partially) **spaCy** can be used to guess multiple languages, present within the same text, but their performance in executing this task needs to be properly tested too.

We haven't discovered any LI tool, which simultaneously has a high Accuracy/F1-Score, can recognize Bulgarian written with Latin letters, and recognizes the languages in multi-lingual posts. This presents an opportunity for creating such a tool.

As future work, we plan to evaluate in more detail the above mentioned functionalities of polyglot, spaCy, and CLD3, and also to implement our own

LI tool.

## 7 Acknowledgements

The work, presented in this article has been supported by the project GATE (funded by Operational Programme Science and Education for Smart Growth under Grant Agreement No. BG05M2OP001-1.003-0002- C01) and is part of the research project TRACES<sup>12</sup>, which has indirectly received funding from the European Union’s Horizon 2020 research and innovation action programme, via the AI4Media Open Call 1, issued and executed under the AI4Media project (Grant Agreement no. 951911).

## References

- Kheireddine Abainia, Siham Ouamour, and Halim Sayoud. 2016. Effective language identification of forum texts based on statistical approaches. *Information Processing & Management*, 52(4):491–512.
- Jannis Androutsopoulos. 2013. 27. Code-switching in computer-mediated communication. *Pragmatics of computer-mediated communication*, page 667.
- Timothy Baldwin and Marco Lui. 2010. Multilingual language identification: ALTW 2010 shared task data. In *Proceedings of the Australasian Language Technology Association Workshop 2010*, pages 4–7.
- Boris Bankov et al. 2017. Extracting top trends from Twitter discussions in Bulgarian. *Izvestia Journal of the Union of Scientists-Varna. Economic Sciences Series*, (2):254–259.
- Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific Twitter collections. In *Proceedings of the second workshop on language in social media*, pages 65–74.
- Atefeh Farzindar and Diana Inkpen. 2017. Natural language processing for social media. *Synthesis Lectures on Human Language Technologies*, 10(2):1–195.
- Archana Garg, Vishal Gupta, and Manish Jindal. 2014. A survey of language identification techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, 6(4):388–400.
- Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.
- Tommi Sakari Jauhiainen, Bo Krister Johan Linden, Heidi Annika Jauhiainen, et al. 2017. Evaluation of language identification methods using 285 languages. In *21st Nordic Conference of Computational Linguistics Proceedings of the Conference*. Linköping University Electronic Press.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Herve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.
- Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.
- Marco Lui and Timothy Baldwin. 2014. Accurate language identification of Twitter messages. In *Proceedings of the 5th workshop on language analysis for social media (LASM)*, pages 17–25.
- Shervin Malmasi. 2017. Open-set language identification. *arXiv preprint arXiv:1707.04817*.
- Shaden Shaar, Maram Hasanain, Bayan Hamdan, Zien Sheikh Ali, Fatima Haouari, Mucahid Kutlu Alex Nikolov, Firoj Alam Yavuz Selim Kartal, Giovanni Da San Martino, Alberto Barrón-Cedeño, Rubén Míguez, Tamer Elsayed, and Preslav Nakov. 2021. **Overview of the CLEF-2021 CheckThat! lab task 1 on check-worthiness estimation in tweets and political debates**. In *Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum, CLEF ’2021, Bucharest, Romania* (online).
- Martin Thoma. 2018. The WiLI benchmark dataset for written language identification. *arXiv preprint arXiv:1801.07779*.
- Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9.

## Appendix A Tool used for manual language annotation

As described in the article, we have used Google Spreadsheets for manually annotating the languages of social media posts. Figure 1 shows the annotation spreadsheet with fall-down menu, containing the annotation categories. The examples of tweets are mock ones, due to Twitter’s restrictions on sharing their data.

<sup>12</sup><https://traces.gate-ai.eu/>

tweet ID	tweet_text	annotation
https://twitter.com	456 от лицата, при които е потвърден COVID-19 у нас,	bg
https://twitter.com	Со звук!	another
https://twitter.com	Има луѓе уште седат дома #stayhome #covid	another-multiling
https://twitter.com	настроение...	unknown
		bg
		another
		unknown
		bg-multiling
		another-multiling

Figure 1: Annotation spreadsheet with fall-down menu.

# Language rehabilitation of people with BROCA aphasia using deep neural machine translation

Kamel Smaili

David Langlois

Peter Pribil

Loria - SMarT Université Lorraine

{kamel.smaili, david.langlois, peter.pribil}@loria.fr

## Abstract

More than 13 million people suffer a stroke each year. Aphasia is known as a language disorder usually caused by a stroke that damages a specific area of the brain that controls the expression and understanding of language. Aphasia is characterized by a disturbance of the linguistic code affecting encoding and/or decoding of the language. Our project aims to propose a method that helps a person suffering from aphasia to communicate better with those around him. For this, we will propose a machine translation capable of correcting aphasic errors and helping the patient to communicate more easily. To build such a system, we need a parallel corpus; to our knowledge, this corpus does not exist, especially for French. Therefore, the main challenge and the objective of this task is to build a parallel corpus composed of sentences with aphasic errors and their corresponding correction. We will show how we create a pseudo-aphasia corpus from real data, and then we will show the feasibility of our project to translate from aphasia data to natural language. The preliminary results show that the deep learning methods we used achieve correct translations corresponding to a BLEU of 38.6.

**Keywords:** Aphasia, Augmentation corpus, Machine translation, Deep learning.

## 1 Introduction

Stroke represents the 2<sup>nd</sup> cause of death in the population, and the 1<sup>st</sup> cause of physical handicap in France. According to World stroke organization, 13.7 million people worldwide will suffer their first stroke on 2022 and 5.5 million will die from it. The incidence of stroke increases significantly with age and, in the West, as people are living longer and longer, stroke is almost becoming a pandemic problem.

Stroke can affect the mechanisms of speech, movement, sensation, and so on. The physical, cognitive and psychic after-effects of a stroke remain,

unfortunately, frequent (30 to 50% of cases). Many survivors will experience some form of lifelong disability or impairment that they will attempt to cure. They are particularly confronted with the reduction of their movements and isolation, among other things, due to their inability to communicate normally. With rehabilitation and specialist support, however, most stroke survivors can return to a near-normal life.

According to the National Aphasia Association<sup>1</sup>, a third of strokes result in aphasia, a major after-effect that greatly affects quality of life (Summers et al., 2009). Aphasia, a term suggested by Armand Trousseau in 1863, is characterized by a disturbance of the linguistic code, affecting the expression and/or the comprehension, and which can concern the oral and/or the written language. It is a localized or diffuse brain damage, generally in the frontal, parietal and/or temporal area of the left hemisphere, essentially of vascular, traumatic or of tumor origin (Marshall et al., 1998).

There are several different types of Aphasia, all of them coming with their own unique side-effects (Clough and Gordon, 2020). Their classification is not a trivial task, however, there is one thing they all share: making communication a difficult task. Findings in current theory (Cho-Reyes and Thompson, 2012) suggest frequent misuses of verbs and nouns, either from a character-mismatch or lexical swap perspective, and heavy syntactic alterations (Garraffa and Fyndanis, 2020). The discourse abilities might also be limited (Armstrong, 2000).

Our ultimate goal is to help People with Aphasia (PwA) to find their words easily by offering them a speech-to-speech system that corrects mispronounced sentences. To achieve this, we first need a parallel corpus where the source is composed by the altered spoken sentences and the target by what should have been spoken. In our knowledge, this

---

<sup>1</sup><https://www.aphasia.org/aphasia-resources/aphasia-factsheet/>



kind of corpus does not exist. What we propose in this article is to create such dataset by starting with sentences pronounced by PwA in speech therapy sessions and their correction, and then augment the corpus with sentence pairs automatically created based on the features of the initial data. We will also perform some preliminary translation experiments to show the overall feasibility of the approach that will lead to an aphasic speech correction system.

Our focus is on Broca patients. We believe that given the nature of popular rehabilitation methods, such as linguistic specific treatment (LST) (Thompson et al., 2003) and mapping therapy (Rochon et al., 2005), both of which are based on repetition of words, similar structures, or giving clues on remembering certain words or phrases, our instant feedback system based on speech translation would be of great help.

## 2 Related works

In natural language processing (NLP) and also in humanities, the availability of corpora is essential for understanding behavior phenomena and proposing tools or softwares based on NLP techniques or machine learning methods that facilitate the comprehension of such phenomena. In this particular topic, the aphasia data are rare and those that exist are not available.

One of the most attractive project developing Aphasia corpora is probably AphasiaBank (Forbes et al., 2012). The objective of AphasiaBank project was to provide researchers and clinicians with a large shared multimedia database of uniform discourse from individuals with and without aphasia. The database includes language samples in English, Spanish, German, Italian, Hungarian, Mandarin and Chinese. The aphasia section of this database contains approximately 180 videos of people with aphasia.

The project RELEASE (Williams et al., 2016) refers to the aphasia dataset of individual patient data for the rehabilitation and recovery of people with Aphasia after stroke. This project seems to be used by clinicians with the objective to study the rehabilitation. No information is given about the transcription of their utterances.

In the Moss Aphasia Psycholinguistics Project (MAPP) (Mirman et al., 2010), the authors provide a searchable database with data from more than 240 patients. The database is made up of the

Philadelphia Naming Test (PNT) results. The PNT is a single-word picture naming test developed to collect a large corpus of naming answers from patients.

Concerning the works that have addressed the problem of aphasia using automatic language processing approaches, we can cite the research below.

Since the grammatical deficiencies depend on the Primary Progressive Aphasia (PPA), in (Themistocleous et al., 2021) the authors propose to classify PPA variants by using part of speech (POS) production and to identify morphological markers that classify them by using machine learning. PPA is a very unique kind of aphasia. It is a form of dementia, and there are no cures available. Eventually, the person with this dementia completely loses their ability to comprehend and produce language due to gradual degradation (Thompson et al., 1997).

The study (Day et al., 2021) combines natural language processing and machine learning methods to predict the severity of PwA, both by score and severity level. The authors used a dataset of 238 participants extracted from AphasiaBank. They took the data from its transcript and composed the dataset by removing stop words and other items not necessary for this task. Stop word lists differ greatly, but they usually contain non-thematic words, like function words (determiners), prepositions (on, it, under), and so on. This is a very questionable decision, given the importance of the already few words people with aphasia are uttering. Stop words could be important indicators.

## 3 Building an Aphasic-French parallel corpus

In this section, we will describe how to build an Aphasic-French corpus (APHAFRECH) which will be used to show the feasibility of developing a communication rehabilitation support system for an aphasic person. To do this, we started by collecting real aphasia data in French that we transcribed, then we developed methods to build a parallel corpus that can be used to develop a machine translation system. We used several sources to build up a corpus for the analysis of aphasic errors. The first source is made up of videos extracted from the Web recorded in therapy sessions between speech therapists and PwA. In each video a speech therapist asked several questions to the PwA such as: *What is your name? How do you feel today? Describe*



what you see in this picture. We transcribed the PwA utterance and we corrected it. We retrieved seven dialogues that last from 3 to 20 minutes, the statistics concerning these videos are given in Table 1.

$d$	65'8''
$ \bar{d} $	8'8''
$ \bar{w} $	349
Males	3
Females	2

Table 1: Statistics about Aphasia videos.  $d$ : duration,  $w$ : word

The second source consists of the transcription of the reading of a text of 131 words by Guy de Maupassant<sup>2</sup> by people with aphasia. This kind of data should be handled with care: reading difficulties might be a by-product of another language disorder frequently accompanying aphasia: alexia. More of it in the next section.

The third source is based on the transcription of two conversations between a PwA and a speech therapist (Colin et al., 2016). This allows the PwA to speak and express themselves without being interrupted.

### 3.1 Analysis of the collected data

We analyzed the transcripts to characterize the effects of aphasia on speech. Several interesting details were observed, among them we can mention that aphasia leads to hesitations, the repetition of the same word or the same syllable, the interruption of speech and the use of periphrases.

In this article, we focus on Aphasia lexical errors. Our objective is to use minimal complexity and confusability in our data as what has been done for the images of PNT (Mirman et al., 2010) in order to facilitate the rehabilitation. In lexical errors, a word form is disturbed at several levels. It may concern the replacement of a character by another (*abricot* becomes *apricot*), swapping of syllables (*télévision* becomes *létévision*). Sometimes the PwA replaced a whole word by another one. This replacement can be explained by the pronunciation proximity (*cigarette* becomes *ciguerapette*) or by a semantic confusion. For example, a word could be replaced by another one semantically close for example *pain* (*bread*) is replaced by *vin* (*wine*) and in addition, in this case these two words are acoustically close to

<sup>2</sup>Pierrot in Contes de la bécasse, 1883

each other. Sometimes the PwAs create new words, it would seem from our study that they maintain the morphology of the words.

It's worth mentioning the influence of a potential aphasia-byproduct language disorder called alexia (Cherney, 2004). There are two main types of alexia: one influences vision in a physical way, the other one damages linguistic processing. Some of the results from reading tasks might be explained by the psycholinguistic deficits caused by a degradation in linguistic processing, and are not necessarily aphasia-related.

We identified from the transcriptions of 43 erroneous words belonging to the class of lexical errors, four categories: substitution, addition, deletion and replacement errors. Figure 1 illustrates the distribution of aphasic errors according to the Levenshtein distance between the correct word and its erroneous aphasic. This figure shows that 67% have a Levenshtein distance smaller or equal to 2 with the correct word.

Therefore, based on these figures, we believe that it is possible to create a large enough aphasia corpus by simulating errors close to those we encountered when analyzing real aphasia data. This will be done by introducing type errors: insertion, deletion and substitution, based on appropriate values of the Levenshtein distance.

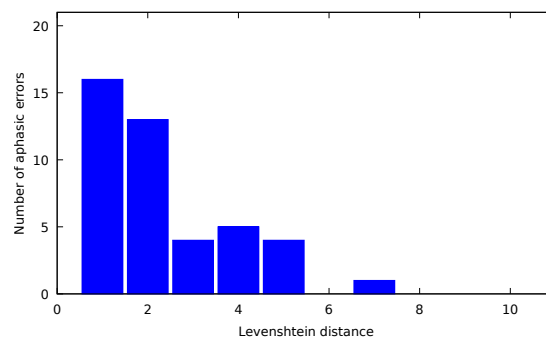


Figure 1: Distribution of aphasic errors according to the Levenshtein distance.

### 3.2 Automatic generation of pseudo-aphasic corpus APHAFRECH

Thanks to the errors studied during the analysis of aphasic data, we propose to create a pseudo-aphasic corpus automatically. In the following, we refer to all aphasic errors and their corrections, described in Section 3.1, as  $\mathcal{A}$  ( $\mathcal{A}$  for Aphasic). This corpus is made up of couples  $(a_i, c_i)$  where  $a_i$  is the  $i^{th}$

aphasic error and  $c_i$  is its correct version proposed by a human annotator.

In order to build APHAFRECH, we propose to start with  $\mathcal{C}$ , a clean corpus totally independent from the aphasia corpus described in Section 3.1. To build  $\mathcal{C}$ , we extracted 2,000 short sentences from the French part of the English-French file produced by Tatoeba project<sup>3</sup>. Then we generated from each sentence of  $\mathcal{C}$  a pseudo-aphasic sentence by applying rules based on the analysis of  $\mathcal{A}$ . For that, we apply the method described in Algorithms 1 and 2. For each sentence in  $\mathcal{C}$ , we randomly select words to alter with a fixed probability  $p$ . Then for each selected  $w$ , we produce  $n$  erroneous words potentially considered as words pronounced by a PwA. These  $n$  words are produced by using substitution, deletion and insertion of letters within  $w$ . Then, among these produced erroneous words, we select the best one  $w'$  which will replace  $w$ . In the alter function (Algorithm 2), for this first experiment, we allow only one alteration, but the algorithm can be later extended to lead to several alterations. We define this best erroneous word by using a scoring that yields the most likely altered word of having been pronounced by a PwA. With this two-step process, we want to give as much freedom as possible to the generation of errors, even if it means generating errors that are actually impossible to pronounce; step 2 then allows us to select plausible errors.

---

**Algorithm 1** Generation of APHAFRECH

---

**Require:** a corpus  $\mathcal{C}$ ,  $p$ ,  $n$   
**Ensure:** parallel corpus APHAFRECH  
 APHAFRECH  $\leftarrow \emptyset$   
**for each** sentence  $s \in \mathcal{C}$  **do**  
    $s' \leftarrow$  empty string  
   **for each** string  $w \in s$  **do**  
     **if** random()  $< p$  **then**  
        $w' \leftarrow$  alter( $w, n$ )  
     **else**  
        $w' \leftarrow w$   
     **end if**  
     add  $w'$  to  $s'$   
   **end for**  
 add couple  $(s, s')$  to APHAFRECH  
**end for**

---

<sup>3</sup><https://www.manythings.org/anki/>

---

**Algorithm 2** Generation word errors (function alter)

---

**Require:**  $w, n$   
**Ensure:** erroneous variant  $w'$  of  $w$   
 $v \leftarrow \emptyset$  (the set of variants)  
**for each**  $i$  from 1 to  $n$  **do**  
**repeat**  
    $w_i \leftarrow w$   
   alteration  $\leftarrow$  random("I","D","S")  
   **if** alteration = "I" **then**  
     insert randomly a character in  $w_i$   
   **else if** alteration = "D" **then**  
     delete randomly a character from  $w_i$   
   **else**  
     replace at random position a character of  $w_i$  by another one randomly selected  
   **end if**  
**until**  $w_i \notin v$  or a maximum number of iterations is reached  
 add  $w_i$  to  $v$   
**end for**  
**for each**  $w_i \in v$  **do**  
 give a score to  $w_i$   
**end for**  
 $w'$  is the  $w_i$  with the best score

---

### 3.2.1 Scoring a variant

Each erroneous variant is assigned a score that indicates to what extent it could have actually been produced by a PwA. Then, in the initial sentence, we replaced the concerned words by those that achieve the best error scores. To measure the quality of a variant, we tested two scoring functions. Actually, the variant  $w'$  is supposed to be pronounced by a PwA and as it is difficult to affirm that, the PwA wanted to say  $w$ , we should score this word. We define a measure  $f(w, w')$  that gives values between 0 ( $w'$  is certainly not an aphasic word spoken in place of  $w$ ) and 1 ( $w'$  could be certainly an aphasic word spoken in place of  $w$ ). In the following, we define the two score measures.

**ngram scoring** For this scoring, the quality of the erroneous string  $w'$  depends only on the likelihood of the character sequence. This likelihood is computed based on a character ngram language model that has been trained on a French novel (Germinal, by Émile Zola). For the smoothing method, we used Katz method (Katz, 1987). In sake of future coverage, the character vocabulary is the set of unicode ids. We propose to define the ngram

score by Equation 1.

$$ngram(w, w') = \frac{1}{m} \sum_{i=1}^m P(w'_i | h'_i) \quad (1)$$

Where  $m$  is the number of characters of  $w'$  and  $h'_i$  is the character sequence preceding  $w'_i$ . In case of  $n$ -gram,  $h'_i$  is truncated to the  $n - 1$  preceding characters.

Let's remark that  $ngram(w, w')$  does not depend on  $w$  because we want only measure the likelihood of  $w'$  independently of the lexical distance between  $w$  and  $w'$  (this distance is fixed to 1 in this experiment).

**soundex scoring** For `soundex`, the words  $w$  and  $w'$  are close if their respective pronunciations are close. To estimate the degree of closeness of words, we compared the soundex encoding of  $w$  and  $w'$ . Soundex (Jacobs, 1982), is a method for indexing words by their sound. Words are encoded by taking advantage of their phonetic form. The encoding is done in both words, the altered and the correct one. The principle of encoding a word consists in deleting spaces, uppercasing the word, keeping the first letter, deleting the vowels, associating digits to each letter in accordance to its phonetic class (see Table 2) and finally by keeping the first four characters.

Phonological group	digit
B,P	1
C,K,Q	2
D,T	3
L	4
M,N	5
R	6
G,J	7
X,Z,S	8
F,V	9

Table 2: Soundex codes for each phonological group

With this encoding function, words like *ballon* and *ballon* will receive the same code B445, while the encoding of the words *farapluie* and *parapluie* will be respectively F614 and P614.

Then we estimate the Soundex closeness of  $w$  and  $w'$  by  $soundex(w, w')$ , defined by Equation 2.

$$soundex(w, w') = \frac{1}{4} \sum_{i=1}^4 \delta(S_i(w), S_i(w')) \quad (2)$$

Where  $S_i(x)$  is the  $i^{th}$  soundex code of the word  $x$ .  $\delta(x, y)$  returns 1 if  $x$  is equal to  $y$ .

**Performance of scoring functions** In order to measure the capability of `ngram` and `soundex` to give a score close to 1 to real aphasic errors, we use  $\mathcal{A}$  as a test corpus. We injected each real aphasic error  $a_i$  into the list of pseudo aphasic errors provided by Algorithm 2. Table 3 shows the average of the inverse rank of  $a_i$  in the list sorted according the `ngram` and `soundex` scores. For `ngram`, we tested several values of  $n$ , the best results have been achieved for  $n = 4$ . The result of the `soundex` function leads to a very low performance compared to the `ngram` function. This is due to the distribution of the `soundex` function scores which has a tiny standard deviation.

Scoring function	Performance
4-gram	0.26
Soundex	0.02

Table 3: Performance of scoring functions on the produced errors

#### 4 A preliminary experience in Machine translation of a pseudo aphasic corpus

In this section, we study the opportunity to translate an aphasic corpus to its corrected counterpart. For that we use APHAFRESH, the parallel corpus we described in Section 3. To generate the aphasic sentences, we used only the `ngram` scoring function since it is the one that achieves the best aphasia errors. Table 4 shows a sample of this corpus.

Pseudo-aphasia sentences	Correct sentences
sois juite	sois juste
j'ai fait sine	j'ai fait signe
je duis calme	je suis calme
je me suis révemillé	je me suis réveillé
je suis detite	je suis petite
si ça ne vous dérande pas, pourrions-nous inspecter votre valise	si ça ne vous dérange pas, pourrions-nous inspecter votre valise ?

Table 4: A sample of the parallel experimental pseudo-aphasic corpus APHAFRESH

Our ultimate objective is to provide an aphasic speech to natural speech translation system. But, in this preliminary experience, we will study the

opportunity to translate an aphasic corpus to its corrected counterpart. For that, we will train a sequence-to-sequence machine translation model, a kind which has been used widely in the literature of machine translation and other NLP applications (Sutskever et al., 2014; Zhang et al., 2015; Nguyen Le et al., 2017; Mao et al., 2020). We used the corpus we created, APHAFRESH, for training, tuning and for testing.

The input of the encoder is the Aphasia sentence and the output is the hidden state and cell state of the LSTM. The decoder has the hidden state and cell state of the encoder as inputs in addition to the input sentence. The results of the decoder LSTM is passed through a dense layer to predict decoder outputs as shown in Figure 2.

In Table 5, we give the different parameters of the neural network architecture we used.

Parameters	Values
Source Maximum Length sentence	13
Target Maximum Length sentence	14
Source Unique words	13,085
Target unique words	8,364
Batch Size	64
Epochs	20
Number of LSTM Nodes	400
Embedding Size	100
SPLIT Training-Tuning	0.1
Test size	2,000

Table 5: The parameters of the sequence-to-sequence model

Concerning the optimizer, in our experiments we tested several methods, the one which achieves the better results is the Adaptive Gradient Algorithm (Duchi et al., 2011). In fact, adaptive gradient algorithms calculate gradient-based updates using the history of gradients, which has the advantage to reduce the inconvenience of manually setting the step size parameter in the stochastic gradient descent optimizer. In addition, AdaGrad is known also for its computational efficiency (Kingma and Ba, 2014). From Figures 3 and 4, we can conclude that the accuracy is high and the model reduced the value of the loss, which means that the model makes small errors on few data and the model predicts well. The training and the validation curves start with relatively high loss at the beginning and gradually decrease as training and validation examples are added and gradually flatten, indicating that

adding more examples does not improve the performance of the model on both data. This leads us to assume that our neural network does not overfit.

We tested this model in a test corpus composed of 2,000 aphasia sentences. The results in terms of cumulative BLEU are given in Table 6.

BLEU1	BLEU2	BLEU3	BLEU4
59.24	51.20	44.39	38.60

Table 6: Cumulative BLEU on the pseudo-aphasia corpus

Figure 5 illustrates the distribution of the BLEU score over the 2,000 sentences of the test corpus. We can observe that more than 31% of the sentences have a BLEU higher than 50 which means that we achieve a very high quality and fluent translation. Only 5% of the translation have a BLEU smaller than 10 which corresponds in general to a useless translation. 19% of the translations have a BLEU between 10 and 19, which corresponds to sentences that are difficult to understand.

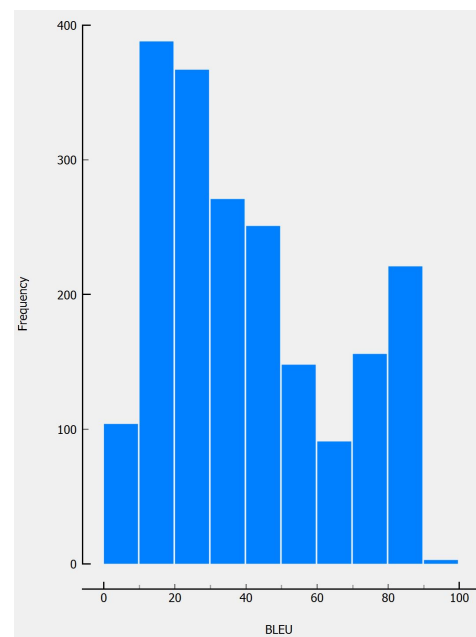


Figure 5: Distribution of the level of understanding of the translation of the 2000 Aphasia sentences

In order to make the reading of Figure 5 easy, Table 8 recalls how to interpret the BLEU score (Noever et al., 2021) accordingly to the quality of the translation.

The global analysis of the BLEU score on the different sentences of the test corpus is illustrated by Table 7.

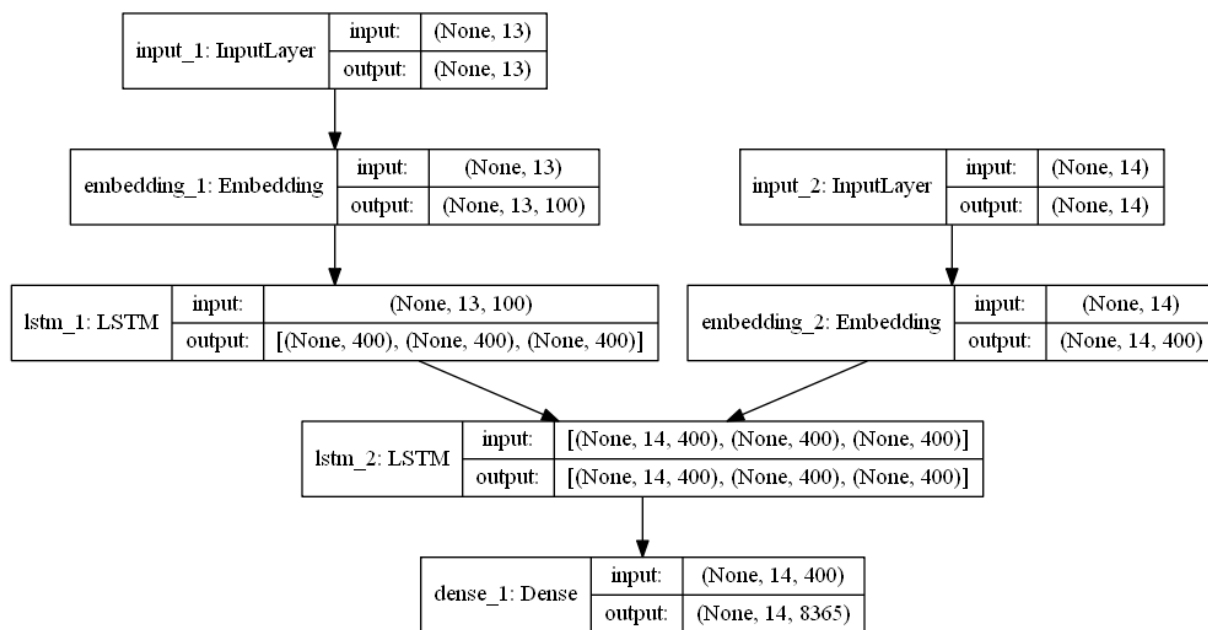


Figure 2: Architecture of the Aphasia sequence-to-sequence model

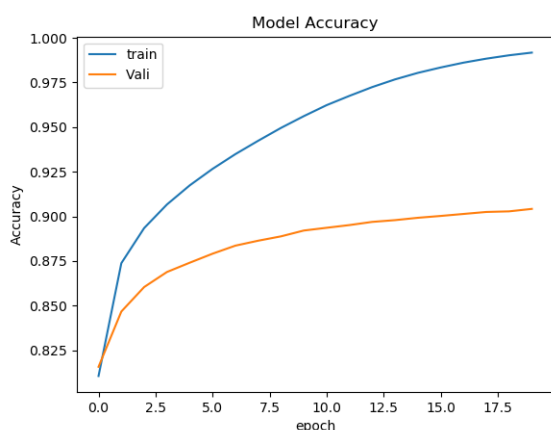


Figure 3: The accuracy on the training and the validation corpus

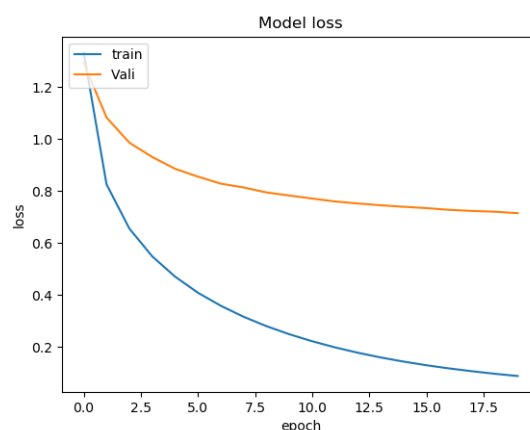


Figure 4: The loss function on the training and the validation corpus

Mean	SD	Max	Min
40.62	24.09	90.48	4.18

Table 7: Some figures concerning the BLEU scores of the Aphasia to natural text machine translation

Figure 6 is a different presentation of Figure 5, it shows the decreasing evolution of the values of BLEU. We can notice that more than 25% of the test corpus was translated with a performance of at least 50 in terms of BLEU.

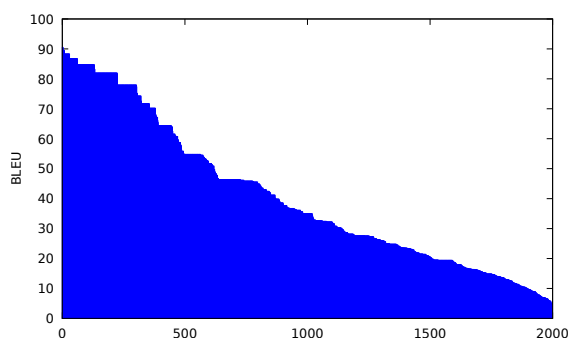


Figure 6: A decreasing distribution of BLEU over 2,000 sentences.



BLEU	Score Interpretation
< 10	Almost useless translations
10 to 19	Difficult to understand
20 to 29	The idea is clear, but it contains many errors
30 to 40	correct translations
40 to 50	High quality translations
50 to 60	Very high quality
> 60	Quality better than a human translation

Table 8: Interpretation of the quality of machine translation depending on the BLEU score

## 5 Conclusion

Aphasia is a unique and rather complex phenomena. There is a great amount of work trying to understand and explain the underlying structural changes from different perspectives. Since there is no general consensus on what the best approach is to therapy, the field remains open for experimentation. We decided to take up the challenge from a machine learning perspective by implementing a method that will eventually allow us to come up with a speech-to-speech system where the input is aphasic speech and the output is a rehabilitated speech. For that, we created an aphasia-like corpus, APHAFRECH, with correct-incorrect sentence pairs, using three different resources. This required the study of errors from aphasic sources in order to understand certain types of errors and to reproduce them automatically. With the created dataset we trained a neural network machine translation that yields very high quality translations on APHAFRECH. The next step will concern the introduction of more complex aphasia errors into APHAFRECH (such as context-dependant errors and semantic based errors) and the study of the quality of the translation by using a more elaborated DNN machine translation.

## References

Elizabeth Armstrong. 2000. [Aphasic discourse analysis: The story so far](#). *Aphasiology*, 14(9):875–892.

Leora Reiff Cherney. 2004. Aphasia, alexia, and oral reading. *Topics in Stroke Rehabilitation*, 11(1):22–36.

Soojin Cho-Reyes and Cynthia K. Thompson. 2012. Verb and sentence production and comprehension in aphasia: Northwestern assessment of verbs and sentences (navs). *Aphasiology*, 26(10):1250–1277.

Sharice Clough and Jean K. Gordon. 2020. Fluent or nonfluent? part a. underlying contributors to categorical classifications of fluency in aphasia. *Aphasiology*, 34(5):515–539.

Capucine Colin et al. 2016. Adaptation du projet aphasiabank à la langue française: Contribution pour une évaluation informatisée du discours oral de patients aphasiques.

Marjory Day, Rupam Kumar Dey, Matthew Baucum, Eun Jin Paek, Hyejin Park, and Anahita Khojandi. 2021. Predicting severity in people with aphasia: A natural language processing and machine learning approach. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 2299–2302.

John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159.

Margaret Forbes, Davida Fromm, and Brian Macwhinney. 2012. AphasiaBank: a resource for clinicians.

Maria Garraffa and Valantis Fyndanis. 2020. [Linguistic theory and aphasia: an overview](#). *Aphasiology*, 34(8):905–926.

J.R. Jacobs. 1982. Finding words that sound alike. the soundex algorithm. *Byte* 7, pages 473–474.

Slava Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3):400–401.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Zhuoyuan Mao, Fabien Cromieres, Raj Dabre, Haiyue Song, and Sadao Kurohashi. 2020. Jass: Japanese-specific sequence to sequence pre-training for neural machine translation. *arXiv preprint arXiv:2005.03361*.

Randolph S Marshall, Ronald M Lazar, and JP Mohr. 1998. Aphasia. *Medical Update for Psychiatrists*, 3(5):132–138.

Daniel Mirman, Ted J. Strauss, Adelyn Brecher, Grant M. Walker, Paula Sobel, Gary S. Dell, and Myrna F. Schwartz. 2010. A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. *Cognitive Neuropsychology*, 27(6):495–504.

An Nguyen Le, Ander Martinez, Akifumi Yoshimoto, and Yuji Matsumoto. 2017. Improving sequence to sequence neural machine translation by utilizing syntactic dependency information. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 21–29, Taipei, Taiwan. Asian Federation of Natural Language Processing.



- D. Noever, Josh Kalin, Matthew Ciolino, Dom Hambrick, and Gerry Dozier. 2021. Local translation services for neglected languages. pages 149–163.
- Elizabeth Rochon, Laura Laird, Arpita Bose, and Joanne Scofield. 2005. [Mapping therapy for sentence production impairments in nonfluent aphasia](#). *Neuropsychological rehabilitation*, 15:1–36.
- Debbie Summers, Anne Leonard, Deidre Wentworth, Jeffrey L Saver, Jo Simpson, Judith A Spilker, Nanette Hock, Elaine Miller, and Pamela H Mitchell. 2009. Comprehensive overview of nursing and interdisciplinary care of the acute ischemic stroke patient: a scientific statement from the american heart association. *Stroke*, 40(8):2911–2944.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.
- Charalambos Themistocleous, Kimberly Webster, Alexandros Afthinos, and Kyrana Tsapkini. 2021. Part of speech production in patients with primary progressive aphasia: An analysis based on natural language processing. *American journal of speech-language pathology*, 30(1S):466–480.
- C. K. Thompson, K. J. Ballard, M. E. Tait, S. Weintraub, and M. Mesulam. 1997. [Patterns of language decline in non-fluent primary progressive aphasia](#). *Aphasiology*, 11(4-5):297–321.
- Cynthia Thompson, Lewis Shapiro, Swathi Kiran, and Jana Sobecks. 2003. [The role of syntactic complexity in treatment of sentence deficits in agrammatic aphasia](#). *Journal of speech, language, and hearing research : JSLHR*, 46:591–607.
- L. Williams, M. Ali, K. VandenBerg, J. Godwin, A. Elders, F. Becker, A. Bowen, C. Breitenstein, M. Gandolfi, E. Godecke, K. Hilari, J. Hinckley, S. Horton, D. Howard, L.M.T. Jesus, M. Jungblut, M. Kambanaros, T. Kukkonen, A. Laska, B. MacWhinney, I. Martins, F. Mattioli, M. Meinzer, R. Palmer, B. Patrício, C. Price, N. Smania, J. Szaflarski, S. Thomas, E. Visch-Brink, L. Worrall, and M. C. Brady. 2016. Creating an international, multidisciplinary, aphasia dataset of individual patient data (ipd) for the rehabilitation and recovery of people with aphasia after stroke (release) project. *International Journal of Stroke*, 11(4).
- Jiajun Zhang, Chengqing Zong, et al. 2015. Deep neural networks in machine translation: An overview. *IEEE Intell. Syst.*, 30(5):16–25.

# Current Shortcomings of Machine Translation in Spanish and Bulgarian Vis-à-vis English

Travis Sorenson

University of Central Arkansas

tsorenson@uca.edu

## Abstract

In late 2016, Google Translate (GT), widely considered a machine translation leader, replaced its statistical machine translation (SMT) functions with a neural machine translation (NMT) model for many large languages, including Spanish, with other languages following thereafter. Whereas the capabilities of GT had previously advanced incrementally, this switch to NMT resulted in seemingly exponential improvement. However, half a dozen years later, while recognizing GT's usefulness, it is also imperative to systematically evaluate ongoing shortcomings, including determining which challenges may reasonably be presumed as superable over time and those which, following a multiyear tracking study, prove unlikely ever to be fully resolved. While the research in question principally explores Spanish-English-Spanish machine translation, this paper examines similar problems with Bulgarian-English-Bulgarian GT renditions. Better understanding both the strengths and weaknesses of current machine translation applications is fundamental to knowing when such non-human natural language processing (NLP) technology is capable of performing all or most of a given task, and when heavy, perhaps even exclusive human intervention is still required.

**Keywords:** Bulgarian, English, Google Translate, machine translation, Spanish

## 1 Theoretical introduction and historical overview

The genesis of this study lies in events that, while years in the making, came to light fully in

late 2016, when programmers behind the scenes switched the online machine translation service Google Translate (GT) from one employing statistical machine translation (SMT) to one relying on the company's newly completed neural machine translation (NMT) system (Lewis-Kraus, 2016). Rather than featuring different modules, NMT utilizes a single, streamlined system that contains only an encoder, which analyzes the training data (mostly bilingual corpora), and a decoder, which applies this analysis to a new source-language text and renders it into the chosen target language. While the encoder assigns individual words and other features numerical qualities, the decoder considers texts to be translated at the full sentence level, rather than according to separate words or phrases as with the SMT models (Poibeau, 2017: 185). This seeming simplicity should not obscure the fact that NMT is not only extremely complex, but, given that the representation of the data is strictly numerical, it is not completely understood even by those who have written the algorithms leading to the vectors of numbers involved in the work of encoding the bilingual texts (193). While NMT can compete with human translators on tasks involving highly repetitive structures (e.g. legal documents, economic texts), less common and more creative, more novel content can lead to serious meaning errors. In other words, adequacy may suffer even if the fluency of the resulting translation may be acceptable. The main cause of this difficulty on the part of computers to engage successfully in natural language processing (NLP) is ambiguity (lexico-semantic, morphosyntactic, etc.) (Koehn, 2020: 37). A number of examples displaying this phenomenon are considered in this paper.

## 2 Overview of research

The purpose of the larger research project – based mainly on English and Spanish – is to determine not only what NMT can and cannot do, both generally and specifically, but also what improvements might occur over the next five years or so. As this research focuses largely on the written word, a thorough assessment of these matters requires a systematic evaluation of the different categories involved, namely expository writing, descriptive/narrative writing, and persuasive writing via texts from numerous subcategories in each case. This allows for methodical contrasting and comparing of the results yielded by GT, some of which are unique to the relationship between English and Spanish, whereas others have features that can be extrapolated to other languages, including Bulgarian.

## 3 Presentation, discussion, and analysis of research data

Initial work drawing on material from the above-mentioned categories will now allow for a discussion of GT results stemming from a variety of texts. While GT is capable of performing felicitous translations in many content areas, and while it has arguably improved vastly in all areas since the switch to NMT beginning in 2016, many renditions continue to be problematic to one degree or another. While many such instances are given throughout this paper, it is appropriate at this point to share two such examples (i.e. one that is considered to be a suitable, even excellent translation by GT, and another that is decidedly flawed). Both happen to be from English to Spanish. The first is from an economic report published by the business and finance website *cnn.com* (Fitzgerald and Stevens, 2021):

- (1) July's Consumer Price Index released Wednesday showed prices jumped 5.4% since last year, compared to expectations of 5.3%, according to economists surveyed by Dow Jones. The government said CPI increased 0.5% in July on month-to-month basis.

### GT (4 Oct 2021):

*El índice de precios al consumidor de julio publicado el miércoles mostró que los precios subieron un 5,4% desde el*

*año pasado, en comparación con las expectativas del 5,3%, según economistas encuestados por Dow Jones. El gobierno dijo que el IPC aumentó un 0,5% en julio mes a mes.*

The GT rendition of passage (1) into Spanish represents an arguably flawless translation, including several important but somewhat subtle details. For instance, the initials CPI for Consumer Price Index have been converted, appropriately, to IPC (*índice de precios al consumidor*). Next, whereas in English no article is used before the expression of percentages, GT inserted an indefinite article in one case (*un 5,4%*), and a definite one in the other (*el 5,3%*). The use of either or both is more common than not in authentic Spanish. Finally, while the decimal separator employed in English is the period, most South American countries use the comma, which is also the case in Spain, which is part of the European Union. Apropos of that, the main reason for the accuracy of this translation, including all the intricacies mentioned, surely lies in the fact that myriad such texts that have been translated between English and Spanish and vice versa – such as those related to the European Parliament's Committee on Economic and Monetary Affairs (ECON), which are found in the Europarl parallel corpus, to which GT has access.

The second example is a set of simple questions that one may easily presume would pose no great difficulty for GT:

- (2) How are you, Dad?

How are you, Father?

How are you, Mom?

How **are you**, Mother?

### GT (26 Mar 2022):

*¿Cómo estás, papá?*

*¿Cómo estás, padre?*

*¿Cómo estás mamá?*

*¿Como **está tu** madre?*

Whereas the first two translated sentences in example (2) are in no way problematic, the final two have a minor and then major errors. While *¿Cómo estás mamá?* inexplicably lacks the comma present in the two previous sentences, it is a detail that does not seriously impede

understanding, essentially continuing to pose the same question. The final sentence, in contrast, suffers a catastrophic semantic change with a shift from second to third person: ¿*Como está tu madre?* means ‘How **is your** mother?’, not ‘How **are you**, Mother?’

The following sections explore various issues, organized by common themes, that arise in GT renditions of original texts in different combinations of Spanish, English, and Bulgarian.

### 3.1 Pronoun-dropping and pronoun confusion between animate and inanimate objects

Whereas pronoun-dropping rarely occurs in English, it is quite common in many other languages, including Spanish and Bulgarian. In these pro-drop languages, other context markers, such as verb conjugations, serve to supplant much of the information carried in the missing pronouns, especially subject pronouns. However, since the context unavoidably becomes more implicit in the absence of the explicit pronouns, ambiguity unavoidably results. Although this type of situation is routinely processed without difficulty by humans, translation platforms such as GT are prone to significant meaning errors under the same conditions, as the following cases from Spanish and Bulgarian into English aptly demonstrate.

Writing about her experience covering the election of Pope Francis, Argentine journalist Elisabetta Piqué wrote the following in her book *Francisco: vida y revolución* (2014):

- (3) *Lo recuerdo bien. **Estaba** en la plaza, embarazada de mi primer hijo de 6 meses, Juan Pablo.*

#### GT (8 Apr 2022):

‘I remember it well. **She** was in the plaza, pregnant with my 6-month-old first child, Juan Pablo.’

Author’s translation:

‘I remember it well. I was in the plaza, six months pregnant with my first child, Juan Pablo.’

Beyond committing the also serious error of stating that a fetus at the sixth-month stage of pregnancy is in reality a sixth-month-old child (a miscalculation of approximately nine months), GT takes what is clearly (to a human reader) a first-person reference and turns it into a third-person one. In the imperfect aspect of the Spanish past tense, the conjugation *estaba* (<infinitive *estar* ‘to be’) corresponds to various potential subjects: ‘I’ (*yo*), ‘he’ (*él*), ‘she’ (*ella*), ‘you’ (formal: *usted*), and ‘it’ (Ø). However, since the initial sentence was ‘I remember it well,’ it is clear that the one following also continues with the first person: ‘**I** was in the plaza...’ Not only would the GT rendition indicate that the father of the expected child was the one narrating and referring to the mother in the third person, but if such were the case the writer would almost surely have used an overt pronoun to make this abundantly clear: *Ella estaba en la plaza...* GT, processing largely at the sentence level, has no intersentential context on which to rely. Curiously, if only the segment *Estaba en la plaza* is processed, GT yields ‘**I** was in the square.’ It is likely, therefore, that the feminine word *embarazada* ‘pregnant’ incorrectly triggered a feminine pronoun: ‘she.’

A similar phenomenon can be witnessed with pronoun-dropping in Bulgarian, such as in the following pair of similar examples, the counterparts of which are also considered in Spanish:

- (4) Виждам **я**. Идва.<sup>1</sup>  
**(I)** see **her**. **(She)** comes.
- (5) Виждам **го**. Идва.  
**(I)** see **him**. **(He)** comes.

#### GT (5 Mar 2022):

‘I see **her**. **It’s** coming.’

‘I see **it**. **It’s** coming.’

- (6) **La** *veo*. *Viene*.  
**Her** **(I)** see. **(She)** comes.
- (7) **Lo** *veo*. *Viene*.  
**Him** **(I)** see. **(He)** comes.

<sup>1</sup> As the author of this paper does not speak Bulgarian, Google Translate was used to aid in the creation of some of the source-language sentences analyzed herein.

**GT (8 Apr 2022):**

‘I see her. She comes.’

‘I see. Comes.’

In the initial two-word sentence in both pairs of examples, it must be supposed that the speaker is referencing the sighting of a man and then a woman, respectively, as the literal gloss indicates. It should also be understood, in the case of the Bulgarian examples, that this language does not have infinitive verb forms; the first person singular conjugation in the present is therefore employed to refer to a verb, after which, as occurs in Spanish, the endings change for other persons and according to tense and aspect (see Table 1 for Bulgarian ‘see’). Nevertheless, (4) and (5) both begin with a verb whose ending, in context, clearly refers, if only by default, to the first person singular. Use of the overt pronoun аз ‘I’ is unnecessary. However, in the second sentence of each example (a one-word verb phrase), the meaning is only implicit, as the verb form идва can mean ‘he/she/it is coming.’ The intent is clear to the speaker, but not to GT, which in both renditions has opted for the impersonal ‘it.’ Regarding example (6), even though the focus of GT’s analysis tends to skew heavily to individual sentences, it seems that the presence of feminine *la* in the first sentence aided its correct choice of ‘She’ in its rendition of the second. However, not knowing if *lo* referred to ‘him’ or ‘it,’ GT omitted both in its version of (7), leaving a first sentence that lacks the needed object pronoun and a second one that is incomplete.

English	Bulgarian
‘I see’	аз виждам
‘you see’ (sing.)	ти виждаш
‘he sees’	той вижда
‘she sees’	тя вижда
‘it sees’	то вижда
‘we see’	ние виждаме
‘you see’ (plur.)	вие виждате
‘they see’	те виждат

Table 1: Present tense of Bulgarian verb ‘see’ (виждам) with overt subject pronouns

As manifested in example (5) in Bulgarian and (7) in Spanish, it is not only the omitted subject pronouns that can prove problematic for machine translation, but also the ambiguity of the clitics that are *not* left out. Whereas English features the unambiguous direct object pronouns ‘me’, ‘you’,

‘him’, ‘her’, ‘it’, ‘us’, and ‘them’, Spanish and Bulgarian both have counterparts of these pronouns that are clear in some instances and ambiguous in others. In Spanish, the equivalent of ‘me’, ‘you’ (sing. informal), ‘us’, and ‘you’ (plur. informal) are *me*, *te*, *nos*, *os*, all of which are distinct and therefore straightforward. However, ‘him’, ‘her’, ‘you’ (sing. formal), and ‘it’ can all be expressed *lo* or *la*, depending on gender, while ‘them’ and ‘you’ (plur. formal) are similarly either *los* or *las*. Faced with this uncertain situation in example (7), GT, as noted above, offered no equivalent pronoun at all in English, leaving only ‘I see,’ an intransitive verb use despite the fact that the sentence calls for a transitive construction, whether it be ‘I see him’ or ‘I see it.’ In Bulgarian, there is also some overlap in direct object pronouns, though it is limited to third person singular forms: ‘him’ and ‘it’ (masc./neut.) are both *го*, and ‘her’ as well as ‘it’ (fem.) are *я* (see Table 2 for all forms). In example (5), GT incorrectly selected inanimate ‘it’ in lieu of animate ‘him,’ even though in (4) it correctly chose animate ‘her.’ Nevertheless, if GT tends to render *я* as ‘she’ in all or most instances involving this type of ambiguity, eventually it will err – as it does in the next example – when this pronoun refers to an inanimate object that is assigned the female gender, the case with many Bulgarian nouns that end in *-a* and *-я*, such as *маса* ‘table’ in the following example given by Leafgren (2011: 74):

(8) Това е новата ни маса.  
 This is new our table.  
 Татко иска да я поставим  
 Father wants (aux.) it put  
 в ъгъла в  
 in (the) corner in (the)  
 кухнята.  
 kitchen.

‘This is our new table. Dad wants us to put it in the corner in the kitchen.’

**GT (7 Apr 2022):**

‘This is our new table. Dad wants us to put her in the corner of the kitchen.’

English	Bulgarian (long)	Bulgarian (short)
‘me’	мен	ме
‘you’ (sing.)	Теб	те

'him'	Него	ГО
'her'	Нея	Я
'it' (mas./neut.)	Него	ГО
'it' (fem.)	Нея	Я
'us'	Нас	НИ
'you' (plur.)	Вас	ВИ
'them'	Тях	ГИ

Table 2: Bulgarian accusative case (direct object) pronouns

### 3.2 Difficulties related to the gender of nouns and adjectives

As seen in the previous section, the use of certain third-person accusative pronouns in both Spanish and Bulgarian depends on the gender of either the person or the inanimate object that they modify. In both languages, gendered nouns themselves (and modifying adjectives) can also lead to difficulties for GT when ambiguities related to them arise in complex source-language material. In this regard, Koehn (2020: 7) proposes the following sentence in English, which is then translated into Spanish and Bulgarian, respectively:

- (9) 'Whenever I visit my uncle and his daughters, I can't decide who is my favorite cousin.'

GT (8 Apr 2022):

*Cada vez que visito a mi tío y a sus hijas, no puedo decidir quién es mi primo favorito.*

Винаги, когато посещавам чичо си и дъщерите му, не мога да реша кой е любимият ми братовчед.

While a human has little problem with the logical deduction that the daughters of the uncle are by necessity female cousins, the link is not explicit enough for GT to avoid falling into the trap, which is set by the fact that English has no endings or any other morphological markings that render nouns and adjectives inherently masculine or feminine. As a result, in each instance, both the noun and its accompanying adjective were rendered in masculine form in the translation. In Spanish, 'female cousin' is *prima* rather than *primo*, and the single feminine form of 'favorite' is *favorita*, not *favorito*. The same order of correct results in Bulgarian is братовчедка rather than братовчед, and любимата instead of любимият.

### 3.3 Lexical differences by regional dialect and the effects of homonymia

An important part of translation entails being able to insert the target language into its appropriate place in terms of culture and geography, a subfield of the discipline called localization. An essential element of this effort has to do with the suitable choice of specific vocabulary. If, for instance, a text in German about a *wohnung* were to be rendered into English, the translator would need to consider not only the target language but the pertinent dialect thereof. For a British audience the term 'flat' would be most appropriate, while US readers would identify with 'apartment.' In Spanish, at least three terms suggest themselves depending on the country or region: *piso* in Spain, *departamento* in Argentina, and *apartamento* in most of the rest of the Spanish-speaking world. The following examples, one from English to Spanish and the other in reverse order, are from the culinary world and show the importance of having certain lexical expertise in Spanish, a language particularly rich in synonym usage:

- (10) 'It is a common practice to sauté mushrooms in butter.'

GT (8 Apr 2022):

Es una práctica común saltar los champiñones en mantequilla.

- (11) *Es más fácil tomar la soda con un carrizo.*

GT (9 Apr 2022):

'It's easier to drink soda with a reed.'

The GT rendition of example sentence (10) would serve well in many Spanish-speaking countries, including certain large ones with high populations such as Mexico and Spain. In others, however, at least one of the food words would be uncommon to point of near non-existence. In Central American countries such as Honduras, Costa Rica, and Panama, the dominant term for 'mushrooms' is *hongos*. In Argentina, Uruguay, and Paraguay, the nearly universal term for 'butter' is *manteca*, despite that fact that in most other countries this name refers to 'lard.' Sentence (11) is one that could be heard throughout Panama, the only country where the default term for 'drinking straw' is *carrizo*, which, while it does mean 'reed' in other dialects,



is employed metaphorically in this Central American country to denote a manmade hollow tube for sipping liquids. If Panamanians need to refer to a ‘reed,’ they have available for this purpose the word *caña* (which, in turn, is at times used in Perú not for a stem from the plant kingdom but, again, for ‘drinking straw,’ though the diminutive *cañita* is much more common for this purpose).

While national boundaries can often determine word usage, such as *carrizo* for ‘drinking straw’ in Panama alone, in some larger countries there may well be various intranational regions for a number of lexical items. For instance, in Spain, while speakers in nearly all areas of the country employ the term *judía* or *judía verde* to denote the ‘green bean,’ in parts of the north, including the Basque Country, the term *vaina* prevails. Something similar is seen in Bulgaria, this time in the animal rather than the plant kingdom, and between the eastern and western zones of the country. Whereas the lexemes *gato* and ‘cat’ are universal in all dialects of Spanish and English, respectively, for the domesticated feline, Garavalova (2020) – referencing the *Bulgarian Etymological Dictionary*/Български етимологичен речник (BER, 1986) and the *Bulgarian Dialect Atlas*/Български диалектен атлас (BDA, 2001) – asserts that while speakers in the eastern two-thirds of Bulgaria tend to utilize the name **котка** (<\*kotja ‘female cat’ <кот <Proto-Slavic \*katъ <Latin *cattus*), in the western third or so of the country it is not uncommon to hear the term **мачка** (etymology uncertain, but shared with Serbian in Cyrillic form and as *mačka* in Croatian, Slovak, and Slovenian with the same pronunciation) (104-106). Of interest, then, is the GT rendition of the following sentence:

(12) ‘I don’t like this **cat**.’

GT (9 Apr 2022):

He харесвам тази **котка**.

If the target audience were speakers in central and eastern Bulgaria, the selection of **котка** would be optimal. If however, the intended group were those in the west, including the capital of Sofia, an acceptable localized rendition for many would be: He харесвам тази **мачка**. It is presumed, nevertheless, that most if not all speakers in western Bulgarian would understand both names – particularly if **котка** is the more normative of these two lexemes – though this is not always the case with dialectally determined

vocabulary, especially with larger languages spread over wide expanses of the globe and several countries, such as English and Spanish.

A return to food vocabulary in the following Spanish-to-English translation will help to illuminate another issue that can arise with the GT renditions of texts involving dialectally based terminology.

(13) ***Las manías** son caras ahora mismo. Las almendras cuestan menos.*

GT (8 Apr 2022):

‘**Crazes** are expensive right now. Almonds cost less.’

Author’s translation:

‘Peanuts are expensive right now. Almonds cost less.’

One immediately notices an incongruence in the machine translation, as ‘crazes’ do not generally carry a price tag and have precious little to do with ‘almonds.’ If, however, it is realized that both sentences in the original text concern a type of ‘nut,’ a bit of research on the matter can lead to a lexical solution in English. While the Náhuatl-derived *cacahuate* is the principle term for ‘peanut’ in Mexico, in the Caribbean, much of Central America, and all of South America, the dominant name is *maní* (plural *manís* or *maníes*), from the now-extinct Amerindian language Taíno. Yet only in Guatemala does one hear the altered form *manías*, as used in sentence (13), which is why GT failed to recognize the term’s true meaning and translated it as ‘crazes,’ since in other contexts Spanish *manía* can signify ‘mania,’ a word denoting ‘madness’ in English that entered both languages via Latin from the earlier Greek. This means that the two cases of *manía(s)* in Spanish are homonyms: lexemes with the same spelling, the same pronunciation, but different meanings (typically with two different etymologies).

The difficulties presented at times by homonymia are not unique to Spanish; the following pair of examples shows that the phenomenon can also occur in English and Bulgarian, though it appears to be much more common in the former than the latter:

(14) ‘The dog was old and sick; its **bark** was very weak.’

**GT (23 Dec 2021):**

*El perro era viejo y estaba enfermo; su corteza era muy débil.*

- (15) ‘I like the feel of this scythe.’

**GT (8 Apr 2022):**

Харесва ми усещането за тази коса.

In the original text of sentence (14), the English word ‘bark’ obviously refers to the sound emanating from the dog’s mouth. The Spanish equivalent of this, however, is *ladrido*. The lexeme given by GT, *corteza*, refers to a protective outer layer of vegetable matter that constitutes the ‘bark’ of a tree. Moving to Bulgarian, the GT rendition from English of sentence (15) is only problematic if the reader does not understand the context, which could be that an individual has mentioned the need to find a useful tool for cutting grass or harvesting grain by hand. In isolation, however, коса could be understood as the homonym ‘hair.’ A person who does not often work with farm implements might well understand the word in its agricultural context and yet go years without encountering it in this setting. In contrast, one may refer to ‘hair’ on a weekly if not daily basis. This is surely the reason for the fact that GT, when processing a back translation of the Bulgarian sentence into English, opts for ‘hair’ as the equivalent of the term in question.

- (16) Харесва ми усещането за тази коса.

**GT (8 Apr 2022):**

‘I like the feel of this hair.’

**3.4 Additional examples in English and Spanish**

Whereas many of the examples of problematic GT renditions shown to this point in the paper have included issues that one might find in relation to Bulgarian, there are myriad others that may pertain less or not at all to this language. If linguists of any native language are to grasp more fully the challenges still presented by machine translation, it is ultimately necessary that they expose themselves to such phenomena in multiple tongues, not to mention the various dialects of each. For instance, examples (15) and (16) above

concerning the use of *koca* to denote both ‘hair’ and ‘scythe’ appears to be a rather rare instance of homonymia in Bulgarian. In contrast, the use of identically spelled and pronounced words in both Spanish and English is quite common. As a mere sampling, Spanish features *partido* (‘game’ or ‘(political) party’), *gato* (‘cat’ or ‘(hydraulic) jack’), and *presa* (‘prey’ or ‘dam’). A small offering of the many cases of synonymia in English, each with at least three meanings, includes ‘date’ (a day on the calendar, a romantic outing, or a fruit; Spanish: *fecha, cita, dátíl*), ‘party’ (a social gathering, a group of people seated together at a restaurant, or a political organization; Spanish: *fiesta, grupo, partido*), and ‘spring’ (a season of the year, a metal coil, or a place where water emerges from the earth; Spanish: *primavera, resorte, manantial*). Yet another example in English (Poibeau, 2017: 171), processed into Spanish, will again demonstrate the possible pitfalls related to such lexemes:

- (17) ‘Little John was looking for his toy box. Finally, he found it. The box was in the pen. John was very happy.’

**GT (9 Oct 2021):**

*El pequeño John estaba buscando su caja de juguetes. Finalmente, lo encontró. La caja estaba en el bolígrafo. John estaba muy feliz.*

The GT rendition of example (17) is illogical to the point of being essentially impossible. The Spanish term *bolígrafo* is a common one for ‘pen’ when it denotes a writing instrument (specifically a ‘ballpoint pen’). Since it is not feasible in any reasonable way for a box of toys to fit inside a ballpoint pen, the ‘pen’ in question surely refers to a child’s playpen (*corralito*), or a pen in which animals are perhaps kept (*corral*), or a similar area of confinement. The obvious problem is that GT, which translates at the sentence level, neither has enough context to know what type of ‘pen’ might be involved nor realizes that what it has proposed is a physical impossibility. This is because a computer system trained to detect patterns does only that and does not “realize” anything the way humans do; it attempts natural language processing without the aid of “natural” (logically intuitive) capacity. Barring some unlikely paradigm shift in this regard, the type of

mistake made by GT in this example seems rather insuperable.

The treatment of homonymia above serves as a segue into a distinct but related phenomenon: polysemy, in which a single word can express partially related, even somewhat overlapping concepts, each of which may have a separate term in another language. For instance, when one uses the English verb ‘to save,’ which in general means to keep something or someone from being harmed or lost, its different specific connotations and appropriate verb translations into Spanish include, to note a few examples, ‘to save a living thing’ (*salvar*) ‘to save money’ (*ahorrar*), or ‘to save a computer file’ or ‘to put something away for safekeeping’ (*guardar*). Without context, GT is incapable of knowing which verb to use when translating from English to Spanish. Speaking of an amount of money (*ahorrar*), or a piece of pizza (*guardar*), one might exclaim: ‘I want to save it,’ which GT renders as *Quiero salvarlo* (as if the speaker wanted to save a puppy). If however, one states, ‘I want to save this piece of pizza,’ a correct translation is given: *Quiero guardar este trozo de pizza* (10 Jun 2022).

While the cases just discussed were created as test samples by the author, the following example, which also involves polysemy, comes from transcribed dialogue in the Spanish sitcom *Aquí no hay quien viva* (Miramón Mendi, 2003). Speaking of the need to use the stairs to ascend to the apartment units above, as the elevator is old and only supposed to be used for descending, one of the actors states:

- (18) ...solo lo utilizamos para **bajar**.  
Se **estropea** mucho.

GT (2 Oct 2021):

‘...we only use it **to download**. It **spoils** a lot.’

Author’s translation:

‘...we only use it to **go down**. It **breaks down** a lot.’

While the speaker did mention the elevator, it was in an earlier sentence, leaving GT, which works at the sentence level, to guess at the intended meaning. In Spanish, the verb *bajar* means not only ‘**to descend**,’ but also to ‘**download**’ (a computer file, for instance). Likewise, *estropear* (perhaps employed more frequently in Spain than the common Latin

American equivalent *dañarse*) can refer to something – organic or inorganic – being damaged, but it is general enough that English requires different specific verbs in translation in order to capture the precise meaning, depending on the context. English speakers may well say that a head of lettuce ‘**spoils**,’ but not an elevator, which ‘**breaks**’ or ‘**breaks down**.’ If one manipulates the original sentences, intentionally joining them and repeating the word *ascensor* (‘elevator’) explicitly – which of course the proficient human translator does not require – then a correct, idiomatic sentence results:

*Solo utilizamos el ascensor para bajar porque se estropea mucho.*

GT (26 Mar 2022):

‘We only use the elevator to go down because it breaks down a lot.’

Two final sets of examples will be given, one in which Spanish differs from English (and perhaps other languages), and the other in which it is English that features the marked construction and is prone to causing erroneous GT renditions. The first example, taken from Bolivia’s *El Espectador* newspaper, concerns time-related references in Spanish:

- (19) *Los tres países afinan detalles para firmar el acuerdo que pondrá en marcha el proyecto para verificar la destrucción de cocales, anunciado **en marzo pasado**.*

GT (1 Apr 2022):

The three countries are fine-tuning details to sign the agreement that will launch the project to verify the destruction of coca crops, announced **last March**.

The otherwise impressive GT rendition into English only becomes problematic at the end of the sentence. The article in question is dated 19 Apr 2011, and the March that is mentioned is the previous month, literally the ‘last’ one to transpire, which is precisely how *pasado* is used in Spanish. In English, however, the correct translation is ‘in March of this year,’ or, in this specific case, ‘last month’ would also suffice. This same issue of temporal orientation can exist

when a Spanish-language text refers to a future date, as seen in the following text from Argentina's *Cronista* newspaper:

- (20) *El **próximo jueves** 24 de marzo se recuerda a las víctimas de la última dictadura con el Día Nacional de la Memoria por la Verdad y la Justicia.*

**GT (1 Apr 2022):**

**Next Thursday**, March 24, the victims of the last dictatorship are remembered with the National Day of Memory for Truth and Justice.

The article in question appeared on Tuesday (22 Mar 2022), which means that the 'next Thursday' would indeed technically be on day 24 of the month. However, the date in question would be accurately expressed in English as 'this Thursday,' or 'Thursday of this week,' since 'next Thursday' would be used to designate Thursday, 31 March, a full week later.

However, it is English that at times poses a unique challenge to GT in its use of the modal verb 'should' to convey not the semantic conditional, but rather the habitual past, such as in this following mini-dialogue of the author's creation:

- (21) Person 1: 'What **would you do** on the weekends?'
- Person 2: 'We **would go** to the beach.'

**GT (5 Mar 2022):**

Person 1: ¿Qué **harías** los fines de semana?

Person 2: **Iríamos** a la playa.

Regarding the translation for Person 1, such a phrase in Spanish would only be used literally, such as in the following sentence expressing a hypothetical: ¿Qué **harías** los fines de semana **si tuvieras más tiempo y dinero?** 'What **would you do** on the weekends **if you had** more time and money?' Of course there is more than one way to express many if not most ideas. For instance, each person in the dialogue could have recast their part thus: 'What did you **use to do** on the weekends?'; 'We **used to go** to the beach.' Just as with the

original phrases, Spanish uses the imperfect aspect of the past tense to accomplish this – *hacías* and *íbamos* or *solías hacer* and *solíamos ir* – never the conditional. A lack of context, however, can cause GT to opt for a literal translation of 'should,' changing the intended meaning entirely.

#### 4 Conclusions

This paper is part of a larger, long-term study whose central focus is the ability (or inability) of Google Translate (GT) to render acceptable translations among multiple written genres between English and Spanish and vice versa. Some of the challenges relating to this pair of languages extend to others. While many GT capabilities have been greatly enhanced since the service's 2016 shift from the use of statistical machine translation (SMT) to a system of neural machine translation (NMT), this does not mean that all such renditions are perfect or even acceptable, or that its performance based on the perceived complexity or simplicity of source texts is predictable. Some economic texts, for instance, are rather intricate, but GT more often than not produces very usable English-Spanish-English results. In contrast, some seemingly simple texts are badly distorted when run through GT, even when no ambiguity is readily apparent. This is surely a reflection of the fact that even apparently uncomplicated human language is more involved than its speakers typically realize. Added to this is the fact that even relatively simple ideas can be expressed in such a variety of ways that no database could contain all the possibilities, let alone adequate past translations of them. This means that any solution to mistranslations could either be years, even decades away, or, surely in some cases, never be attainable at all, signifying that human translators will need to continue occupying an indispensable role in the translation process for the foreseeable future. While ascertaining some of these matters to the degree possible is the objective of the larger study, various patterns have already begun to suggest themselves and have been demonstrated to a modest degree in this paper via several examples featuring Spanish, English, and Bulgarian.

## References

- El Cronista*. 22 Mar 2022. 'Próximo Feriado del 24 de marzo: ¿es puente, hay fin de semana largo?' Retrieved 15 March 2022.
- El Espectador* (EFE). 19 April 2011. 'Bolivia acepta ayuda económica de EE.UU. para destruir la coca.' Retrieved 30 March 2022.
- 'Érase una mudanza.' *Aquí no hay quien viva*. Season 1, episode 1, Miramón Mendi S.L., 2003.
- Fitzgerald, Maggie and Stevens, Pippa. 10 Aug 2021. 'Dow rises 220 points to new record after inflation report is not as bad as feared.' *cnn.com*. Retrieved 4 Oct 2021.
- Garavalova, Iliyana. 2020. 'The Bulgarian dialect names of the cat.' *Papers of BAS. Humanities and Social Sciences*, 7(2), pp. 103-116.
- Koehn, Philipp. 2020. *Neural Machine Translation*. Cambridge, UK: Cambridge University Press.
- Leafgren, John. 2011. *A Concise Bulgarian Grammar*. Durham, NC: Slavic and Eurasian Language Resource Center.
- Lewis-Kraus, Gideon. 14 Dec 2016. 'The Great A.I. Awakening.' *The New York Times*.
- Piqué, Elisabetta. 2014. *Francisco: vida y revolución: Una biografía de Jorge Bergoglio*. Chicago: Loyola Press.
- Poibeau, Thierry. 2017. *Machine Translation*. Cambridge, MA: MIT Press.

# A Myriad of Ways to Say: “Wear a mask!”

**Cvetana Krstev**

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

**Duško Vitas**

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

## Abstract

This paper presents a small corpus of notices displayed at entrances of various Belgrade public premises asking those who enter to wear a mask. We analyze the various aspects of these notices: their physical appearance, script, lexica, syntax and style. A special attention is paid to various obligatory and optional parts of these notices. Obligatory parts deal with wearing masks, keeping the distance, limiting the number of persons on premises and using disinfection. We developed local grammars for modelling phrases that require wearing masks, that can be used both for recognition and for generation of paraphrases.

**Keywords:** short messages, local grammars, phrase generation, paraphrasing.

## 1 Introduction

Short messages have been attracting attention of linguists and researchers in natural language processing (NLP) for some time. One of the reasons is that it is a widespread type of communication, and is not limited to use among young people and entertainment. It has been noticed that short messages use a specific language and a particular style. For that reason, numerous corpora of short messages have been collected and can be explored by means of NLP tools. A corpus of 88,000 French SMS messages was collected, anonymized and made available for research purposes (Panckhurst, 2017). Petrović et al. (2010) presented a large Twitter corpus of 97 million posts and made it publicly available for researchers working in social media, NLP and large-scale data processing.<sup>1</sup>

In many cases researchers tailored their own corpora of short messages suiting their purposes. Bernicot et al. (2012) compiled a corpus of 864

<sup>1</sup>The corpus is no longer available due to change in Twitter policy.

SMS messages produced by French-speaking adolescents in order to analyze the effects of writers' characteristics on message length, dialogue structure, and message function. A corpus of French Twitter posts containing complaints regarding railway services was used to investigate linguistic directness and indirectness and differentiate them from perceived (im)politeness (Depraetere et al., 2021).

Graffiti on the walls of urban spaces are also a type of short messages; moreover, they have a much longer history than messages on today's social media. They have been analyzed from various perspectives: political, sociocultural, and linguistic (Alonso, 1998). The graffiti for analysis are often selected from a specific area, such as a university campus. Authors applied corpus method to analyse 378 graffiti found on walls of two Jordan universities (Al-Khawaldeh et al., 2017), and found that they express different themes: personal, social, national, religious, political etc.

A formulaic short messages are often sent on the occasion of Christmas and New Year. Christmas and New Year messages issued by important persons, like heads of state, which are far from being formulaic, have received more than their fair share of attention (Sauer, 2007). Still, some researchers were interested in the formulaic expressions: Deng et al. (2010) analysed how the language of Chinese SMS messages conveying Christmas wishes reflects a shift in cultural values and customs, while Włosowicz (2011) analysed how foreign language learners' mother tongue and cultural background influence their Birthday, Christmas and New Year's wishes.

Linguists showed interest in short messages written in Serbian as well. The use of shortening, clipping and elliptic constructions in text messages was analysed by several authors (Polovina and Jelić, 2020; Jelić and Vekarić, 2019). Graffiti



that emerged in Serbia in its transition era were analysed and messages were classified in overarching categories, as hate messages and love messages (Mršević, 2014). NLP specialists collected corpora of various types of short messages in order to solve different problems. Šandrih (2018) prepared a corpus of 5,500 Serbian SMS messages to test the system for detecting their sender. A corpus of 9,059 Serbian Twitter posts was collected in order to determine how their sentiment is affected by the use of negation (Ljajić and Marovac, 2019). Jokić et al. (2021) collected and manually annotated for hate-speech 6,436 tweets to be used for training hate-speech detection applications.

In this paper we are interested in notices that announce precaution measures related to the Covid-19 pandemic at front doors of public premises. These notices are similar to graffiti as they are public and are not a product of social media. However, contrary to graffiti, their content is restricted and in that respect they are closer to Christmas and New Year wishes.

Our paper is organized as follows. Section 2 presents a small corpus of notices related to precaution measures against Covid-19 virus. In Section 3 the lexica used in these notices is analysed. Sections that follow are restricted to the part of these notices that concern protective masks: their basic syntactic patterns (Section 4) and their semantic value (while in Section 5). Finally in Section 6 we show by generating mask messages of determined structures that there are myriad ways to say: “Wear a mask!”. In Section 7 we discuss avenues for future research.

## 2 About the Corpus

Our corpus is based on notices that were photographed between 21 January 2021 and 30 March 2022 and then re-typed. We considered as a single notice everything written on one sheet of paper. Not all notices were originally composed; instead, many were acquired from Internet and other sources, and used on entrances to many different facilities. We filtered only different notices from all notices photographed. We do not claim that our set of notices is in any way representative since all notices were photographed in the center of Belgrade at the walking distance from the place of the residence of the authors of this paper. The total number of photographed notices is 231.

**Physical appearance** – The majority of notices

were printed on a paper (207), 23 notices were handwritten. There was one 3D notice – a word “Obavezno” (obligatory) written on an actual mask.

**Capitalization** – The majority of notices (165) were written using only upper-case letters, 45 notices were written using lower-case letters, while in 21 notices only some parts were written using only upper-case letters for emphasis.

**Script** – The majority of notices (187) were written using Latin script. Among them 18 did not use diacritics. 44 (19%) of notices were written in Cyrillic script. Among 23 handwritten notices, 8 (34.7%) were written in Cyrillic. These findings are in line with the conclusions reached by Ivković (2013) that the Latin alphabet dominates over Cyrillic on Serbian news websites and the use of non-standard Latin orthographic variants (diacritics omission) is becoming stable.

**Emphasizing** – Various methods of emphasis were used. Many notices included images (of a mask, people maintaining a distance, etc.) – 68 (29.4%) such notices were in the selected set. Other means of emphasizing were: using bold font (18), underlining (23), colors (mostly red) (29), increasing the font size (13). In many cases more than one mean of emphasizing was used. The other means of emphasizing the message of the notice was the use of the exclamation mark. It was used in 52 notices, either only once (40) or repeatedly (two times, three times, and up to 16 times). It is interesting to note that only once an emoticon was used (a smiley).

**Multilinguality** – All notices were written in Serbian. However, some of them had translations in English, and they vary in form as much as those written in Serbian that we will explore in following sections. These 5 notices in English concerning the use of masks were: “*Please wear protective mask in public areas*”, “*No entry without face mask*”, “*use a protective mask*”, “*Face mask required*”, “*Please don’t enter without a face mask*”.

In some notices images were intertwined with words to convey the meaning. For instance, <img of a person> 5 <img of a mask> Hvala (Thank you). Such notices were excluded from further analysis, as well as notices or their parts written in English.

The length of notices is between one word (a word “Obavezno” (obligatory) written on an actual mask) and 84 words. The average length of notices is 14.3 words, while two thirds of them were

written using no more than 15 words.

Each notice mentioned one to four protective measures: (a) wearing a mask – 207 notices; (b) number of persons allowed on premises – 64 notices; (c) keeping the distance – 39 notices; (d) disinfection – 17 notices.

The majority of notices (157) listed only one protective measure, 49 notices listed two, 17 listed three, while 8 notices listed all four measures. The parts of notices concerning masks were the shortest (6.74 words), followed by disinfection (8.29 words), number of people on premises (8.39 words), and keeping a distance (8.74 words).

Besides parts of notices listing the protective measures, some of them have one or more additional parts. They are:

**Attracting attention and addressing those who enter** – This part was represented in 70 messages (30.3%). It is simple and did not vary much in form. To attract attention several words were used: *Obaveštenje* (Notice) (12), *Pažnja* (Attention) (9), *Važno* (Important) (2), *Stop* (1). Three forms were used to address those who enter premises: *Poštovani*, (Respected,) (11) – this is a very formal and impersonal form of address that is often used for written official communication with unknown persons; A slightly less formal *Poštovani kupci*, (Respected customers,) (30) where the most frequently used word *kupci* can be replaced with *potrošači* (consumers), *klijenti* (clients), *posetioci* (visitors), *gosti* (guests), *sladokusci* (gourmand); *Dragi kupci*, (Dear customers,) (5) where the word *kupci* can be replaced with some of the previously listed words – the more informal form of address, but still very polite. These findings lead us to the question of whether Serbian society still belongs to “solidarity cultures” of the East rather than “distance cultures” of the West (Schlund, 2014). In some notices parts for both attracting attention and addressing customers were used.

**Invoking authority and explanations** – This part occurred in 47 of selected notices (20%). It is rather long and without a strict form. It conveyed the reasons for the protective measures and/or who has prescribed them. The statement expressing reasons for prescribing necessary measures started usually with *zbog/usled pandemije...* (due to the pandemic...) or *u cilju/radi sprečavanja pandemije* (in order to prevent pandemics...). The statements invoking authorities started with *Po/Prema/Na osnovu/U skladu sa odlukom...* (According/On the

basis/In accordance with the decision...). The most frequently mentioned authority is the government of the Republic of Serbia (14), and besides it the Ministry of Health (1) and *Krizni štab* (Crisis Response Team) (1). In the cases when a specific authority is not mentioned, a particular decision or regulation published in *Službeni glasnik* (Official Gazette) is listed (3). In one case the precise article of the regulation is mentioned with no less than 13 issues of the Official Gazette. Both messages – authority and explanation – sometimes occur together: *Usled odluke Vlade Republike Srbije, a u cilju suzbijanja epidemije...* (Due to the decision of the Government of the Republic of Serbia, and in order to suppress the epidemic...). It is legitimate to ask which of the features mentioned by Njegovan et al. (2011): credibility, exclusiveness, uniqueness, omnipresence, validity that characterize bureaucratic authority, are most likely to appeal to customers to obey to precaution measures.

**Gratitude** – This part occurred in 51 of selected notices (21.6%). It is simple and does not vary much in form. The used expressions are: *Hvala* (Thanks) (23) – neutral, *Hvala Vam* (Thank you) (1) – a slightly more personal, *Hvala (Vam) (lepo)/Zahvaljujemo se na razumevanju* (Thanks/We thank (you) (nicely) for your understanding) (22) – apologetic (we apologize that you have to wear a mask and we thank you for understand it), *Hvala unapred/Unapred hvala* (Thanks in advance) (3) – appealing to customers’ conscience because they accept the gratitude before they have done what is asked of them. Finally, the form *Zahvaljujemo što poštujete navedenu meru* (We thank you for complying with this measure) was used twice. It is interesting to note that the form *Hvala lepo* (Thank (you) nicely) once popular in everyday communication was encountered only once.<sup>2</sup>

**Miscellaneous** – Occasionally some miscellaneous information was added to notices (38 cases). This information was sometimes completely unrelated to the precaution measures, e.g. working hours of a shop. In a number of cases some additional information is added to a certain precaution measure, like *Zadržavanje u radnji je do 10 minuta!* (Staying in the shop is up to 10 minutes!) or *Ukoliko nemate masku, dobićete je u knjižari...* (If you don’t have a mask, you will get it in the bookstore...). In some notices (24) additional statements

<sup>2</sup>This form has 54 occurrences in the SrpKor2013 (<http://www.korpus.matf.bg.ac.rs/>), and 4,622 in the SrpKor2021 (<https://noske.jerteh.rs>).

NOUN	Freq	VERB	Freq	ADJ	Freq	ADV	Freq
<i>maska</i>	207	<i>moliti</i>	48	<i>obavezan</i>	122	<i>obavezno</i>	25
mask		to request		mandatory		mandatorily	
<i>nošenje</i>	91	<i>dozvoliti</i>	27	<i>zaštitni</i>	84	<i>istovremeno</i>	16
wearing		to permit		protective		simultaneously	
<i>hvala</i>	50	<i>moći</i>	27	<i>poštovan</i>	40	<i>maksimalno</i>	7
gratitude		can		respected		maximally	
<i>ulazak</i>	49	<i>nositi</i>	25	<i>drugi</i>	16	<i>najmanje</i>	7
entering		to wear		other		at least	
<i>objekt</i>	39	<i>držati</i>	19	<i>maksimalan</i>	11	<i>najviše</i>	7
facility		to keep		maximal		at the most	
	<b>1272</b>		<b>291</b>		<b>417</b>		<b>82</b>

Table 1: The most frequent nouns, verbs, adjectives and adverbs in the Mask corpus

were used as encouragement for people to respect imposed measures: *Budimo odgovorni* (Let's be responsible) or *Čuvajmo sebe i druge* (Let's take care of ourselves and others).

**Signature** – 42 notices were signed by the facility which attached a notice. Even this part of selected notices was not completely uninteresting. More than half of the signed notices (22) reveal appealing foreign or foreign-like firm names: *Beauty and the beast center*, *Beomelody d.o.o.*, *Ušće Shopping center*.

Serbian has a binary pronominal system of address which employs one pronoun (second person singular – *Ti*) for familiar address and another (second person plural – *Vi*) for formal address. Although studies (Milosavljević, 2018) have shown that the informal address has been gaining attraction in everyday communication over the last couple of decades, e.g. in media were a show host addresses a guest, in our notices a familiar address using the *Ti* pronoun was never used. Some statements are ambiguous, such as *Molimo vas da se pridržavate mera zaštite radi sprečavanja širenja zaraze koronavirusom* (Please adhere to protection measures to prevent the spread of coronavirus infection), which can refer both to one person addressed by the *Vi* pronoun and to a group of people. However, a statement like *Obavezno nosi<sub>sing</sub> masku<sub>sing</sub>* (Be sure to wear a mask) was not found in our selection of notices.

### 3 Lexica used in notices

When preparing our tiny little corpus the notices were typed as they were and corrected only evident typos, leaving grammatical and orthographic errors. The corpus consists of 3,581 tokens and

3,285 words. Among words, there were 193 different nouns, 49 different verbs, 57 different adjectives and 15 different adverbs. The five most frequently used nouns, verbs, adjectives and nouns are listed in Table 1.

In our sample specific groups of nouns were identified. The first group containing 34 nouns were used to refer to places to which notices about protective measures apply. Here one can distinguish the most general concepts: *objekat* (facility) (39), *prostor* (space, area) (13), *prostorija* (room) (9), *zgrada* (building) (2), *mesto* (place) (1). These concepts can be further qualified: *prodajni prostor* (shopping area), *poslovni prostor* (business area), *javni prostor* (public area), *zatvoreni prostor* (enclosed area) *prodajni objekat* (shopping facility), *maloprodajni objekat* (retail shopping facility), *radna prostorija* (working space), *javno mesto* (public place). The more specific concept was represented by three nearly synonymous words *radnja* (32) *prodavnica* (12) (shop), *maloprodaja* (retail) (1). The remaining 26 nouns were used to name a place for a specific activity, like *menjačnica* (exchange office), *pekara* (bakery), *fakultet* (faculty).

The other group of specific nouns refers to people to whom notices are addressed. Here also very general concepts were used, lexicalized by *osoba* (39) and *lice* (8) (person), *gradjanin* (1) (citizen), *ljudi* (people) (4). Besides them, more specific concepts were used for potential: *kupac* (shopper) (28), *potrošač* (consumer) (14), *mušterija* (customer) (3), while other concepts were specific to particular activity, like *gost* (guest) (1) and *posetilac* (visitor) (7).

It should be noted that concepts referred to by these two specific groups of nouns are related in

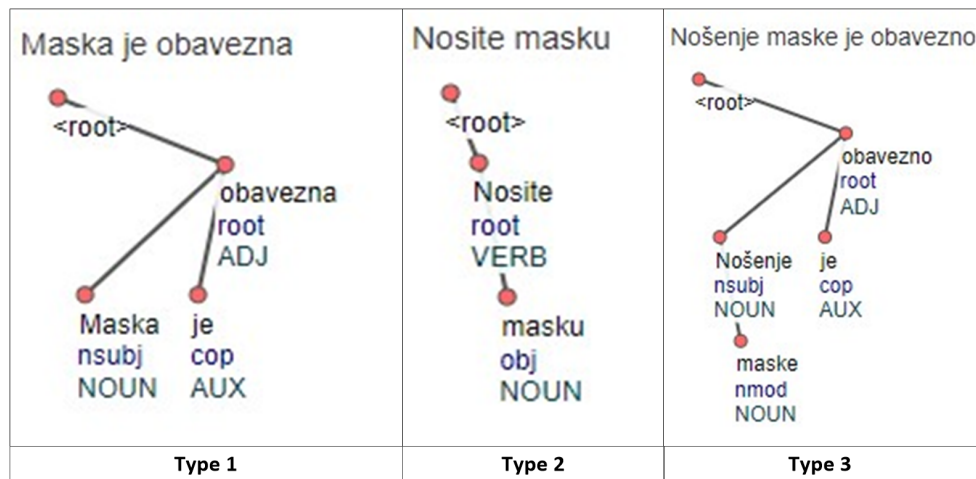


Figure 1: The basic syntactic structures of mask messages.

a specific way. *Osoba* and *lice* (person) are applicable to all types of premises, while visitors of *biblioteka* (library) and *sala* (hall) cannot be designated as *kupac*, *potrošač* or *mušterija*. Moreover, even though *mušterija*, *kupac* and *potrošač* can all be translated as ‘customer’, only *mušterija* is applicable to *menjačnica* (exchange office) and *frizerski salon* (hairdresser), while *kupac* and *potrošač* are not. On the other hand, the library patrons are customarily designated as *posetilac* (visitor) or *član* (member), while *mušterija* cannot be used. The larger and more versatile corpus is needed to fully investigate relations between designations for types of visitors and types of venues.

Finally, even in this small corpus we have ambiguity: *lice* can mean both face and person. In the former meaning it was used 9 times, e.g. *maska za lice* (face mask), while in the later case it was used 8 times.

#### 4 The Analysis of Mask Messages

In this section we will present basic syntactic patterns of one section of selected notices, namely the one related to wearing masks. There were 207 such statements, with a total of 1,466 tokens, and 1,365 words. The most frequent of 72 different nouns is *maska* (204), followed by *nošenje* (wearing) (91) and *ulazak* (entering) (37). The most frequent of 22 different verbs is *moliti* (to request) (29), followed by *nositi* (to wear) (25) and *staviti* (to put) (10). The most frequent of 21 different adjectives is *obavezan* (mandatory) (124), *zaštitni* (protective) (83) and *zatvoren* (enclosed) (8). In total, 8 different adverbs have been identified, of which only 2 occurred more than once: *obavezno* (mandatorily)

(14) and *isključivo* (exclusively) (2).

The most frequent noun *maska* is often characterized as *zaštitna maska* (protective mask), *higijenska maska* (hygienic mask) and *maska za lice* (face mask), and can be additionally described as *ličan* (personal). In the descriptions below [maska] stands for all these possibilities. Also, in this context *korišćenje* and *upotreba* (usage) are treated as synonyms for *nošenje* ([korišćenje]); similarly, synonym of *nositi* (to wear) is *koristiti* (to use), while *staviti* (to put) is also used in a similar context ([koristiti]).

As for their syntactic patterns, the majority of mask messages has one of 9 general forms discussed below. Their basic syntactic patterns were analyzed using the UDPipe (Straka and Straková, 2017) and they are presented in figures 1–3.<sup>3</sup> All these basic sentences can be modified with additional phrases: *Molimo Vas...* (Please...), *U ovom objektu...* (On this premises), *Pri ulasku u ...* (On entering in...), and *Svi kupci* (All customers...). For the recognition of these basic patterns with their various realizations we developed within Unitex/Gramlab<sup>4</sup> local grammars that are supported by Serbian morphological dictionaries (Stanković et al., 2021).

**1. [maska] je [obavezna]** – or “Mask is mandatory”. In this statement *maska* and *obavezan* can be replaced with synonyms (see Section 3), and the whole statement can be in plural: *maske su obavezne* (masks are mandatory). The auxiliary is sometimes omitted. This form was adopted by 25

<sup>3</sup>We analysed basic sentences using (Straka, 2020)

<sup>4</sup>Unitex/Gramlab – the Multilingual Corpus Processing Suite ([unitexgramlab.org](http://unitexgramlab.org))



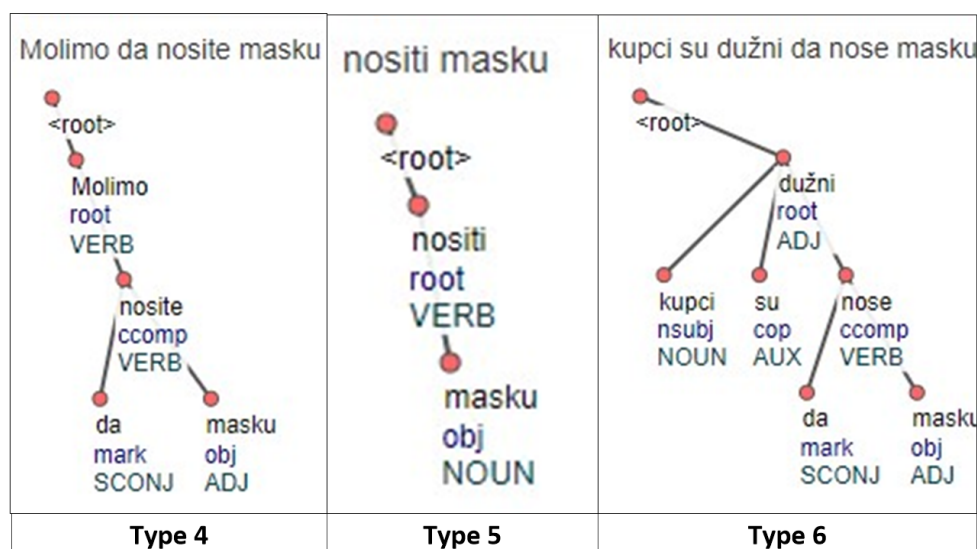


Figure 2: The basic syntactic structures of mask messages (continuation).

notices, two of which are: *Molimo Vas - maske su obavezne pri ulasku* (Please - masks are required upon entry) and *U lokalu obavezne zaštitne maske* (In premises protective masks mandatory).

**2. [nošenje] [maske] je [obavezno]** – or “Wearing a mask is mandatory”. Notices with this form occurred 99 times, and two of them are: *Nošenje maske u poslastičarnici je obavezno* (Wearing a mask in the pastry shop is mandatory) and *tokom boravka na fakultetu neophodno je nošenje maske* (it is necessary to wear a mask during your stay at the faculty premises). No message of this type was used with a phrase *Molimo Vas...* (Please...).

**3. [nosite] [masku]** – or “Wear a mask”. In this case verb *nositi* is in the imperative mood, 2nd person plural. There were 19 notices of this form, and two of them are: *Molimo, stavite masku pre ulaska u apoteku* (Please put on the mask before entering the pharmacy) and *Pri ulasku u radnju obavezno koristite masku* (Be sure to use a mask when entering the store).

**4. [Molimo da] [nosite] [masku]** – or “We entreat you to wear a mask”. In this case verb *nositi* is in the present tense, 2nd person plural. The structure remains the same if *Molimo da...* (We entreat you to...) is replaced by *Hvala što...* (Thank you for...). There were 15 notices of this type: *Hvala što nosite masku* (Thanks for wearing a mask) and *Molimo Vas da prilikom ulaska u prodajni prostor nosite zaštitnu masku* (Please when entering the shopping area wear a protective mask).

**5. [nositi] [masku]** – or “to wear a mask”. In this case verb *nositi* is in the infinitive. There were 2 notices of this type: *Masku staviti pre ulaska u agenciju* (Put on the mask before entering the agency) and *Obavezno koristiti zaštitnu masku* ((It is) Obligatory to use a protective mask).

**6. [kupci] su [dužni] da [nose] [masku]** – or “customers are required to wear a mask”. Here [dužni]={dužni, obavezni}, while [kupci] stands for all types of persons entering premises. In this case verb *nositi* is in the present tense, 3rd person plural. There were 5 notices of this type: *Kupci su obavezni da imaju zaštitnu masku* (Customers are required to have a protective mask) and *Sva lica dužna su da pri ulasku u objekat nose masku* (All persons are required to wear a mask when entering the facility).

**7. zabranjen je [ulaz] bez [maske]** – or “entry without a mask is prohibited”. Here [ulaz]={ulaz, ulazak, dolazak}. The auxiliary can be omitted. There were 9 notices of this type: *Strogo zabranjen ulaz bez maske* (entry without a mask is strictly prohibited) and *Ulaz u maloprodaju je zabranjen licima bez zaštitne maske!* (Entry into retail store is prohibited to persons without a protective mask). A variant of this structure is negated: **nije dozvoljen [ulaz] bez [maske]**, one of 4 retrieved examples is: *U knjižaru nije dozvoljen ulaz bez maske* (It is not allowed to enter the bookstore without a mask).

**8. [ulaz] je dozvoljen sa [maskom]** – or “entry allowed with a mask”. [ulaz] has the same values as before. The auxiliary can be omitted. There were

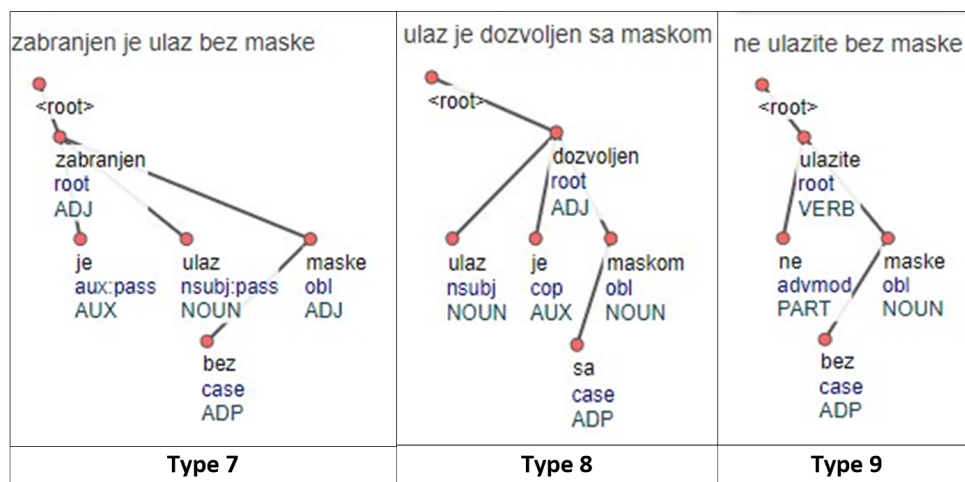


Figure 3: The basic syntactic structures of mask messages (continuation).

3 notices of this type: *ulazak je dozvoljen samo osobama sa zaštitnom maskom* (entry is allowed only to persons with a mask).

9. There were 7 notices that use the verb *ulaziti* (to enter). When used without negation these notices are similar to 4 – **Molimo da ulazite sa [maskom]** (Please enter with a mask). An example is *Molimo Vas da u galeriju ulazite s maskom* (Please enter the gallery with a mask). The form with the negation is **ne ulazite bez [maske]** (do not enter without a mask) where verb is in the imperative mood, 2nd person plural. An example is: *u radnju ne ulazite bez zaštitne maske* (Do not enter the shop without a protective mask). There were also examples with the verb in the infinitive (similar to 5) and in the 3rd person plural (similar to 6).

These 9 basic syntactic patterns describe 188 out of 207 messages about wearing a mask. Some of the remaining notices either do not use the word *maska* but a more general term, e.g. *Obavezno koristite mere zaštite od koronavirusa* (Be sure to use coronavirus protection measures), use some shortened expressions like *nošenje maske* (wearing a mask), or some specific form *Bez maske ne primamo u objekat* (We shall deny entry into the facility (to those) without a mask).

### 5 Hidden messages and sentiment values

Although the parts of notices related to mask wearing are very short, less than 7 words on average, some of them convey additional meaning to the main one, which is that one is requested to wear a mask. For instance, in two cases the statement contained *obavezno nošenje maski preko nosa i*

*usta* (mandatory to wear a mask over the nose and mouth) meaning that it is not enough to wear a mask, but also to wear it properly. Moreover, if notice said *ulaz u objekt s maskom* (entry in the facility (allowed only) with a mask), it would be possible to interpret it as if one was required to wear a mask only while entering the facility, but not throughout the visit. Therefore, some notices added explicit clarification: *Nije dozvoljen ulazak odnosno boravak lica bez ličnih zaštitnih maski* (It is not allowed for persons to enter or stay without personal protective masks).

The tone of messages related to masks vary across our corpus, and goes from very polite to severe or even unpleasant. Notes starting with *Molimo da...* or *Hvala Vam što...* always have a positive tone. However, some expressions like *Zabranjeno je...* (It is forbidden...) are never used with “please”, and cases like that were not found in our set. Also, a neutral expression like *Nošenje maske je obavezno* (Wearing a mask is obligatory) can be made more severe by adding *bez izuzetka* (without exceptions) or similar. The level of severity of mask notices according to their lexica and syntax can be ranked in the following way strating from those most strict:

- Messages using *zabranjeno je* (*Zabranjen ulazak bez maske* – Entry without a mask is forbidden);
- messages using imperative mood and/or negation (*Stavite masku* – Put a mask, *Ne ulazite bez maske* – Do not enter without mask);
- neutral messages (*Maska je obavezna* – Mask is obligatory);



- “Please” with the imperative mood and/or negation (*Molimo, stavite masku* – Please put a mask, *Molimo da ne ulazite bez maske* – Please, do not enter without a mask);
- “Please” addressing persons indirectly (*Molimo kupce da nose masku* – We ask customers to wear a mask);
- “Please” addressing persons directly (*Molimo Vas da nosite masku* – We ask you to wear a mask).

There is a Serbian proverb *Lepa reč gvozdена vrata otvara* (A nice word opens an iron gate). In this case, we cannot say which announcements, polite or strict, appealed to people more to respect the measures.

## 6 Generating mask messages

We used the local grammars developed for the recognition of 9 basic patterns with their various realizations that were systematized on the basis of data found in our corpus to generate possible mask messages within Unitex. In order to avoid an excessive number of possibilities, synonyms retrieved in our set, sometimes also hiponyms and hiperonyms, designating concepts *maska, kupac/osoba, objekat/radnja, ulazak* were not used for generation. By doing so we obtained:

**Type 1** – 19,520 messages, for example: *u ovoj prostoriji maska preko nosa i usta je obavezna* (in this room a mask over a nose and a mouth is obligatory);

**Type 2** – 10,944 messages, for example: *upotreba maski na licu u ovom objektu je obavezna* (the use of masks over the face in this facility is obligatory);

**Type 3** – 110,784 messages, for example: *molimo, stavite masku preko nosa i usta obavezno ako ulazite u ovu radnju* (please, put a mask over your nose and mouth obligatorily if you enter this shop);

**Type 4** – 62,160 messages, for example: *molimo da imate vašu masku na licu pri ulasku i za vreme boravka u ovoj prostoriji* (we ask you to have your mask on (your) face when entering and during (your) stay in this room);

**Type 5** – 14,240 messages, for example: *u objektu obavezno morate nositi masku* (in the facility you must be sure to wear a mask);

**Type 6** – 35,328 messages, for example: *mole se svi kupci da prilikom ulaska u ovu radnju obavezno stave masku* (all customers are asked to be sure to put a mask when entering this shop);

**Type 7** – 6,544 messages, for example: *ulazak nije dozvoljen osobama bez maske* (entering is not allowed to persons without a mask);

**Type 8** – 11,520 messages, for example: *dozvoljen je ulazak isključivo osobama sa maskom na licu* (only persons with a mask on (their) face are allowed to enter);

**Type 9** – 19,710 messages, for example *mole se kupci da ne ulaze u ovaj objekat bez maske* (the customers are asked not to enter this facility without a mask).

In a total, we produced 290,750 ready to use, correct messages all conveying the same basic meaning: “wear a mask”.

## 7 Conclusion

In this paper we presented the analysis of a set of notices collected from front doors of various premises that require compliance with protective measures against Covid-19. We analysed lexic and syntactic patterns of mask notices in more detail which enabled us to generate notices featuring one of their basic structures.

Our next step will be the production of paraphrased sentences with a full morphosyntactic description. Besides that we will analyse in the similar way messages about other Covid-19 protective measures. Moreover, we will collect other public announcements that emerge spontaneously and convey the similar meaning, like *Zatvarajte vrata za sobom* (close the door behind you) or *Ne primamo reklame* (We do not accept advertisements). The goal of our future project is to produce a big and versatile set of paraphrases.

## References

- Nisreen Naji Al-Khawaldeh, Imad Khawaldeh, Baker Bani-Khair, and Amal Al-Khawaldeh. 2017. An exploration of graffiti on university’s walls: A corpus-based discourse analysis study. *Indonesian Journal of Applied Linguistics*, 7(1):29–42.

- Alex Alonso. 1998. Urban graffiti on the city landscape. *San Diego State University*.
- Josie Bernicot, Olga Volckaert-Legrier, Antonine Goumi, and Alain Bert-Erboul. 2012. Forms and functions of SMS messages: A study of variations in a corpus written by adolescents. *Journal of Pragmatics*, 44(12):1701–1715.
- Jing Deng et al. 2010. Texting Christmas wishes in China: A view from pragmatics. *Bucharest Working Papers in Linguistics*, (2):115–127.
- Ilse Depraetere, Sofie Decock, and Nicolas Ruytenbeek. 2021. Linguistic (in)directness in Twitter complaints: A contrastive analysis of railway complaint interactions. *Journal of Pragmatics*, 171:215–233.
- Dejan Ivković. 2013. Pragmatics meets ideology: Digraphia and non-standard orthographic practices in Serbian online news forums. *Journal of Language and Politics*, 12(3):335–356.
- Gordana Jelić and Gordana Vekarić. 2019. Elliptical Constructions in SMS Communication. In *Značenje u jeziku – Od individualnog do kolektivnog. Zbornik radova sa međunarodnoga znanstvenog skupa Hrvatskoga društva za primenjenu lingvistiku održanoga od 16. do 18. svibnja 2019. u Rijeci*, pages 89–102.
- Danka Jokić, Ranka Stanković, Cvetana Krstev, and Branislava Šandrih. 2021. A Twitter Corpus and lexicon for abusive speech detection in Serbian. In *3rd Conference on Language, Data and Knowledge (LDK 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
- Adela Ljajić and Ulfeta Marovac. 2019. Improving sentiment analysis for Twitter data by handling negation rules in the Serbian language. *Computer Science and Information Systems*, 16(1):289–311.
- Marija S Milosavljević. 2018. Addressing the interlocutor in informative programmes. *Reči (Beograd)*, 10(1):15–29.
- Zorica Mršević. 2014. The beauty of graffiti. <http://zoricamrsevic.in.rs/english/GraffitiBeautyMrsevic4.pdf>.
- Biljana Ratković Njegovan, Maja Vukadinović, and Leposava Grubić Nešić. 2011. Characteristics and types of authority: the attitudes of young people. a case study. *Sociológia*, 43(6):657–673.
- Rachel Panckhurst. 2017. A digital corpus resource of authentic anonymized French text messages: 88milSMS—What about transcoding and linguistic annotation? *Digital Scholarship in the Humanities*, 32(suppl.1):i92–i102.
- Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. The Edinburgh Twitter corpus. In *Proceedings of the NAACL HLT 2010 workshop on computational linguistics in a world of social media*, pages 25–26.
- Vesna Polovina and Gordana Jelić. 2020. Shortening and Clipping in Serbian Text Messaging. *Anali Filološkog fakulteta*, 32(2):321–343. 18.
- Branislava Šandrih. 2018. Fingerprints in SMS Messages: Automatic Recognition of a Short Message Sender Using Gradient Boosting. In *3rd International Conference Computational Linguistics in Bulgaria (CLIB 2018)*. Department of Computational Linguistics at the Institute for Bulgarian Language with the Bulgarian Academy of Sciences: Sofia, Bulgaria, pages 203–210.
- Christoph Sauer. 2007. Christmas messages by heads of state. *Pragmatics & Beyond New Series (P&BNS)*, page 227.
- Katrin Schlund. 2014. Aspects of linguistic politeness in Serbian. A data-based comparison with German. *Linguistik online*, 69(7).
- Ranka Stanković, Cvetana Krstev, Rada Stijović, Mirjana Gočanin, and Mihailo Škorić. 2021. Towards Automatic Definition Extraction for Serbian. In *Proceedings of the XIX EURALEX Congress of the European Association for Lexicography: Lexicography for Inclusion (Volume 2)*, pages 695–704. Democritus University of Thrace.
- Milan Straka. 2020. UDPipe Croatian: Morphosyntactic Analysis of Raw Text. <https://live.european-language-grid.eu/catalogue/tool-service/437>.
- Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.
- Teresa Maria Włosowicz. 2011. Ways of expressing birthday, Christmas and New Year’s and Easter wishes in L2 and L3: Cross-cultural transfer and interlanguage pragmatics. In *Aspects of culture in second language acquisition and foreign language learning*, pages 217–231. Springer.

# Image Models for large-scale Object Detection and Classification

**Jordan Kralev**

Technical University, Sofia  
jkrulev@tu-sofia.bg

**Svetla Koeva**

Institute for Bulgarian Language, BAS  
svetla@ddcl.bas.bg

## Abstract

Recent developments in computer vision applications that are based on machine learning models allow real-time object detection, segmentation and captioning in image or video streams. The paper presents the development of an extension of the 80 COCO categories into a novel ontology with more than 700 classes covering 130 thematic subdomains related to Sport, Transport, Arts and Security. The development of an image dataset of object segmentation was accelerated by machine learning for automatic generation of objects' boundaries and classes. The Multilingual image dataset contains over 20,000 images and 200,000 annotations. It was used to pre-train 130 models for object detection and classification. We show the established approach for the development of the new models and their integration into an application and evaluation framework.

**Keywords:** image dataset, image models, object detection, object classification

## 1 Introduction

The shift of traditional data fusion methods challenged by multimodal big data motivates the creation of a new image corpus, the Multilingual Image Corpus, which is characterised by carefully selected images that illustrate thematically related domains and precise manual annotation for the segmentation and classification of objects in the images.

Recent developments in computer vision applications that are based on machine learning models allow real-time object detection, segmentation and captioning in image or video streams (Kasapbaşı et al., 2022; Cameron et al., 2019). We developed an image processing pipeline for object detection and object segmentation using pre-trained models. We also delivered a reliable service for automatic annotation of objects in images using advanced

deep learning techniques (Michelucci, 2019) and some existing tools and machine learning frameworks.

The paper presents the development of an extension of the 80 COCO categories into a novel ontology with more than 700 classes covering 130 thematic subdomains related to Sport, Transport, Arts and Security (presented in Section 2). The development of an image dataset for object segmentation and classification (The Multilingual image Corpus, MIC21) was accelerated by machine learning for automatic generation of objects' boundaries and classes (presented in Section 3). The MIC21 image dataset containing more than 20,000 images and 200,000 annotations was used to pre-train 130 models for object detection and classification. We show the accepted approach for the development of machine learning models and their integration into a framework for the evaluation and running of models (in Section 4).

In other words, we will demonstrate the application of models for the prediction of object outlines and classes in images as part of the development of the Multilingual Image Corpus, and then we will show how the new dataset, in turn, can be used to pre-train existing models so that they predict large number of object classes.

## 2 Multilingual Image Corpus in brief

The Multilingual Image Corpus offers data to train models specialized in object identification, segmentation and classification by providing fully annotated objects within images with segmentation masks categorised according to an Ontology of Visual Objects. The Multilingual Image Corpus is distinguished by the following key features: a) large image collection containing thousands of images and annotations; b) an Ontology of visual objects specifically created for object classification;

c) preparatory automatic object segmentation and classification evaluated by experts; d) translation of object classes and attaching definitions of concepts in 25 languages.

The dataset contains images from 4 thematic domains (Sport, Transport, Arts and Security), which represent highly related objects such as Tennis player and Soccer player, Limousine and Taxi, Singer and Violinist, Fire engine and Police boat grouped in 130 subsets of images. The images in the dataset are collected from a range of repositories offering API: Wikimedia, Pexels, Flickr, Pixabay, Creative Commons Search. Each image is equipped with a metadata description in JSON format. The metadata include fields such as: the name of the sub-dataset, sub-dataset id, image author, author’s web address, image original size, file name, image license, image source, last access to the source, source’s web address, MIC21 project url, etc. (Koeva et al., 2022)

The selected classes for annotation are organized into an Ontology of visual objects (Koeva, 2021). The Ontology consist of 706 classes that describe visual objects, 147 classes that represent their hypernyms, 14 relations between concepts and axioms that make explicit claims about the relations between concept classes. The Ontology classes correspond (but are not limited) to WordNet concepts (Fellbaum, 1999; Miller et al., 1990) which can be represented by visual objects (almost half of the Ontology classes are not contained in the WordNet). Two of the relations and their properties are also inherited from WordNet.

For example, the dominant class **Accordionist** is represented in WordNet, while the dominant class **Handball player** – not. For new dominant classes the appropriate hypernym in the WordNet structure is determined, in this case – **Athlete**. The attribute classes for Handball player are: Handball referee, Handball court, Handball, Handball goal, Handball jersey, Handball pants, Handball shorts, Handball shoe, Handball sock, Race number, Knee pad, from which only Handball and Knee pad are part of the WordNet. Ontology relations between the dominant class and its attribute classes link depicted relations between visual objects, for example: Handball player is next to Handball referee, Handball player plays at Handball court, Handball player plays with Handball and so on.

The use of the Ontology of visual objects ensures the selection of mutually exclusive classes,

the interconnectivity of classes by means of formal relations and an easy extension of the Ontology with more concepts corresponding to visual objects.

All Ontology classes have been translated into 25 languages using publicly available wordnets and BabelNet: English, Albanian, Bulgarian, Basque, Catalan, Croatian, Danish, Dutch, Galician, German, Greek, Finnish, French, Icelandic, Italian, Lithuanian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, and Swedish (Koeva et al., 2022).

An image processing pipeline for object detection and object segmentation was developed. Two software packages – Yolact (Bolya et al., 2019) and Detectron2 (Wu et al., 2019), and Fast R-CNN (Girshick, 2015) models trained on the COCO dataset were used for the generation of annotation proposals. The COCO format is a commonly used format for the instance segmentation representation (Sun et al., 2022; Amo-Boateng et al., 2022; Conrady et al., 2022; Cui et al., 2022).

The task for the annotators was to correct, reject or create new polygons for individual objects in the image and to classify the objects against the classes from the predefined Ontology. Table 1 displays the Multilingual Image Corpus’s current status.

Domain	Images	Annotations
Sport	6,915	65,482
Transport	7,710	78,172
Arts	3,854	24,217
Security	2,837	35,916
MIC21	<b>21,316</b>	<b>203,797</b>

Table 1: The Multilingual Image Corpus in Numbers

The metadata of images, Ontology, object annotations and multilingual descriptions of Ontology classes are available to be downloaded, copied, modified, distributed, displayed and used in accordance with the Creative Commons Attribution-ShareAlike 4.0 International License.<sup>123</sup>

The Multilingual image dataset can be implemented in: automatic identification and annotation of objects in images (a prerequisite for effective search of images and (within) video content), automatic annotation of images with short descriptions in European languages.

<sup>1</sup><https://doi.org/10.57771/belg-vm57>

<sup>2</sup><https://doi.org/10.57771/hxe0-4826>

<sup>3</sup><https://doi.org/10.57771/v36v-yb33>



We developed a framework based on FiftyOne, Yolact and Detectron2, and implemented it over Mask R-CNN on Python3, Keras and TensorFlow. We pre-trained Fast R-CNN models using the Detectron2 framework with ground truth annotations, which resulted in 130 models that generate bounding boxes and segmentation masks for each instance of a particular object within an image. The framework maintains an API functionality for processing new images with any of the three models: Yolact, Detectron2 and MIC21. The MIC21 framework allows for evaluation and comparison of the grand truth and MIC21 models annotations as well as for running the models on new image datasets.<sup>4</sup>

We will present in more detail the integration of existing models in our Image Processing Pipeline in order to automatically predict the objects' boundaries in images and their classes, as well as the development of 130 models based on the Multilingual Image Corpus, which can be used for object recognition and classification and for future experiments.

### 3 Image Processing Pipeline

The Image Processing Pipeline contains a number of modules that support the work of annotators in several ways: by predicting object outlines and by managing images and annotations.

The Multilingual Image Corpus is organized into manageable in size datasets containing at least 100 images (in rare cases) and in the most common case – about 150 images. The initial datasets roughly correspond to the final thematic subdomains that will be formed. However, in the preliminary stage of the work, it was acceptable to have several datasets representing a single thematic subdomain, images classified in inappropriate datasets, etc. Therefore, the initial organization of the images into datasets is mostly with respect to the collection methodology and the decision on the size of the data. After the initial processing of the images and the manual annotation, some images are reorganized, if necessary, which reflects the final content of the thematic subdomains.

All input images are represented in raster image compression formats such as: Portable Network Graphic (PNG), Joint Photographic Experts Group (JPEG) or Tag Image File Format (TIFF). The size of images varies considerably between 2 to 11 megapixels. The small image dimensions

may affect the quality of the annotated regions. On the other hand, when the image dimensions are too large, the amount of allocated memory, as well as the processing time, increases exponentially. Another requirement for the input image is its colour space format to be in red, green and blue (RGB) channels without additional or missing channels.

For an effective processing of images with a convolution model they have to be in proper dimensions and in the RGB color space. Hence, the first module of the pipeline examines each image and performs the necessary transformations (resizing and/or color space mapping); in case the transformations are not possible, the image is excluded from the dataset. As a result, the images are described by their attributes as JSON objects. The pre-processing step is automated by a Python script, which calls some of the OpenCV<sup>5</sup> routines for performing the operations over the images.

The open source Yolact (Bolya et al., 2019) provides several convolution neural network models for object detection and segmentation within the COCO domain (Lin et al., 2014). The model we have employed for the automatic annotation is Resnet50-FPN. The notation indicates that each convolution layer from the backbone component of the model includes a feed-through connection from the input to the output of the layer. Such structure is appropriate for training highly stacked models (as those used for computer vision tasks) in order to improve the numerical stability of the optimization procedure and to prevent over-fitting. The intuition behind this is that each convolution layer from the stack is approximating only the residual error between the target and the output of the previous layer. First the data is processed through a stack of input convolutional layers with decreasing resolution. Consequently, a stack of output convolutional layers with increasing resolution is applied. The dimension of each layer from an input stack is matched by a layer from the output stack. In addition, 1x1 convolutional connections are established between the corresponding layers from the input and the output stacks.

The result from dataset processing through Yolact software is the instance segmentation and classification within the COCO domain. The results are stored in MS COCO JSON format. The format provides two options for recording the bounding contour of each detected object – run-

---

<sup>4</sup><https://mic21.dcl.bas.bg>

<sup>5</sup><https://github.com/opencv/opencv>

length encoding (RLE) and point coordinates. The RLE is a compression format over the point coordinates and allows for more compact representation; however, not all systems are able to work with it directly. A useful tool for converting between formats is the Python library `pycocotools`. In some cases, more processing is required because the raw segmentation mask resulting from the convolution model is a binary mask. For the conversion of a binary mask to an object contour a useful routine from OpenCV library, `findContours`, is used.

As noted, the object detection model produces annotation data in the domain of the 80 COCO categories. The automatically obtained annotation data are imported in an open-source annotation software, the COCO-annotator (Brooks, 2019). The process is automated through a bash script, which connects to the database docker of the annotator, creates a new dataset, copies the images and imports the annotation data. The COCO-annotator features multi-user environment composed of a mongoDB database, Flask backend and Vue front-end employing a worker processing model. In the COCO-annotator, software images are organized into datasets and the front-end provides a tool for performing manual editing of annotation contours creating/deleting annotations or changing/assigning object labels.

To further accelerate manual annotation an automatic relabelling of the imported annotations in the coco-annotator database is implemented. It takes as an input a dictionary that states the relabelling rules specific for a sub-dataset. For example, the category 'Person' in the sub-dataset 'Basketball' is replaced with the class 'Basketball player' and the identifier of the new class replaces the identifier of all annotations 'Person' within the sub-dataset 'Basketball'.

Certain manipulations during the manual annotation were performed to provide functionalities that are not implemented in the COCO-annotator software. We have developed a dedicated Python script performing such operations by connecting to the mongoDB engine of the annotation software using PyMongo Python library:

- Import images and annotations when creating a new dataset;
- List annotated images by class label and store them into a file on the disk;
- List hyperlinks to images in the database ac-

ording to their thematic subdomain;

- List annotated images by thematic subdomain;
- Generate statistical reports for the annotated images;
- Merge two thematic sub-datasets into a single one;
- Move one or several images from one thematic sub-dataset to another, together with their associated annotations;
- Remove images, annotations or categories from a dataset based on various criteria;
- Export all dataset images and annotations into JSON COCO format;
- Remove images from the dataset marked as deleted in the COCO-annotator software;
- Replace labels in a thematic sub-dataset according to specific rules;
- Scan image paths in the database for inconsistency and fix them if necessary;
- Change classes of particular images or differentiate labels between two distinct thematic sub-datasets.

Python scripts<sup>6</sup> execute each of the described operations.

After manual annotation and database post-processing, the images from the resulting 130 thematic subdomains are exported together with their ground truth annotations represented in MS COCO format. The structure of the MIC21 dataset is as follows:

```
-thematic_field_name
- data
  - image_1.jpg
  - image_2.png
  ...
thematic_field_name_gt.json
```

The data sub-directory for the respective thematic subdomain contains the images in jpg, jpeg or png format. The `*gt.json` field is a COCO format JSON file describing the polygonal segmentation of objects in images.

<sup>6</sup>[https://github.com/link\\_will\\_be\\_provided](https://github.com/link_will_be_provided)



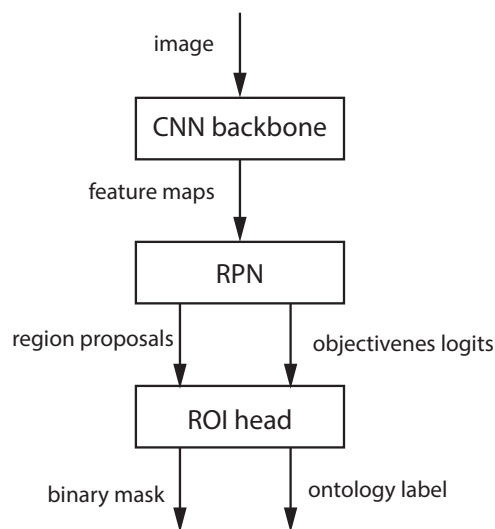


Figure 1: Structure of the model

#### 4 MIC21 Models

We trained domain specific models, which are able to detect and label objects in an image with classes from the MIC21 ontology. The benefits of such models are twofold – first, they allow further acceleration of the manual annotation reflecting the MIC ontology of visual objects; second, they represent an extension of the standard COCO classes to 130 thematic domains. For the training of the domain specific models we use Detectron2 framework of the Facebook research group<sup>7</sup>, which is an open-source Python software based on PyTorch library. The general structure of the object detection model is presented in Figure 1.

It is composed of a convolution neural network (CNN) backbone component, a proposal generator and a region of interest (ROI) head (Redmon et al., 2015). Detectron2 supports 3 backend structures - Resnet, Regnet and FPN. The backbone is represented as stacked convolution layers with different interconnections depending on the structure type.

In the Resnet structure, the residual building block has an option for a direct shortcut connection from the input of the layer to its output, i.e. projecting the input features into the output, and the actual network is keeping the difference between the input and the target. The output of the backend component comprises selected feature maps from the stack of layers depending on the network design. Usually, in addition to the output of the last layer, 3 to 4 of the output feature maps from the

deeper layers are selected.

Region proposals in the modern object detection networks are generated through a region proposal convolution network (RPN), composed of a 3x3 convolution layer followed by 1x1 convolution layers for the generation of object box deltas and an objectiveness score for a box. As a basis for region proposals, a set of anchors is generated within the image, for example, by dividing the image into a grid of large boxes and putting an anchor point at the centre of each box. Then the anchor boxes are refined during the training of the network by fitting them to the ground truth. The result of the proposal generator layer is a list of box coordinates and an objectiveness score indicating whether the respective box contains an object or not.

The third stage in the object detection framework is a ROI head network, which iterates over each of the generated proposal boxes and performs per region classification and binary mask extraction. The prediction is based on the features from the backbone layer constrained to the current examined box.

Each of the trained models is characterised by input data, output data and parameters. The parameters represent the internal weights of the model obtained during training, which are specific for each thematic subdomain and have to be loaded through a `DetectionCheckpointer` class. The input for the model is a Python list structure `list[dict]`, where each element of the list is a dictionary field `image`, representing a 3-dimensional array with colours for each pixel from the image in RGB colour space, and also width and height attributes for the image in pixels. If the model weight is updated during the training, the input dictionary for each image has to include a field `instances` describing the coordinates of the bounding boxes for the ground truth objects in the image, as well as a class label for each object in the range `[0, num_categories]` and a ground truth binary mask for each object.

The training of the models is performed with the Detectron2 framework by inheriting the `DefaultTrainer` class. The training loop and each of the network layers are aligned with the PyTorch requirements for building neural network models. Each layer has to provide a `loss` function, which calculates the residual error given the training targets and network outputs, and additionally to provide a `forward` function, which calculates the

<sup>7</sup><https://github.com/facebookresearch/detectron2>

layer outputs from inputs. To compute the gradients during the backward network pass the PyTorch features an `autograd` engine, which (when enabled) is able to track each arithmetic operation during the forward pass and to obtain the gradient of the residual error of the layer with respect to the parameters.

During the training, the model is evaluated periodically when a certain iteration count is passed by tracking the intersection over union (IoU) metric by category. The library `pycocotools` contains useful routines for comparing results from computer vision models either by using bounding boxes or binary masks. The IoU metric is defined as:

$$S_{IoU}(Z, T) = \frac{\mathcal{A}(Z \cap T)}{\mathcal{A}(Z \cup T)}, \quad (1)$$

where  $Z$  is the model bounding box or mask detection,  $T$  is the corresponding ground truth instance from the same ontology class and  $\mathcal{A}$  is the area calculating operator, which usually is expressed by the number of pixels in a region. In order to calculate this metric for multivariate models (with several ontology classes), first a correspondence between model outputs and ground truth targets has to be established by comparing the region overlap. When a model output region is matched to a ground truth region for a given IoU threshold, 4 metrics can be calculated:

- TP – true positive – when  $Z$  and  $T$  are from the same class;
- FP – false positive – when  $Z$  and  $T$  are matched, but the model is wrong for the class of  $Z$ ;
- FN – false negative – no region  $Z$  is matched to a ground truth  $T$ ;
- TN – true negative – an object that is not part of the ground truth is also left undetected by the model.

With respect to the classification outcome for a given IoU threshold, three additional metrics for the model are commonly examined, which are model precision

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}}, \quad (2)$$

reflecting how good the model is in producing correct labelling for the detected regions, model recall,

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}}, \quad (3)$$

which is about how good the model is in detecting the correct objects from a category and general model detection accuracy expressed by

$$A = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{FP} + N_{FN} + N_{TN}}, \quad (4)$$

where  $N_{\bullet}$  denotes the number of detections over the whole dataset from TP, FP, FN or TN category. Note that each of the metrics P, R and A are functions of the IoU threshold level  $S_{IoU}(Z, T)$  used to perform the matching between the model detection and ground truth regions. By selecting different IoU thresholds we will get different model performance, hence to obtain a more complete picture of the model detection capabilities  $P$  and  $R$  are evaluated for the whole range of IoU values from 0 to 1, producing the so-called precision recall curve for a model.

## 5 Framework for running and evaluation the MIC21 models

For the purposes of presentation, comparison and evaluation the dataset is organized into a system of components, called 'MIC21 framework' (Figure 2). The framework is composed of a backend (processing service) and frontend (visualization service) component. The processing service is implemented as a Flask server implemented in Python, which is able to run the Yolact, Detectron2 and MIC21 pre-trained domain specific models (Figure 2). The processing service offers a set of Web APIs implemented over an HTTP, with the following functions:

- Prediction of annotations using the Yolact software;
- Prediction of annotations using the Detectron2 software;
- Prediction of annotations using the MIC21 trained models;
- Import of new images, their ground truth and predictions into FiftyOne framework (into a new or already existing dataset);
- Initial loading of datasets into FiftyOne;
- A simple interface to upload a new image to a dataset;
- Evaluation of predictions against the ground truth and storing the results into the FiftyOne framework. Print the evaluation statistics.

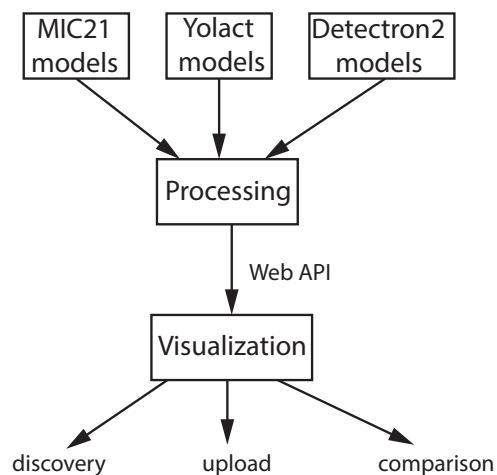


Figure 2: MIC21 framework components

The framework frontend service is based on the FiftyOne software (Moore and Corso, 2020), which is an open-source tool for building high-quality datasets and computer vision models. In the FiftyOne frontend service the Yolact, Detectron2 and MIC21 pre-trained models can be compared with the ground truth annotations for 130 subdomains in the Sport, Transport, Arts and Security thematic domains.

The repository for the framework is freely accessible at GitHub and includes: source code of Mask R-CNN built on FPN and ResNet101, training code for MS COCO, pre-trained weights for MS COCO and MIC21 classes, Jupyter notebooks for visualization the detection pipeline and evaluation routines for MS COCO metrics integrated in the FiftyOne.

The original FiftyOne code is extended with a function, which allows a user to upload a new image into a selected MIC21 sub-dataset. When uploaded, the image is automatically processed by the backend service and it is annotated independently by 3 different models (Yolact, Detectron2 and MIC21). The results can be compared in the FiftyOne.

## 6 Results and Evaluation

The MIC21 object detection models produce the output in four components:

- Bounding box, described with its coordinates and its width and height;
- Polygon of points outlining the object contours;
- Class label of the detected object;

- Confidence score between 0 and 1 – how certain the model is about the predicted class.

The FiftyOne framework integrates functionalities to compare different object detection models (in our case MIC21 model outcomes and the ground truth annotations). As each image can represent many objects from different classes, the comparison shows the correct classification for a given object within an image. Overall, the results can be summarized as follows:

- If a model could find the object location, the object is assigned a class;
- The classes that have strong correlation with the COCO classes (i.e. *baseball player* vs. *human*; *soldier* vs. *human*, etc.) are recognized with a better precision, over 90 %, such classes represent 27 % from the MIC21 Ontology classes;
- Classes representing objects that are not categorized in the COCO dataset are recognized and classified with accuracy over 50 %. Figure 3 represents details for four randomly selected sub-datasets.

Category	Accur.	Precis.	Recall	Support
Sport	0.60	0.83	0.63	1663
Transport	0.59	0.84	0.63	1421
Arts	0.59	0.89	0.63	855
Security	0.20	0.39	0.24	1973
Total	0.5	0.76	0.53	1478

Table 2: Average metrics for the 130 MIC21 models

The results depend on the selected model parameters, number of training epochs, batch size and also on the structure of the train and the validation datasets. In our experiment, an initial training is performed with a fixed number of 1500 epochs. The resulting models can be re-trained further by the code templates provided within the framework. After the initial training, we have calculated average accuracy, precision and recall metrics for each Ontology class represented in the MIC21 dataset. The low accuracy and recall for some classes from the domains Security and Arts is due to the small number of the ground truth instances. Methods for training models with small datasets have already been developed (Chen et al., 2021; Hu et al., 2020).

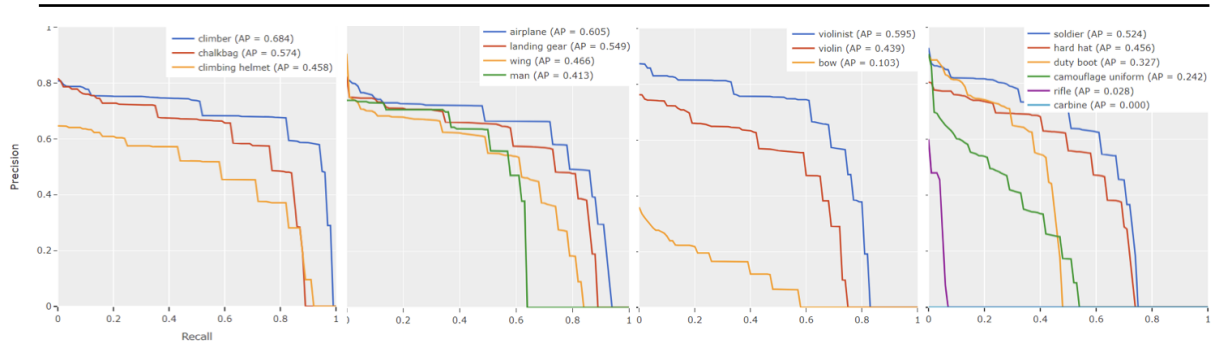


Figure 3: Precision-recall relationship for 4 randomly selected sub-datasets. From left to right: Climbing, Airplane, Violinist and Soldier

During the training we have used a fixed number of 1 500 epochs to generate results for the 130 models in reasonable time frame. However, training for a fixed number of iterations leads to a different performance of the model over the dataset. We summarize the results from the evaluation of the trained models over the target domains in figures 4 to 7. For this purpose we calculate average precision and recall metrics using the official COCO API library (Dollar and Lin, 2014). The library provides the class `COCOEval`, which takes ground truth and model detection arrays as inputs and evaluates each image and category in the dataset over specified surface area ranges. The matching between ground truth and detection masks is determined by a range of intersection-of-unions thresholds (IoT). In our evaluation scenario, we evaluate all area ranges, and the IoT range from 0.5 to 0.95 for both precision and recall. The per image metrics are then accumulated for the whole dataset.

To highlight the capabilities of the trained models, we have performed evaluation over a single class per dataset, which was selected as a dominant class for the particular subdomain. These classes are usually with high number of instances in the subdomain. For some of the subdomains the selection of a dominant class can be ambiguous. However, the main rule during that decision process was that the dominant class must uniquely identify the respective subdomain. Another rule we have observed was to use different sets of classes for different subdomains (to the possible extend). The resultant sets of classes can be tracked in the Figures 4 to 7.

In the Figure 4 we compare the subdomains from the domain Sport. We can see that for the most of the dominant classes we have reached average precision and recall of about 0.4. The highest average precision is reached for the category **Golf player**

in the subdomain **Golf**, and the lowest precision and recall are for the class **Race driver** in the subdomain **Car racing** (not a dominant class), which is due to uncommon for COCO models pose of the object **Person** within the images and provided limited training epochs. Hence, if we deviate more from what is typical of the initial training of the model, we have to perform deeper changes in the layers with the additional training in order to preserve the level of fit to ground truth. It is noticeable that we have big difference between average precision and recall for some classes such as **Volleyball player**, **Soccer player**, **Hockey player**, **Cricketer**, etc. In all cases, the recall is about 30% lower than the precision, meaning that when the model detects the respective instance it is correctly classified. However, not all objects from a particular class have been detected. This can be related to how the ground truth objects are selected, in terms of a sufficient number of images depicting the object in a particular situation, or can be attributed to overlapping between two objects in some situations. Such events can lead to lower confidence score from the model. We evaluate all MIC21 models for a confidence score of 0.9, which is quite high, to highlight the differences between the models.

Figure 5 contains the results of evaluation over the domain Transport. The pre-trained models have reached a performance between 0.4 and 0.8 for average precision and recall for that domain. The highest result on precision is for the subdomain **Tram** of about 0.85 and the highest recall is for the subdomain **Convertible** of about 0.84. It is interesting to note that, while for the Sport subdomains we always have higher precision than recall, for the Transport subdomains we have many cases when recall is higher than the precision. This indicates that, while selected models are better at recognizing



ing automobiles, it is more difficult to identify them than people. The lowest metrics for Transport are for the subdomain **Car transporter**, which can also be attributed to the untypical objects we try to identify using a model that was first trained for the 80 COCO classes.

Results from the evaluation of the models in the domain Arts are presented in the Figure 6. With some models targeting higher and some models aiming lower, the average precision and recall are about 0.4-0.5. We cannot see a precision over recall dominance as in Sport because the dominant class in those subdomains is once again Person. With recall and precision close to 0.85 and 0.83, the subdomain **Photographer** achieves the higher metrics.

Interesting finding is that as with Cellists, Ballet dancers, and Percussionists, lower recall levels unnecessarily reduce precision levels. In other words, if the model is good in detecting a particular object, it has good chances to properly classify it. This can be related to the fact that we modify only the ROI head sub-component of the model, without targeting the lower detection layers. In our dataset we have a few classes from the domain Security visualized in the Figure 7. Classes from the domain Security show similar metrics with other examined domains with both precision and recall ranging around 0.5.

## 7 Conclusion and Future Work

The Multilingual Image Corpus offers data to train models specialized in object detection, segmentation, and classification by providing fully annotated objects within images with segmentation masks, categorised according to an Ontology of Visual Objects. The Ontology of visual objects allows easy integration of annotated images in different datasets as well as learning the associations between objects in images.

Models trained on the COCO dataset were used for the generation of annotation proposals. We developed a framework based on FiftyOne, Yolact and Detectron2, and implemented it over Mask R-CNN on Python3, Keras and TensorFlow. We pre-trained Fast R-CNN models with the MIC 21 dataset, which resulted in 130 models that generate bounding boxes, segmentation masks and object classes.

The MIC 21 framework supports web-based visualization, evaluation and comparison of different

models together with the ground truth annotations.

We can provide a number of alternatives for completing multimodal tasks using the created datasets, including automatic image caption generation aligning sentences with images in various multimodal documents and visual question answering. Interpreting an image and the brief text that goes with it, such as a caption, a question or a description of the objects in the image, can be a supporting task.

Prospective developments also include: automatic extension of the dataset using the pre-trained models, which will considerably accelerate manual annotation in the target thematic domains; training models for automatic image captioning (Li et al., 2020) or question answering (Wu et al., 2021; Liu et al., 2021); adaptation of the pre-trained models for video processing (Zhao et al., 2021); identification (automatic generation) of images representing particular objects or particular textual descriptions; application in motion analysis systems.

## Acknowledgments

The Multilingual Image Corpus (MIC21) project was supported by the European Language Grid project through its open call for pilot projects. The European Language Grid project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 825627 (ELG).

## References

- Mark Amo-Boateng, Nana Ekow Nkwa Sey, Amprofi Ampah Amproche, and Martin Kyereh Domfeh. 2022. [Instance segmentation scheme for roofs in rural areas based on Mask R-CNN](#). *The Egyptian Journal of Remote Sensing and Space Science*.
- Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. 2019. [Yolact: Real-time instance segmentation](#). In *ICCV*.
- Justin Brooks. 2019. [COCO Annotator](#).
- James A.D. Cameron, Patrick Savoie, Mary E. Kaye, and Erik J. Scheme. 2019. [Design considerations for the processing system of a CNN-based automated surveillance system](#). *Expert Systems with Applications*, 136:105–114.
- Tingkai Chen, Ning Wang, Rongfeng Wang, Hong Zhao, and Guichen Zhang. 2021. [One-stage CNN detector-based benthonic organisms detection with limited training dataset](#). *Neural Networks*, 144:247–259.

- Christopher R. Conrady, Şebnem Er, Colin G. Attwood, Leslie A. Roberson, and Lauren de Vos. 2022. [Automated detection and classification of southern African Roman seabream using mask R-CNN](#). *Ecological Informatics*, 69:101593.
- Fan Cui, Muwei Ning, Jiawei Shen, and Xincheng Shu. 2022. [Automatic recognition and tracking of highway layer-interface using faster R-CNN](#). *Journal of Applied Geophysics*, 196:104477.
- Piotr Dollar and Tsung-Yi Lin. 2014. [Microsoft COCO Toolbox. Version 2.0](#).
- Christiane Fellbaum, editor. 1999. *WordNet: an Electronic Lexical Database*. MIT Press, Cambridge, MA.
- Ross Girshick. 2015. [Fast R-CNN](#).
- Xiaodong Hu, Xinqing Wang, Fan-jie Meng, Xia Hua, Yu-ji Yan, Yu-yang Li, Jing Huang, and Xue-mei Jiang. 2020. [Gabor-CNN for object detection based on small samples](#). *Defence Technology*, 16:1116–1129.
- Ahmed Kasapbaşı, Ahmed Eltayeb Ahmed Sibushra, Omar Al-Hardanee, and Arif Yilmaz. 2022. [Deep-ASLR: A CNN based human computer interface for American Sign Language recognition for hearing-impaired individuals](#). *Computer Methods and Programs in Biomedicine Update*, 2:100048.
- Svetla Koeva. 2021. [Multilingual Image Corpus: Annotation Protocol](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 701–707, Held Online. INCOMA Ltd.
- Svetla Koeva, Ivelina Stoyanova, and Jordan Kravev. 2022. [Multilingual Image Corpus – Towards a Multimodal and Multilingual Dataset](#). In *Proceedings of the Language Resources and Evaluation Conference*, pages 1509–1518, Marseille, France. European Language Resources Association.
- Ruifan Li, Haoyu Liang, Yihui Shi, Fangxiang Feng, and Xiaojie Wang. 2020. [Dual-CNN: A Convolutional language decoder for paragraph image captioning](#). *Neurocomputing*, 396:92–101.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. [Microsoft COCO: Common Objects in Context](#). In *European Conference on Computer Vision (ECCV)*, pages 740–755, Zürich.
- Yun Liu, Xiaoming Zhang, Qianyun Zhang, Chaozhuo Li, Feiran Huang, Xianghong Tang, and Zhoujun Li. 2021. [Dual self-attention with co-attention networks for visual question answering](#). *Pattern Recognition*, 117:107956.
- Umberto Michelucci. 2019. *Advanced Applied Deep Learning: Convolutional Neural Networks and Object Detection*. Apress.
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to Wordnet: an on-line lexical database. *International journal of lexicography*, 3(4):235–244.
- Brian Moore and Jason Corso. 2020. [Fiftyone](#). *GitHub*.
- Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2015. [You Only Look Once: Unified, Real-time Object Detection](#).
- Yuxin Sun, Li Su, Yongkang Luo, Hao Meng, Wanyi Li, Zhi Zhang, Peng Wang, and Wen Zhang. 2022. [Global Mask R-CNN for marine ship instance segmentation](#). *Neurocomputing*, 480:257–270.
- Yirui Wu, Yuntao Ma, and Shaohua Wan. 2021. [Multi-scale relation reasoning for multi-modal Visual Question Answering](#). *Signal Processing: Image Communication*, 96:116319.
- Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. [Detectron2](#).
- Guoping Zhao, Mingyu Zhang, Yaxian Li, Jiajun Liu, Bingqing Zhang, and Ji-Rong Wen. 2021. [Pyramid regional graph representation learning for content-based video retrieval](#). *Information Processing & Management*, 58(3):102488.

**Appendix A Evaluation over the classes in the four main domains: Sport, Transport, Arts, Security**



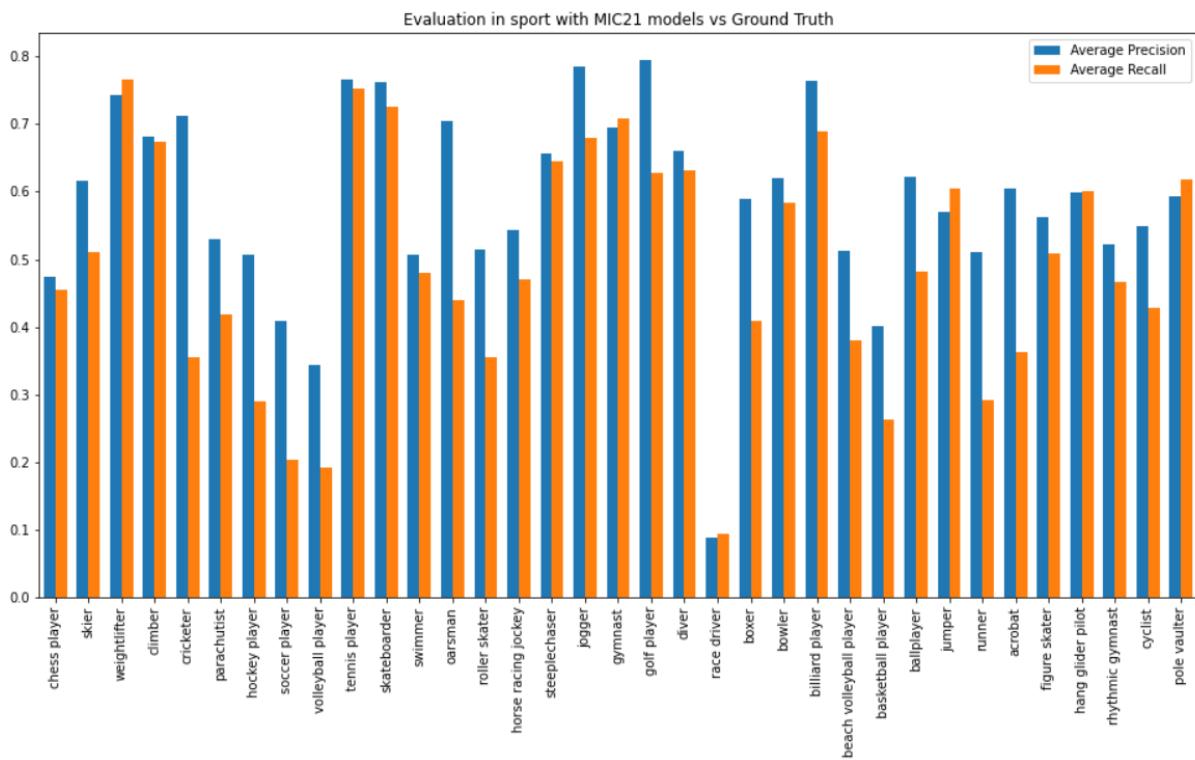


Figure 4: Evaluation over Sport categories

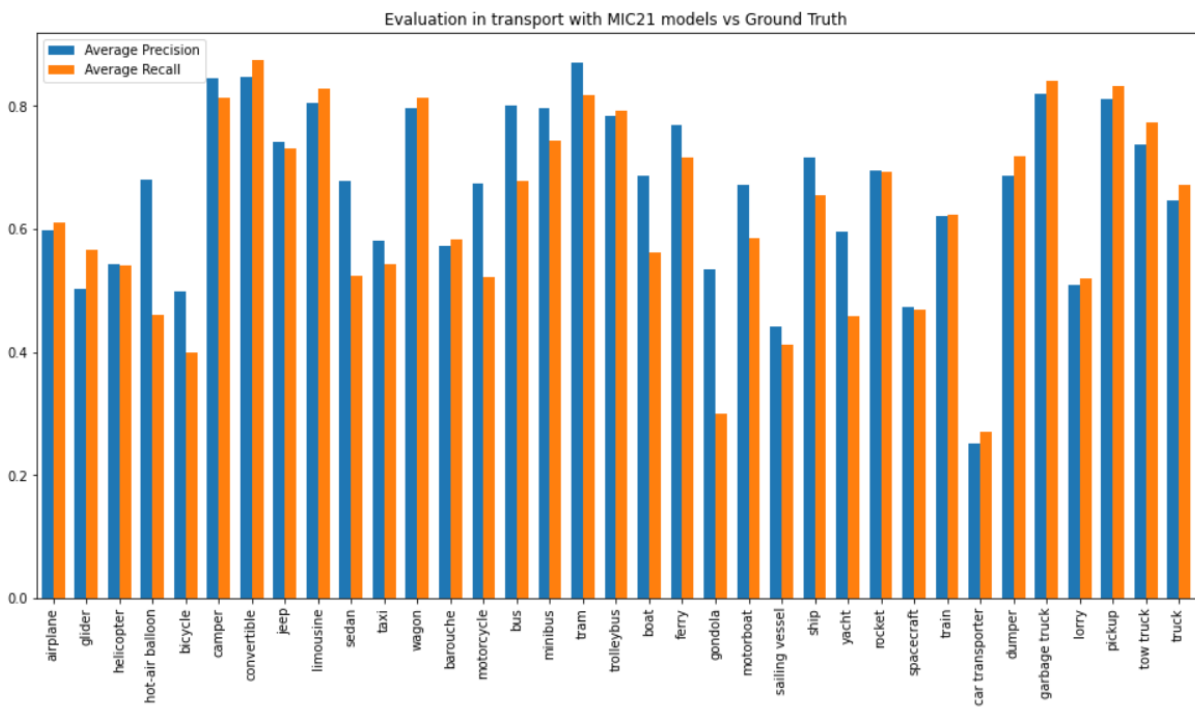


Figure 5: Evaluation over Transport categories

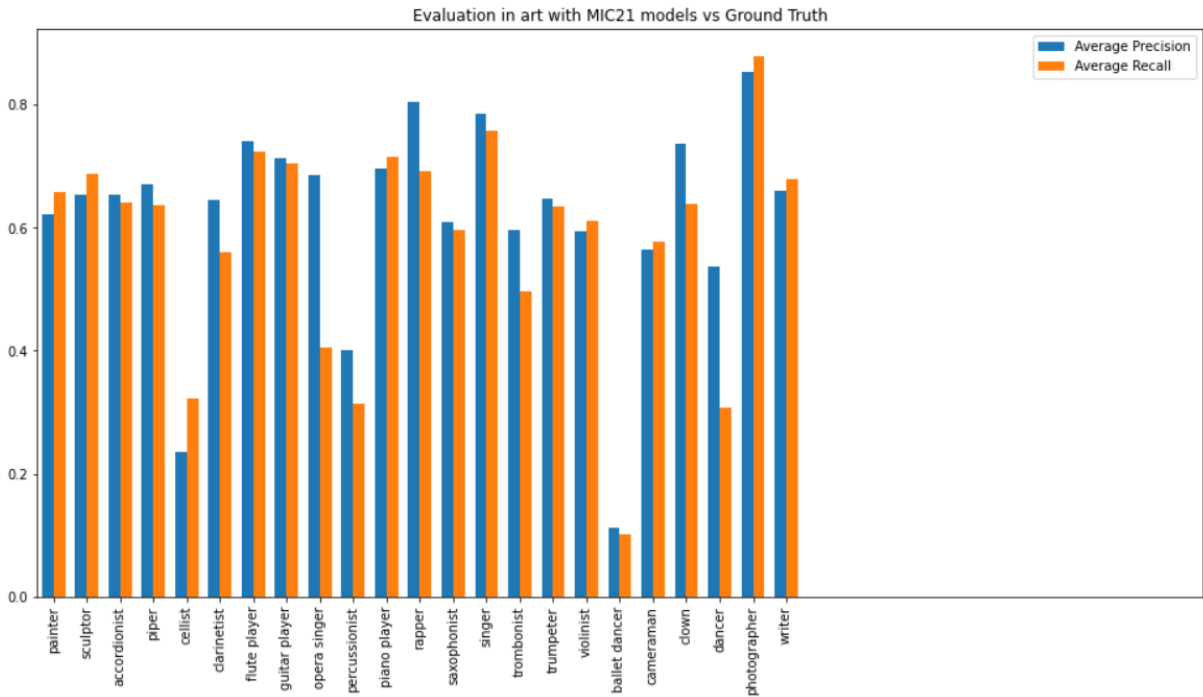


Figure 6: Evaluation over Arts categories

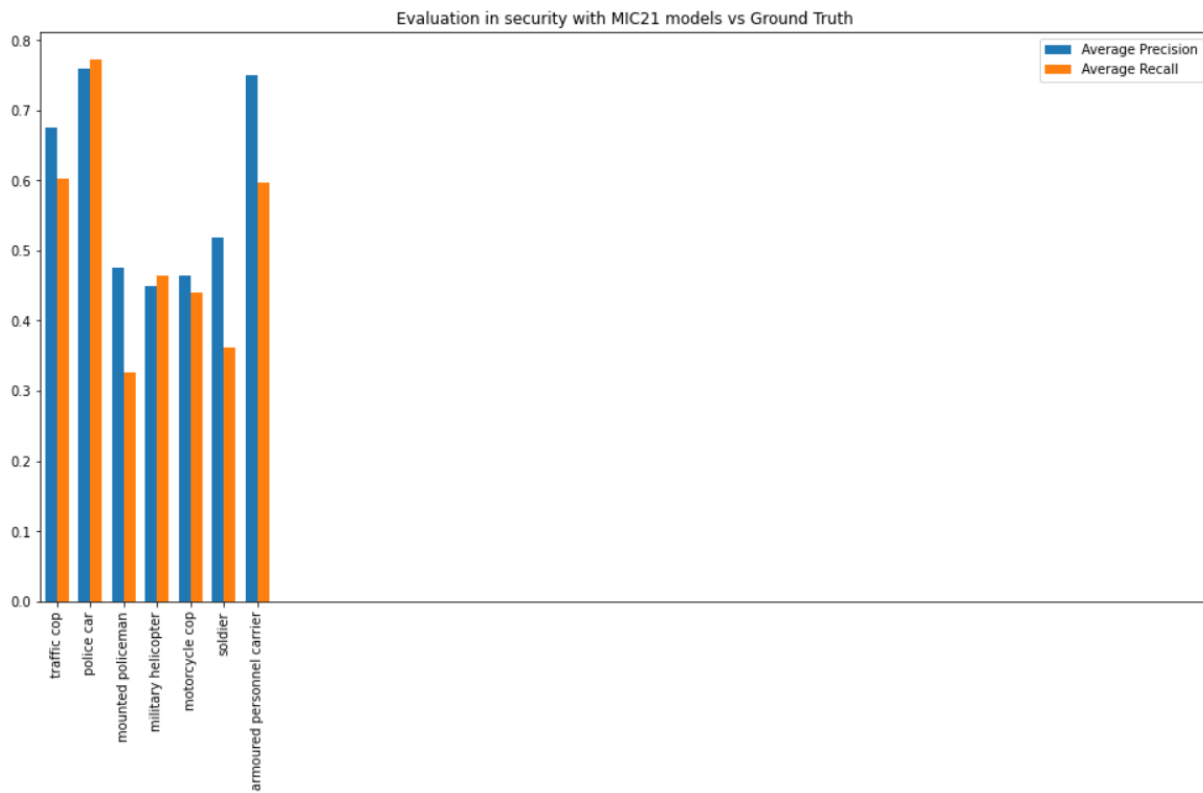


Figure 7: Evaluation over Security categories

---

**SPECIAL SESSION ON  
WORDNETS, FRAMENETS AND ONTOLOGIES**

---

# Ontology Supported Frame Classification

**Svetla Koeva**

Institute for Bulgarian Language, BAS  
svetla@dcl.bas.bg

**Emil Doychev**

Plovdiv University Paisiy Hilendarski  
e.doychev@uni-plovdiv.bg

## Abstract

We present **BulFrame** – a web-based system designed for creating, editing, validating and viewing conceptual frames. A unified theoretical model for the formal presentation of Conceptual frames is offered, which predetermines the architecture of the system with which the data is processed. A **Conceptual frame** defines a unique set of syntagmatic relations between verb synsets representing the frame and noun synsets expressing the frame elements. Thereby, the notion of Conceptual frame combines semantic knowledge presented in WordNet and FrameNet and builds upon it. The main difference with FrameNet semantic frames is the definition of the sets of nouns that can be combined with a given verb. This is achieved by an ontological representation of noun semantic classes. The framework is built and evaluated with Conceptual frames for Bulgarian verbs.

**Keywords:** Conceptual frames, ontology of noun semantic classes, verb semantics

## 1 Introduction

There are many rich semantic resources (mainly for English but also for other languages) that include different types of semantic information: WordNet (Miller et al., 1990), FrameNet (Baker et al., 1998), VerbNet (Kipper et al., 2007), PropBank (Palmer et al., 2005), Ontonotes (Weischedel et al., 2011), Pattern Dictionary of English Verbs (Hanks, 2004), Yago (Suchanek et al., 2007), BabelNet (Navigli and Ponzetto, 2012), VerbAtlas (Di Fabio et al., 2019), SynSemClass (Uresova et al., 2020), among others.

**BulFrame**<sup>1</sup> is a web-based system designed for creating, editing, validating and viewing Conceptual frames. A unified theoretical model for the formal presentation of conceptual frames is offered,

<sup>1</sup><https://dcl.bas.bg/bulframe/>

which predetermines the architecture of the system with which the data is processed. In this regard, several fundamental theoretical models focused on verb semantics have been taken into account – among the most famous research in this field are Charles Fillmore’s theory of frame semantics (Fillmore, 1982), the description of verb classes and possible alternations by Beth Levin (Levin, 1993), the concept of representation of verb frames in FrameNet (Baker et al., 1998; Fillmore and Baker, 2001) and others.

Some of the main advantages of both resources (WordNet and FrameNet) with regard to the conceptual description of the predicate – argument structure are complemented and upgraded to expand WordNet with Conceptual frames that represent verb predicate-argument syntagmatic relations. The main advantages of WordNet for semantic analysis focused on introducing Conceptual frames are: a) the large number of concepts organized in a semantic network and b) the grouping of concepts in semantic classes according to their general meaning. The main advantages of FrameNet for implementing Conceptual frames are: a) the extensive description of semantic knowledge about an event type and its participants and b) the linking of semantic frames with semantic relations (Koeva, 2021).

The paper is organized as follows: we begin with a brief introduction to the notion of **Conceptual frame** in Section 2. In Section 3 we present the design of the **BulFrame** system. Section 4 is dedicated to the linguistic interpretation of Conceptual frames with a special focus on the ontology of semantic classes of nouns. Finally, related work (section 5), conclusions and future directions of our work (section 6) are presented.

Our main contributions are: (a) identification of verbs that evoke a particular FrameNet semantic frame; (b) detailed ontological representation

of semantic classes of noun synsets; (c) specification of frame elements relevant to the expression of syntagmatic relations; (d) assigning the frame elements with noun semantic classes or a combination of classes ensuring the words' compatibility in Bulgarian; (e) definition of Conceptual frames depicting semantics of Bulgarian verbs.

## 2 The Notion of Conceptual Frames

**Conceptual frames** are abstract structures that define the semantic and syntactic compatibility between verb predicates and noun arguments. A particular Conceptual frame is: associated with a semantic class that expresses its general semantic properties; represented by a set of verbs organized in the WordNet synonym sets, and described by a set of frame elements. The verbs in the same frame can be one or several: linked between each other with lexical relations (synonymy, antonymy) and/or hierarchical relations (hypernymy, troponymy, entailment). The Conceptual frame elements roughly correspond to the FrameNet core elements; however, there is no one-to-one correspondence between FrameNet Semantic frames and Conceptual frames (because of some differences in conceptualization in different languages and because of differences between the two theoretical representations) (Koeva, 2020, 2021).

Each Conceptual frame element is associated with a set of nouns that are compatible with the verb predicate. Again, the set could contain a single noun or several nouns linked between each other with lexical relations (synonymy, antonymy) and/or hierarchical relations (hypernymy, hyponymy). The association between the frame (verb synsets) and its elements (noun synsets) can be explicitly introduced in WordNet by means of syntagmatic relations. If more than one noun synset can express the frame element (which is the usual case), the syntagmatic relation links the verb synset with the top-most noun synset of the hierarchy grouping nouns with the same semantic properties (semantic class). The diversity in the compatibilities between representatives of verb classes and noun classes drives the necessity for a detailed **Ontology of semantic classes of nouns**.

We can generalize that a **Conceptual frame** defines a unique set of syntagmatic relations between:

- verb synsets representing the frame, and
- noun synsets expressing the frame elements.

Thereby, the notion of Conceptual frame com-

bines semantic knowledge presented in WordNet and FrameNet and builds upon it.

The main difference between Conceptual frames and the FrameNet Semantic frames (Ruppenhofer et al., 2016) is that Conceptual frames are explicitly linked with the noun synsets representing the words with which the verb predicate can be combined (to the extent this is possible due to WordNet structure and content and metaphoric language use). For example, a Conceptual frame that roughly corresponds to the FrameNet semantic frame **Experiencer focused emotion**<sup>2</sup> is represented by the verb synsets: **dislike** 'have or feel a dislike or distaste for'; **hate, detest** 'dislike intensely; feel antipathy or aversion towards'; **like** 'find enjoyable or agreeable'; **love** 'have a great affection or liking for'. The Conceptual frame elements are Experiencer and Content (if we keep the names of the FrameNet core elements). The semantic classes of nouns that they could be expressed with are **[Human]**, **[Animal]**, **[Physical entity]**, and **[Abstraction]** and the combinations are the following:

- **Experiencer:** person — **Content:** physical entity and abstraction
- **Experiencer:** animal — **Content:** physical entity.

## 3 BulFrame Design

**BulFrame** is a system whose functionality is designed for the definition and description of Conceptual frames. The functionality is divided into three main modules: (a) definition of the abstract structure; (b) description of particular Conceptual frames based on the defined structure; and (c) public access to the Conceptual frames, with a read-only restriction.

### 3.1 Abstract structure

The abstract structure of the system provides a complete set of components and operations for setting up any hierarchical structure. Moreover, it can be changed over time taking into account the risk of information loss after certain operations.

#### 3.1.1 Objects

The abstract structure has only one object type, which is defined by attributes related to the object with system internal relations. Thus, the difference between the object and the attribute is that the object does not have a parent, or in other words, it is

<sup>2</sup><https://framenet.icsi.berkeley.edu/fndrupal/frameIndex>

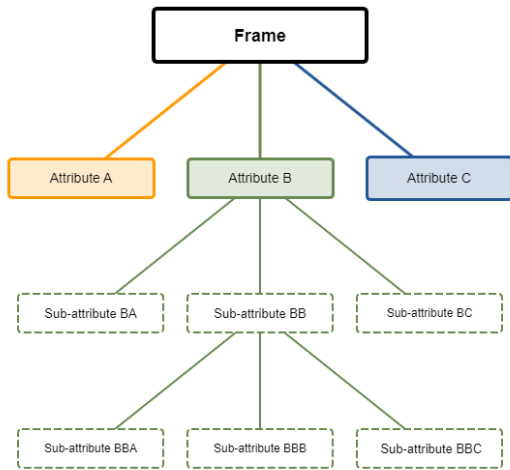


Figure 1: Abstract Structure of the Object FRAME

always a root, while the attribute is always related either directly to the object or to the other attribute (the attribute can never be a root). In **BulFrame** implementation of the abstract structure the object has two instances: **WORD** and **FRAME**. The object instances are linked with specially defined relations by means of the following formal properties: equivalence, reflexivity, transitivity. The relations might be: FRAME to FRAME, WORD to WORD and FRAME to WORD.

The abstract structure is represented as a strictly hierarchical structure. This is illustrated at Figure 1, in which the root is the object FRAME and the other nodes represent the frame attributes. The hierarchical organization of the abstract structure is achieved in two ways: by nesting of object attributes and by encoding taxonomy relations between objects.

### 3.1.2 Attributes

The attributes have a uniform structure represented by the pair **key : value**, where the key is the name of the attribute and the value determines the way the value is defined: directly by a value or by a sub-attribute. Defining the attribute value as a sub-attribute forms complex (nested) attributes.

Each attribute, as a separate element of the hierarchical structure, is defined in tables from the database as shown in Figure 2.

The value types that are supported by the framework are: text, number, relation and sub-attribute (for complex attributes). In addition to defining the value type of the attributes, the framework also provides the opportunity to define the type of the visual component with which the corresponding value has

to be represented. The supported components are:

- for the value **text**: a text box, a text box with autocomplete function based on the existing values for the same attribute, a drop-down with a single select option based on a predefined list of values and a drop-down with multiple select options based on a predefined list of values;
- for the value **number**: a numeric box;
- for the value **relation**: a combo-box based on the predefined FRAMES and WORDS.

Table 1 contains the general information about the attributes (the combinations of possible values and their interpretation). In addition to the name and the value of the attributes, there are some restrictions (minimal occurrences, maximal occurrences) that determine whether the attribute is mandatory and how many times it can be repeated in the frame description.

Attribute	MIN	MAX
Value	Meaning	Meaning
Null	Not allowed	Unlimited
Digit X	At least X	Not more than X

Table 1: Definition of attributes, where MIN states for minimum value occurrences, and MAX – for maximum.

Additional elements describing the attributes are:

- **position** – associates the attributes and their parents in the user interface;
- **import/export** – the name of the XML / JSON element that is responsible for the data import/export;
- **code** – a system element enabling the implementation of functionality linked with a specific attribute;
- **parent object type** – a system element ensuring the hierarchical structure (parent\_obj\_type\_id).

For the value **text**, the component drop-down selection with single/multiple select options requires the definition of the list of possible values. For the value **relation**, the specific relations have to be defined. The reference to a relation is defined by the **relation type** (relation\_type\_id), which is constituted by several components:



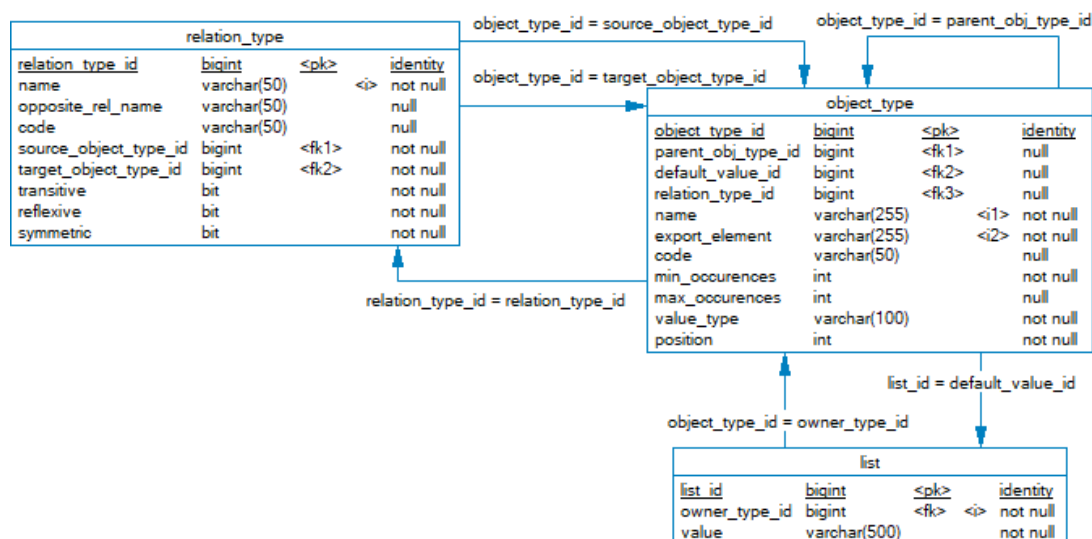


Figure 2: BulFrame database

- **name** – the name of the relation;
- **reverse relation name** – the name of the reverse relation (reverse\_rel\_name), if any;
- **source / target object** – the type of objects that are linked by the relation: either WORD or FRAME (source / target\_object\_type\_id);
- **relation properties** – transitive, reflexive, symmetric.

Figure 3 in the Appendix A shows the definition of the Conceptual frame structure within the BulFrame system.

The BulFrame allows the following actions: import verbs and Conceptual frames, edit existing entries and delete verbs with no associated Conceptual frames and Conceptual frames with no associated verbs.

#### 4 Linguistic interpretation of Conceptual frames

Conceptual frames represent the lexical meaning and morphological features of (Bulgarian) verbs which actually predict the syntactic realization and semantic combinability of their arguments (core frame elements), which are also the subject of description (Koeva, 2021). The structure of a Conceptual frame consists of the following sections: Lexical, Morphological, and Frame (Syntactic and Semantic) section.

#### 4.1 Lexical section

The Lexical section embraces the information about the verb lemmas (object WORD). The unique interpretation of a verb is ensured by its WordNet ILI (Inter-Lingual-Index) number, WordNet sense number and definition. The WordNet ILI has two purposes: it links the synonyms in a synset and shows the mapping to the respective synset (concept) from the Princeton WordNet (Vossen et al., 1998).

The Lexical section includes: verb lemma (literal), whether the verb is a multiword expression or not, part of speech, WordNet ILI to which the verb belongs, sense number, sense definition, synset semantic class, stylistic or usage note, and relations with literals from other synsets.

The verb multiverb expressions can be classified mostly as non-fixed lexicalized expressions: **reflexiva tantum se**: smeya se ‘laugh’, izpravyam se ‘stand up’; **reflexiva tantum si**: vaobrazyabvam si ‘imagine’; **reciproca tantum se**: sastezavam se ‘compete’; **reciproca tantum si**: govorya si ‘talk to oneself’; **accusativa tantum**: marzi me ‘feel lazy’; **dativa tantum**: hrumna mi ‘it occurred to me’; **reflexiva dativa tantum** gadi mi se ‘feel sick’; **with obligatory preposition(al phrase)**: privezhdam v sila ‘enforce’; **with obligatory noun (phrase)**: podavam zhalba ‘file a complaint’, davam si smetka ‘realise’; **with obligatory adverb(ial phrase)**; stoya nastrana ‘stand aside’.

**Perfective and imperfective verbs** in Bulgarian express different meanings, although the verb aspect pairs are closely related, for example the

verbs (rolya) ‘give birth’ and (razhdam) ‘am giving birth’. The definition should describe the meaning in a way that uniquely distinguishes a verb from other senses of the same word; the definition also reflects the morphological features of the verbs (for example, the limited person paradigm as third-person, impersonal and plural personal) and the lexical-grammatical category aspect.

The verbs in the (Princeton) WordNet are organized into semantic classes (primitives): generic concepts, perceived as unique roots (beginners) of separate hierarchies, and the verbs belonging to the hierarchies are subsumed under the common **semantic class**: bodily care and functions, change, cognition, communication, competition, consumption, contact, creation, emotion, motion, perception, possession, social interaction, weather verbs, state (Fellbaum, 1990). One and the same semantic class might be assigned to many verb roots and as close to the root the concept is as abstract or general its meaning is.

The note to the literal can express: the belonging to **non-standard lexis** – a dialectal word, a folk word, a word with an undesired use; use in a **specific functional style** – a colloquial word; a poetic word; a literary word; term; the **historical period of use** – an obsolete word; a new word; the **expressive properties** of the literals – a word with pejorative meaning; the **frequency of use** of the literals – a rare word; the **nuances in the use** of the literals – a figuratively used word. It should be pointed out, however, that stylistic marking usually excludes words from the core vocabulary, so although the information is intended to be transferred from WordNet, it is not expected that there will be many such cases, so far among the 639 verbs described with Conceptual frames (as of May 2022), only 4 have been marked as belonging to colloquial speech.

#### 4.1.1 Selection of Verbs

The Bulgarian verbs included in the database of Conceptual frames are selected according to several criteria.

- **Presence in Age of Acquisition (AoA) test** – the school level at which a word (meaning of a word) must be studied or mastered. The resource includes a list of 44,000 entries (31,000 words and compound words; not only verbs) compiled by Dale and O’Rourke’s Living Word Vocabulary (Dale and O’Rourke, 1981) and supplemented by estimates from other authors (Goodman et al.,

2008; Morrison et al., 1997). For example, the Age of Acquisition ratings are a self-esteem given by adults (mostly students) about the age at which they learned a word, which is also further adjusted by other assessments and experiments (Kuperman et al., 2012).

- **Root distance** – the distance of the synset to the root of the local tree (the hierarchical substructure in Wordnet in which the corresponding synonym set is included). The distance is represented by the number of nodes (synsets) between the node in which the corresponding verb is located and the respective root, a node with an abstract meaning in WordNet.

- **Presence in Base concepts** – targeting maximum overlap and compatibility across wordnets of different languages (Vossen et al., 1998). 1,024 Base Concepts are identified on the basis of English, Dutch, Spanish and Italian along the following criteria: high position in the semantic hierarchy and maximum number of relations with other concepts in the WordNet. New Base Concepts have been added of second and third batch on the basis of data from Bulgarian, Greek, Romanian, Serbian and Turkish (Tufiş et al., 2004) and the first batch of Base Concepts has been expanded to 4,689. The following additional criteria were used to identify the main concepts of the second and third batch: a) the most common words in large representative corpora and b) the hyperonyms of already selected synsets to the root of the corresponding local tree.

- **Relative frequency** – represents a) frequency of verbs in the Bulgarian National Corpus<sup>3</sup> (in the whole corpus, in fiction texts and in news); and b) frequency of verbs in Bulgarian textbooks from 1st to 4th grade. The frequency is calculated at the level of lemma; however, some noise is left due to lack of sense disambiguation.

The presented measures were evaluated by experts in order to select a proper set of basic verb vocabulary for Bulgarian:

- If the following criteria are fulfilled: the verb is part of the AoA, the distance to the root is 0 or 1 and the verb is a member of the Base concepts (batch 1 or batch 2), the verb, accompanied with its sense number, ILI record and definition, is selected.

- If the verb is present in the AoA, but the other two criteria are not met, the expert judges according to the frequency of use and his/her personal intuition.

<sup>3</sup><https://search.dcl.bas.bg>

– If the verb is not present in the AoA, the other criteria are used in the following order: root distance, member of the Base concept lists, frequency of use.

– The following principles have been also adopted during the selection: if a perfect verb is selected, the corresponding imperfect verb is also included; secondary imperfect verbs are not selected (at the moment).

#### 4.1.2 Semantic Relations

The semantic relations are inherited from WordNet and inserted in the database. Taxonomic relations are: inverse and transitive (**is hypernym of** and **has a hypernym, has a troponym and is a troponym of**); meronymic relations are also inverse and transitive (**has subevent** and **is subevent of**). Non-hierarchical relations are: symmetric, reflexive, transitive and Euclidean (**synonymy**), symmetric, irreflexive and non-transitive (**antonymy**); symmetric, irreflexive and Euclidean (**also see, verb group**). The relations in WordNet are defined between synsets. As the basic unit in our system is the verb (WORD) and in the Bulgarian wordnet verbs with a different lexical aspect are grouped in one and the same synset, the following rules are implemented while inheriting the WordNet relations (Table 2).

Verb	Hypernym	VerbG	Antonym
1: Imperfect	1,2	All	1,2
2: ImperfT	1,2	All	1,2
3: Perfect	3,4	All	3,4
4: PerfT	3,4	All	3,4

Table 2: Verb to verb semantic relations, ImperfT stands for Imperfectiva tantum, PerfT – for Perfectiva tantum, VerbG – for Verb group.

## 4.2 Morphological section

A morphological classification of a target verb is necessary because the grammatical and morphosyntactic features determine in some cases the syntactic structures associated with a given word. We have distinguished four groups of grammatical subclasses of Bulgarian verbs depending on the subject: **personal, impersonal**: *zazoryava se* ‘it dawns’, **third personal singular and plural**: *rekata se vлива v moreto* ‘the river flows into the sea’, and **plural personal**: *sabirame se okolo masata* ‘we gather around the table’.

The different meaning of **verb aspect pairs** is reflected at both the morphological and the syntactic levels: the paradigms of the perfective and imperfective verbs are different – perfective verbs do not have the so-called independent present tense, and they do not form either present participles (agentive and adverbial) or negative imperative forms; the derivational potential of the perfective and imperfective verbs is different – perfective verbs do not form some types of deverbal nouns and some nouns denoting professions; perfectivity is directly related to the syntactic realization of obligatory complements – direct objects of perfective verbs cannot remain implicit and perfective verbs cannot be complements of phase predicates; perfectivity is also directly related to the possibility for different types of verb diathesis: perfective verbs do not form middles, optatives or impersonals (Koeva, 2010).

In the Bulgarian WordNet verb aspect pairs are included in one and the same synsets, although the perfective and imperfective members of a pair are not cognitive synonymous, and as a consequence only some of the literals are translation equivalents to the respective synonyms in English. For the differentiation of verbs of different aspect, a literal note is attached to each verb indicating its aspect: **perfective verb**: (*zapeya*) ‘start singing’; **imperfective verb**: (*zapyavam*) ‘sing off’; a **simultaneously perfective and imperfective verb**: (*pensioniram*) ‘retire’; an **imperfective verb with no perfective equivalent**: (*vali*) ‘it rains’; a **perfective verb with no imperfective equivalent**: (*povyarvam*) ‘get to believe’. The values of the category verb aspect are transferred directly from WordNet.

Verbs are also classified according to their transitivity.

## 4.3 Frame section

One part of the elements in the **Frame section** are inherited from the Berkley FrameNet, another part is constructed in compliance with the FrameNet organisation and yet another one is specific for the organisation of the Conceptual frames. The FrameNet related parts are: frame name, frame definition, frame-to-frame relations, and frame elements with their names, status (core, non-core and extra-thematic) and definition.

Several types of frame-to-frame relations are defined, of which for the definition of the Conceptual

frames the important ones are: **Inheritance** (an is-a relation, the child frame is a subtype of the parent frame), **Using** (the child frame presupposes the parent frame as background); **Inchoative of** and **Causative of** (Ruppenhofer et al., 2016). Inheritance is the strongest relation between frames corresponding to an **is-a** relation in many ontologies. The basic idea of the inheritance relation is that each semantic fact about the parent must correspond to an equally specific or more specific fact about the child (Ruppenhofer et al., 2016). The origin of the information is marked: inherited from FrameNet; from FrameNet with modifications; completely new information.

So far, 104 different semantic frames were used as basic structures for defining Conceptual frames. 105 unique frame elements were used, among which the most frequently selected are: **Agent** – 175 instances, **Experiencer** – 81 instances, **Cause** – 66 instances, **Stimulus** – 57 instances and so on. Together with the frame elements that can be encountered in different semantic frames, there are 30 cases of rare use of a particular frame element – 1 or 2 times. For example, frame elements **Intervention**, **Medical condition**, **Medical professional** and **Result** are so far selected only once.

#### 4.3.1 Frame element Syntactic Structure

The phrases that express the frame elements may be **obligatorily explicit** (in rare cases in Bulgarian) or **non-explicit**, which means that the potential for a syntactic realization of the phrase exists, but its explicitness is not mandatory because it is understood from the context in a broader sense (verb morphology, preceding text, extralinguistic information, etc.), a special case is **pronominal drop** in the subject position.

The **syntactical phrases** that can be candidates for arguments in Bulgarian are: **NP** (noun phrase), **PP** (preposition phrase), **AdvP** (adverb phrase), **S** (sentence), **SC** (small clause), **ACCCL** (obligatory accusative clitic), **DATCL** (obligatory dative clitic). For a single verb with a unique meaning, there might be more than one combination of obligatory environments. Each personal verb incorporates an argument – a noun phrase (NP) or a sentence (S) that are realized as the subject in the sentence. The subject may not be explicitly stated – with personal verbs the information for person and number of the omitted pronoun subject is expressed by the verb inflexion.

The frame elements related to the subject of Bul-

garian verbs can be characterized as follows: (a) with an explicitly or implicitly expressed subject with a full paradigm of the category of person; (b) with an explicitly or implicitly expressed third-person subject; (c) with no subjective argument. The frame elements related to the complements of Bulgarian verbs can be classified as follows: (a) with a single NP complement; (b) with an NP complement and an S complement; (c) with an NP complement and PP complements, regardless of their number; (d) with an NP complement, PP complements, regardless of their number, and an S complement; (e) with PP complements, regardless of their number; (f) with PP complements, regardless of their number, and an S complement; (g) with an S complement; (h) with an AdvP predicate modifier; (i) With an SC (small clause) NP argument; (j) with an SC (small clause) PP argument; (k) with an SC (small clause) AP argument; (l) with no complements.

The **syntactic functions** (names of syntactic positions taken from traditional grammar) are subject, direct object, indirect object, adverbial, subject clause, object clause, adverbial clause and small clause. The syntactic structure is described by information about the phrases: explicitness (checkbox), syntactic category (checkbox) and syntactic function (checkbox).

#### 4.3.2 Frame element Semantic Structure

FrameNet allows for the characterization of ‘role fillers’ by semantic types of frame elements, which ought to be broadly constant across uses (Ruppenhofer et al., 2016). However, not all frame elements are supplied with a semantic type or the semantic types are too general, and in some cases, they do not show the actual restrictions for lexical combinations. For example, the following frame elements of the semantic frame **Experiencer focused emotion** are equipped with semantic types:

**Content** with the semantic type [Content]; **Event** with the semantic type [State of affairs]; **Experiencer** with the semantic type [Sentient]; **Degree** with the semantic type [Degree]; **Explanation** with the semantic type [State of affairs]; **Manner** with the semantic type [Manner]; **Time** with the semantic type [Time].

We call selectivity the **semantic restrictions** to a given argument in a certain context selectivity. Due to the fact that selective restrictions act between a concrete predicate and the arguments that belong to it, they can be different for each separate case. The



most general semantic classification distinguishes among **abstract** and **concrete** nouns. On their part, concrete nouns can be **animate** or **inanimate**. Animate nouns may be classified as **persons** and **non-persons**, **persons** as **agents** or **experiencers**. This classification tree is convenient but too shallow to represent the selective restrictions that act with verbs and nouns. Besides the general cases, there may also be cases where concrete restrictions are required, as for example **liquid**, **food**, etc. That is why we include the link to the top most synset (or the conjunction/disjunction of top most synsets) taken from the Bulgarian WordNet. The top most synset should dominate all appropriate synsets for a given syntactic slot, i.e., **liquid is a hypernym of water, milk, liquor**, etc.

The semantic classes of nouns in WordNet might be subdivided into a set of semantic subclasses. For example, within the semantic class **[Food]** we can introduce the sub-class of **[Beverage]** for nouns associated with verbs like **stir**, **sip**, **drink**, **lap**, etc. Such representation aims to specify the organization of concepts into an **ontological structure** which allows inheritance between the semantic classes down the hierarchy and ensures more precise specification of verb – noun compatibility.

One potential to extend the repository of WordNet semantic classes is to map the WordNet synsets to an existing hierarchy of semantic types, such as the CPA types (Hanks, 2004). The semantic types (e.g. **[Human]**, **[Animal]**, **[Part]**, etc.) refer to properties which can be expressed by words regularly found to participate in particular verb pattern positions (Hanks 2012: 57–59). In other words, the semantic types state the semantic preferences of verbs that determine the sets of nouns and noun phrases that are normally found in a particular clause role depending on a verb predicate.

Some verb patterns may contain very general preferences, i.e., the semantic type **[Anything]**, while others impose preferences for a limited set of lexical units grouped into more particular semantic types. For example, some verbs are associated with nouns characterised as **[Body part]**. However, the verb **shampoo** is associated with a more particular semantic type **[Hair]**; the same is referred to the verb **nod**, which is associated with the type **[Head]**, etc. Some verb patterns require a very small set of lexical units for a particular slot and in this case, a semantic type is not formulated; instead, the concrete lexical units are listed in the verb pattern. The

expansion of WordNet semantic classes with CPA semantic types is performed manually by matching the CPA semantic types with WordNet synsets and choosing the most appropriate ones (Koeva et al., 2018).

The 253 CPA semantic types are manually mapped onto the respective WordNet concepts (synsets) as follows: 199 semantic types are mapped directly to one concept, i.e., **[Permission]** is mapped to **permission** ‘approval to do something, semantic class noun communication’; **[Dispute]** is mapped to **disagreement** ‘the speech act of disagreeing or arguing or disputing’, semantic class noun communication; 39 semantic types are mapped to two WordNet concepts, i.e., **[Route]** is mapped to **road; route** ‘an open way (generally public) for travel or transportation, semantic class noun artefact, and **path; route** ‘an established line of travel or access’, and semantic class noun location, and so on. Automatic mapping of hyponym synsets to the inherited semantic types was performed. In the cases where a semantic type and its ancestor were both mapped to the same synset, the ancestor was removed. 82,114 WordNet noun synsets were mapped to the 253 semantic types of the CPA ontology, resulting in 172,991 mappings. As there are multiple hypernymy relations in WordNet, some of the inheritances are not correct; furthermore, the inheritance by multiple hypernyms will be manually evaluated, and if necessary, adjusted (Koeva et al., 2018).

Some of the initially selected classes were not chosen as dominant classes for nouns compatible with particular verbs, for example the class **[Plant]** (eng-30-00017222-n), the class **[Abstract object]** (eng-30-00019128-n), and so on. This obviously is a consequence of the selections of the verbs. On the other hand, 84 unique selective restrictions were used identifiable by a representative noun and its ILI number. Some new classes were introduced (28 altogether, which constitutes 35,7 % of the total number of classes used so far. For example, new classes are: **[Text]** (eng-30-06387980-n), **[Examination]** (eng-30-07197021-n), **[Fire]** (eng-30-07302836-n), and so on. Still, the abstract notions show more instances in the dataset: **[Person]** — selected 850 times, **[Entity]** — selected 249 times, **[Object]** — selected 175 times, **[Physical object]** — selected 170 times and so on.

The concrete prepositions for a given frame element expressed with a prepositional phrase are to

be selected from a list box. The same holds for frame elements that express the obligatory noun (phrase) or adverb (phrase). The types of subordinate clauses depend on the method of linking – interrogative pronouns or conjunctions, thus the respective linking phrase or complementizers are to be selected (more than one choice is permissible).

## 5 Related work

**FrameNet** is the most famous language resource that contains lexical and conceptual knowledge (Ruppenhofer et al., 2016). FrameNet can be viewed as a semantic network (or a set of small semantic nets), whose nodes indicate the semantic frames and whose arcs represent semantic relations between frames. For the purposes of the presented research, the following information is employed: the sets of verb lexical units related to semantic frames, the inheritance relation between semantic frames, and the description of core and peripheral frame elements and their semantic types. The FrameNet annotation is mostly used for automatic role labelling while we offer the definition of noun sets compatible with verbs from a particular Conceptual frame (and such approach offers much more training data for automatic processing). FrameNets for languages other than English are being developed, including for Bulgarian.

**VerbAtlas** is a relatively new, hand-crafted lexical-semantic resource, whose goal is to bring together WordNet verbal synsets into semantically-coherent frames (Di Fabio et al., 2019). The frames define a common, prototypical argument structure, while at the same time provide new concept-specific information. VerbAtlas also offers an explicit, cross-frame set of semantic roles linked to selectional preferences expressed in terms of WordNet synsets, and is the first resource enriched with semantic information about implicit, shadow and default arguments. The main difference between the VerbAtlas and the presented framework is that the VerbAtlas selectional preferences are too general, similarly to the semantic types of core elements in FrameNet, in comparison to the extensive semantic information provided within the BulFrame to ensure accurate noun-to-verb compatibility.

Some efforts to describe Bulgarian frame lexicon were also shown, and we believe our work draws on the best approaches in the field.

## 6 Conclusion and Future work

The presentation of Conceptual frames of Bulgarian verbs provides opportunities for the enrichment of already existing resources (Wordnet and Framenet) with new semantic information (in the direction of completeness and structural expansion), developing a detailed ontology of the semantic classes of nouns and linking it to the hierarchical structure of WordNet and the frame elements of FrameNet.

The main characteristic of the approach we have taken is the manner of connecting FrameNet and WordNet – not by assigning frames to synsets, i.e., not in the usual way, but by showing which WordNet subtrees are suitable to take one or another syntactic position in which a frame element is realized. The morpho-syntactic features that are specific for Bulgarian are shown in detail; selective restrictions are specified so that the resource can be used for automatic prediction of semantic roles in any text.

As future work, we plan to take full advantage of the semantic features available in BulFrame, such as wide-coverage selectional preferences and verb level grammatical information, by employing them in semantic role labelling tasks.

## Acknowledgments

This research is carried out as part of the project *Enriching Semantic Network WordNet with Conceptual frames* funded by the Bulgarian National Science Fund, Grant Agreement No. KP-06-H50/1 from 2020.

## References

- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. [The berkeley FrameNet project](#). In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.
- Edgar Dale and Joseph O’Rourke. 1981. *The Living Word Vocabulary: A National Vocabulary Inventory*. World Book-Childcraft International, Chicago.
- Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. [VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong Kong, China. Association for Computational Linguistics.



- Christiane Fellbaum. 1990. [English Verbs as a Semantic Net](#). *International Journal of Lexicography*, 3(4):278–301.
- Charles J. Fillmore. 1982. *Frame semantics*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.
- Charles J. Fillmore and Collin F. Baker. 2001. [Frame Semantics for Text Understanding](#). In *Proceedings of WordNet and Other Lexical Resources Workshop*, Pittsburgh. NAACL, NAACL.
- Judith C. Goodman, Philip S. Dale, and Ping Li. 2008. [Does frequency count? Parental input and the acquisition of vocabulary](#). *Journal of Child Language*, 35(3):515–531.
- Patrick Hanks. 2004. [Corpus pattern analysis](#). In *Proceedings of the 11th EURALEX International Congress*, pages 87–97, Lorient, France. Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.
- Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2007. [A large-scale classification of English verbs](#). *Language Resources and Evaluation*.
- Svetla Koeva. 2010. *Bulgarian FrameNet*. Prof. M. Drinov Academic Publishing House, Sofia.
- Svetla Koeva. 2020. [Semantic Relations and Conceptual Frames](#). In Svetla Koeva, editor, *Towards a Semantic Network Enriched with a Variety of Semantic Relations*, pages 7–20. Sofia: Professor Marin Drinov Publishing House of BAS.
- Svetla Koeva. 2021. [Towards Expanding WordNet with Conceptual Frames](#). In *Proceedings of the 11th Global Wordnet Conference*, pages 182–191, University of South Africa (UNISA). Global Wordnet Association.
- Svetla Koeva, Cvetana Dimitrova, Valentina Stefanova, and Dimitar Hristov. 2018. [Mapping WordNet concepts with CPA ontology](#). In *Proceedings of the 9th Global Wordnet Conference*, pages 69–76, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.
- Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. [Age-of-acquisition ratings for 30,000 English words](#). *Behavior research methods*, 44(4):978–990.
- Beth Levin. 1993. *English Verb Classes and Alternations A Preliminary Investigation*. University of Chicago Press, Chicago and London.
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. [Introduction to WordNet: An On-line Lexical Database](#). *International journal of lexicography*, 3(4):235–244.
- Catriona M. Morrison, Tameron D. Chappell, and Andrew W. Ellis. 1997. [Age of Acquisition Norms for a Large Set of Object Names and Their Relation to Adult Estimates and Other Variables](#). *The Quarterly Journal of Experimental Psychology Section A*, 50(3):528–559.
- Roberto Navigli and Simone Paolo Ponzetto. 2012. [BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network](#). *Artificial Intelligence*, 193(0):217 – 250.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. [The Proposition Bank: An annotated corpus of semantic roles](#). *Computational Linguistics*, 31(1):71–106.
- Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, , Collin F. Baker, , and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California.
- Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. [Yago: A Core of Semantic Knowledge](#). In *Proceedings of the 16th International Conference on World Wide Web, WWW '07*, pages 697–706, New York, NY, USA. ACM.
- Dan Tufiş, Dan Cristea, and Sofia Stamou. 2004. [BalkanNet: Aims, methods, results and perspectives. a general overview](#). *Romanian Journal of Information Science and Technology*, 7(1-2):9–43.
- Zdenka Uresova, Eva Fucikova, Eva Hajicova, and Jan Hajic. 2020. [SynSemClass linked lexicon: Mapping synonymy between languages](#). In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 10–19, Marseille, France. European Language Resources Association.
- Piek Vossen, Laura Bloksma, Horacio Rodriguez, Salvador Climent, Nicoletta Calzolari, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, and Wim Peters. 1998. [The EuroWordNet base concepts and top-ontology](#). Technical report, GPLN - Grup de Processament del Llenguatge Natural and TALP - Centre de Tecnologies i Aplicacions del Llenguatge i la Parla.
- Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin Sameer Pradhan Lance Ramshaw, and Nianwen Xue. 2011. [Ontonotes: A large training corpus for enhanced processing](#). *Joseph Olive, Caitlin Christianson, and John McCary, editors, Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*.

## **Appendix A Conceptual frame structure (some parts of the structure are not presented)**

Object/Characteristic	Type
<ul style="list-style-type: none"> <li> <span style="font-size: 1em;">▼</span>  Frame           </li> </ul>	Text
<ul style="list-style-type: none"> <li>  Lemma           </li> </ul>	Relation to word
<ul style="list-style-type: none"> <li>  Frame name           </li> </ul>	Text
<ul style="list-style-type: none"> <li>  Origin           </li> </ul>	Single choice from list
<ul style="list-style-type: none"> <li>  Semantic type           </li> </ul>	Single choice from list
<ul style="list-style-type: none"> <li>  Definition           </li> </ul>	Text
<ul style="list-style-type: none"> <li>  Relation Inheritance           </li> </ul>	Relation to frame
<ul style="list-style-type: none"> <li>  Relation Uses           </li> </ul>	Relation to frame
<ul style="list-style-type: none"> <li>  Relation Inchoative of           </li> </ul>	Relation to frame
<ul style="list-style-type: none"> <li>  Relation Causative of           </li> </ul>	Relation to frame
<ul style="list-style-type: none"> <li> <span style="font-size: 1em;">▼</span>  Frame element           </li> </ul>	Node
<ul style="list-style-type: none"> <li>  Element name           </li> </ul>	Autocomplete text
<ul style="list-style-type: none"> <li>  Status           </li> </ul>	Single choice from list
<ul style="list-style-type: none"> <li>  Semantic type           </li> </ul>	Autocomplete text
<ul style="list-style-type: none"> <li>  Element definition           </li> </ul>	Autocomplete text
<ul style="list-style-type: none"> <li>  Syntactic obligatoriness           </li> </ul>	Single choice from list
<ul style="list-style-type: none"> <li> <span style="font-size: 1em;">▼</span>  Syntactic category           </li> </ul>	Node
<ul style="list-style-type: none"> <li> <span style="font-size: 1em;">▼</span>  NP           </li> </ul>	Node
<ul style="list-style-type: none"> <li>  Syntactic function           </li> </ul>	Single choice from list
<ul style="list-style-type: none"> <li>  Selective restrictions           </li> </ul>	Relation to word

Figure 3: Conceptual frame structure

# Linked Resources towards Enhancing the Conceptual Description of General Lexis Verbs Using Syntactic Information

**Svetlozara Leseva**

Institute for Bulgarian Language  
Bulgarian Academy of Sciences  
zarka@dcl.bas.bg

**Ivelina Stoyanova**

Institute for Bulgarian Language  
Bulgarian Academy of Sciences  
iva@dcl.bas.bg

## Abstract

The paper describes the linking of three previously aligned resources (FrameNet, VerbNet and WordNet) both by expanding their coverage (by means of enhancing existing alignments) and by mapping elements of the semantic and syntactic description of the lexical items: FrameNet frame elements and VerbNet semantic roles, FrameNet valency patterns and VerbNet syntactic patterns. The study focuses on general lexis verbs as being more representative across languages. After describing the used resources and their interaction, we go on to outline the mapping procedures and the elements of the resulting resource. The discussion sums up the main challenges encountered in carrying out the described tasks.

**Keywords:** linked resources, FrameNet, VerbNet, WordNet, semantic description, syntactic patterns

The paper deals with linking complementary semantic and syntactic resources (FrameNet, VerbNet and WordNet) through aligning relevant elements of their semantic and syntactic description. We take as a point of departure previously made alignments between these resources where WordNet synsets or synset members have been assigned a FrameNet frame and/or a VerbNet class on the basis of equivalent or similar meaning. Through its membership to a class or a frame a lexical unit (in this case a verb belonging to a verb synset in WordNet) inherits the semantic and syntactic description associated with them. While the syntactic and semantic knowledge from FrameNet and VerbNet informs a rich linguistic description associated with each verb, there are a number of challenges to this approach: such a description includes a lot of information couched in different terms which on the one hand may be redundant and on the other does not provide linking between corresponding elements of meaning; such elements include: the semantic roles (SR)

describing the argument structure of each verb in VerbNet, the frame elements (FEs) in FrameNet, part of which (roughly speaking the core FEs) represent VerbNet SR counterparts; the selectional restrictions defined for semantic roles and the relevant semantic types of FrameNet FEs; the syntactic patterns that are associated with the contextual realisations of the verbs in the two resources.

In this work we describe a linked resource in which not only lexical units are aligned but also the elements of the semantic and the syntactic description associated with them. We use an aligned version of VerbNet and FrameNet and propose a methodology for linking semantic and syntactic knowledge in the resources so as to reduce redundant information and make the best use of both of them. We focus on general lexis verbs selected from WordNet using various criteria.

## 1 Related Work

Significant efforts have been invested in aligning and in some cases expanding the mapping between semantic and syntactic resources in the past two decades and this interest has been growing in recent years. A number of proposals have brought together the advantages of conceptual and lexical information encoded in resources such as WordNet, FrameNet, VerbNet and others. Such works include the mapping of WordNet, FrameNet and VerbNet by [Shi and Mihalcea \(2005\)](#), the elaboration of WordFrameNet<sup>1</sup> by [Laparra and Rigau \(2010\)](#) and MapNet<sup>2</sup> by [Tonelli and Pighin \(2009\)](#), the implementation of other FrameNet-to-WordNet mappings, such as the one by [Ferrández et al. \(2010\)](#). More enhanced linked resources include Semlink<sup>3</sup> ([Palmer, 2009](#)), which unifies WordNet, FrameNet

<sup>1</sup><http://adimen.si.ehu.es/web/WordFrameNet>

<sup>2</sup><https://hlt-nlp.fbk.eu/technologies/mapnet>

<sup>3</sup><https://verbs.colorado.edu/semliink/>

and VerbNet with PropBank, and its follow-up Semlink+ that brings in a mapping to Ontonotes (Palmer et al., 2014).

More recently, the SynSemClass lexicon<sup>4</sup> has marked a distinguishable effort towards combining the rich semantic description in the Vallex dictionary family with conceptual and syntactic information from external semantic resources in order to create a multilingual contextually-based verb lexicon. The aim of the lexicon is to provide a resource of classes of verbs that compares their semantic roles as well as their syntactic properties (Urešová et al., 2020a). In addition, each entry is linked to FrameNet, WordNet, VerbNet, OntoNotes and PropBank, as well as the Czech VALLEX.

VerbAtlas<sup>5</sup>, proposed by Di Fabio et al. (2019), is a hand-crafted lexical semantic resource which represents synsets as clusters with prototypical argument structures presented as frames, to a large extent inspired by VerbNet roles and semantic restrictions.

One of the main concerns related to resource alignment has been the limited coverage. Hence, another line in the research on semantic resource linking has been the expansion of inter-resource coverage. Burchardt et al. (2005) have proposed a method for enriching FN frame-to-WN synset alignment based on exploring the structural features of the two resources. In particular, they study candidate frames evoked by literals (individual members of synsets) related to a target literal through certain semantic relations, such as synonymy, hypernymy, antonymy, and assign weights to them according to the adopted methodology.

Di Fabio et al. (2019) adopt a strategy of clustering WordNet synsets according to semantic similarity and associating them with frames that describe the predicate-argument structure and selectional restrictions of each cluster. While these frames are inspired by VerbNet, the clustering algorithm achieves much better coverage of WordNet synsets as compared with WordNet-VerbNet mappings relying on the lexical correspondence of the units in the two resources.

Another proposal for expansion of the mapping between FrameNet and WordNet proposed by Leseva et al. (2018b) and further refined in Leseva and Stoyanova (2019) makes use of the relational structure of the two resources. The method involves the

mapping of FrameNet frames to WordNet synsets on the basis of the inheritance of conceptual features in hypernym trees, i.e., by assigning frames from hypernyms to hyponyms where possible and implementing a number of validation procedures based on the structural properties of the two resources, primarily the relations encoded in them.

Another venue of research has been to map relevant information representing fragments of meaning associated with lexical units across resources, especially essential components of the semantic and the syntactic description such as semantic roles or their counterparts in the respective resources (e.g. frame elements, argument positions, valency slots). Alignments at the verb arguments' level have been carried out as part of the Semlink project and its more recent version Semlink 2.0. (Stowe et al.). The alignments described there include PropBank to VerbNet mappings (PropBank roleset – VerbNet senses, PB arguments – VerbNet semantic roles) as well as VerbNet to FrameNet mappings (VerbNet senses – FrameNet frames, VerbNet semantic roles – FrameNet frame elements). Another similar task, which makes use of the linking of various semantic resources (FrameNet, WordNet, VerbNet, OntoNotes and PropBank), has been implemented in the development of the SynSemClass Lexicon (Urešová et al., 2020a,b): the more general SynSemClass valency slots have been mapped to relevant FrameNet frame elements.

In this paper we build upon previous efforts in aligning and expanding the coverage of semantic resources by mapping semantic and syntactic elements of the description of their basic units, in particular: FrameNet frame elements and VerbNet semantic roles (along with the selectional restrictions defined for them in the two resources) and the syntactic patterns associated with the verbs in the respective FrameNet frame and VerbNet class. Instead of using it directly, we employ the mapping provided in Semlink and Semlink+ (Palmer, 2009; Palmer et al., 2014) as a reference set to compare to our own mapping for a couple of reasons: some classes are only marginally corresponding to a given frame so the alignment needs to be considered more carefully; as FrameNet's and VerbNet's descriptions do not always correspond straightforwardly, the semantic roles and frame elements may have one-to-many or many-to-many mappings or actually not be counterparts of each other despite a seeming coincidence or similarity in the names or

<sup>4</sup><https://ufal.mff.cuni.cz/synsemclass>

<sup>5</sup><http://verbatlas.org/>



definitions. The availability of a mapping to compare our independent results will make the analysis of debatable cases more reliable.

In addition, we map the syntactic patterns capturing the expression of the semantic roles for the verb classes in VerbNet and the valency patterns for the verbs in a given frame in FrameNet. This procedure is aimed at providing an additional syntactic level of comparison between the two resources that may inform studies and applications both for English and for other languages. The syntactic correspondences are also applicable in semantic role – frame element mapping or mapping validation procedures, especially in cases where the semantic roles and the frame elements are not successfully mapped but have equivalent syntactic expression.

We consider two main research questions:

1. How can we integrate semantic and syntactic information to enhance the conceptual description of WordNet synsets and literals?
2. To what extent is syntactic information language independent and can it be transferred from English to less-resourced languages such as Bulgarian?

The contributions of the paper include:

- Mapping of VerbNet classes and their roles and FrameNet frames and frame elements – although initially relying on existing alignments, we extend them using additional mappings of FrameNet to WordNet which allows us to expand the dataset;
- Enhancing conceptual description of WordNet synsets and literals with syntactic patterns facilitates tasks such as syntactic and semantic parsing and semantic role labelling;
- Mapping of general (largely language-independent) and (language-)specific syntactic patterns for Bulgarian and English allows for cross-linguistic analyses, transferring valid patterns and adapting them for low-resourced language such as Bulgarian with limited resources on valency and syntactic realisation of conceptual frames.

## 2 Resources

Below we describe in brief the used resources and how they are integrated with each other.

### 2.1 WordNet

WordNet<sup>6</sup> (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 cognitive synonyms (synsets) linked by means of a number of conceptual-semantic and lexical relations such as hypernymy, meronymy, antonymy, etc. Of the three resources employed in this work, WordNet provides the greatest lexical coverage; the verbs represented in it are organised in 14,103 synsets. We use both the Princeton WordNet and the Bulgarian WordNet, which are aligned at the synset level by means of unique synset identifiers.

### 2.2 FrameNet

FrameNet<sup>7</sup> (Baker et al., 1998; Baker, 2008) is a lexical semantic resource which couches lexical and conceptual knowledge in the terms of frame semantics. Frames are conceptual structures describing types of objects, situations, or events along with their components (frame elements) (Baker et al., 1998; Ruppenhofer et al., 2016). Depending on their status, frame elements (FEs) may be core, peripheral or extra-thematic (Ruppenhofer et al., 2016). We deal primarily with core FEs, which instantiate conceptually necessary components of a frame, and which in their particular configuration make a frame unique and different from other frames.

### 2.3 VerbNet

VerbNet (Kipper-Schuler, 2005; Kipper et al., 2008) is a hierarchical network of English verbs which represents their syntactic and semantic patterns<sup>8</sup>. It is organised into 274 classes extending Levin’s classification (Levin, 1993) through refining and adding subclasses so as to provide better syntactic and semantic coherence among members of a class. VerbNet explicitly projects semantic relations onto syntactic structures and encodes information about thematic roles, arguments’ selectional restrictions and syntactic frames. While the syntactic dimension of the resource is more specific to English, the semantic roles and the selectional restrictions employed provide well-motivated semantic generalisations.

---

<sup>6</sup><https://wordnet.princeton.edu/>

<sup>7</sup><https://framenet.icsi.berkeley.edu/fndrupal/>

<sup>8</sup><https://verbs.colorado.edu/verbnet/>

Besides the rich lexical description (glosses, examples, semantic primitive) and the encoded relations, WordNet’s main contribution to this work is the rich lexical coverage of verbs, including information about the membership of synsets to the so-called base concepts – a cross-lingual selection of synsets which we use as an approximation (together with other selection criteria) for establishing a set of general lexis verbs. Our focus on general lexis stems from the interest in studying the semantic and syntactic (a)symmetries in the vocabulary cross-linguistically. While we use wordnets for English and Bulgarian, any available wordnets for other languages (aligned at the synset level) can be used instead as at least the semantic components and for a number of languages – a part of the syntactic component may be used both for monolingual and comparative/contrastive research and applications.

FrameNet and VerbNet bring in rich semantic description in terms of aligned inventory of: (i) frames, frame elements and semantic restrictions associated with FN lexical units and detailed valency patterns representing the syntactic realisation of the frame elements for each verb (in the form of annotated sentences); (ii) verb classes, predicate-argument structures (in the form of semantic role configurations), selectional restrictions and syntactic patterns realising the arguments of the verbs pertaining to the classes defined in the VerbNet lexicon. In implementing the task of aligning the lexical items in FrameNet and VerbNet we focus particularly on mapping core frame elements as they are most likely to represent a verb’s arguments and hence – counterparts of the semantic roles. Differences between frames’ core FEs sets and corresponding predicate argument structures reveal valuable language- and resource-specific features of the semantic and syntactic description.

As we use an expanded synset-to-frame mapping between WordNet and FrameNet (Leseva et al., 2018b; Leseva and Stoyanova, 2019), the number of verbs associated with a FN frame and all the information pertaining to it is larger than in the original mappings. An interesting research question to be tackled in the future is to what extent the indirectly aligned WordNet verbs (especially ones that do not correspond to a lexical unit in FrameNet or VerbNet) may be satisfactorily described semantically and syntactically by means of the information already available in the mapped

resources.

### 3 Dataset Compilation

The three resources have been mapped automatically using existing mappings or newly designed procedures in such a way that WordNet synsets are assigned corresponding verb classes from VerbNet and frames from FrameNet where possible. The previously implemented mappings have been supplemented and partially validated. In particular, the following have been employed: a mapping of the VerbNet 3.4 verb classes to WordNet synsets, as well as two types of mappings of the frames in FrameNet and the synsets in WordNet: indirectly via SemLink and directly through the system described by Laparra and Rigau (2010). In addition, in order to increase the inter-resource coverage between WN synsets and FN frames, we have used an expanded synset-to-FrameNet frame mapping described in detail in (Leseva and Stoyanova, 2020).

The focus of the study are general lexis verbs in WordNet. We are aiming at compiling a lexical resource of verbs of high frequency and wide usage supplied with conceptual description and syntactic frames. The main source of the description is the information from the FrameNet frame and VerbNet class aligned to the WordNet synset. The resource will serve as a model and can be further expanded to cover other verbs.

#### 3.1 General Lexis Verbs and their Representation in WordNet

First, we identify verbs in WordNet that potentially belong to the general lexis using several criteria:

- verbs labelled as base concepts (BCS) in WordNet;
- verbs with high frequency in the Bulgarian National Corpus (considering the usage of all their senses);
- verbs identified in primary school textbooks in Bulgarian;
- verb senses included in Concepticon;
- verb senses marked with age of acquisition in primary school age;
- verb synsets that have been assigned FrameNet frames with high frequency (50+ verified occurrences assigned to WordNet



synsets), which in most cases have general meaning.

Base Concepts<sup>9</sup> were introduced within the WordNet research framework (Vossen et al., 1998) as the building blocks for constructing wordnets for different languages. Base Concepts typically satisfy two main criteria: a high position in the semantic hierarchy and having many relations to other concepts. WordNet synsets lexicalising Base Concepts are therefore among the likeliest candidates for general lexis.

The Bulgarian National Corpus (Koeva et al., 2012) consists of 5.4 billion words (1.2 billion for Bulgarian) and represents the lexis of contemporary Bulgarian. For our purposes, we extracted verbs with high frequency (over 10 per million words) that are found across different domains, text types and genres (at least two different domains, one of them being either fiction or news articles).

We also cross-checked the identified verbs against a small corpus of primary school textbook texts (for children aged 7 to 11 years old) in 5 different subjects. Verbs of high frequency appearing in textbooks in at least two subjects are deemed to belong to the general lexis. For verbs occurring in more than one synset we have manually selected the more general and frequent senses (based on human expert evaluation).

Concepticon is an open-source online lexical database of linguistic concepts which links concept labels from 160 concept lists (compiled from various sources and for various purposes) to 2495 concept sets (structured by defining different relations between the concepts) (List et al., 2016)<sup>10</sup>. In essence, it is a concept meta-resource which is applicable across various languages and is also linked to lexical-semantic resources such as WordNet and BabelNet.

Kuperman et al. (2012) present a data resource of 30,000 English words labelled with age of acquisition (AoA) information. The initial list was compiled by selecting base words (lemmas) appearing with high frequency in an English corpus of movie and TV series subtitles and the AoA data was collected using web-based crowdsourcing. The data includes the mean AoA ratings (in years of age) and standard deviations (attesting to the reliability of judgement), as well as the number of respondents who gave ratings to the word. Addi-

tionally, we have lemmatised the AoA word list, extracted the verbs rated with AoA of up to primary school age (up to the age of 11) and matched them to WordNet synsets.

Finally, we have identified FrameNet frames and the corresponding VerbNet classes that have a high coverage in terms of WordNet synsets (synsets assigned the respective frames and/or classes) as established in the extended inter-resource mappings used in this study. The assumption is that the most populated frames and classes represent the general part of the lexicon.

Using the criteria above, we have identified a dataset of 4,927 verb synsets of which: (a) 2,362 belong to the category of base concepts; (b) 1,800 have a high frequency in the Bulgarian National Corpus (frequency of 200+ counted as accumulative frequency of all literals in the synsets across all of their possible senses); (c) 1,470 synsets whose literals appear in primary school textbooks (frequency of 20+ in the textbook collection counting all occurrences of the synset literals across all of their possible senses); (d) 322 are included in Concepticon; (e) 252 have age of acquisition in primary school years; (f) 1,844 verb synsets have been assigned a high frequency frame. 1,405 synsets (28.5% of the dataset) are confirmed by at least 3 of the features, 212 (4.3%) are confirmed by 4 or more, which shows that the features are complementary for the purpose of general vocabulary extraction.

### 3.2 Conceptual Description of General Lexis Verbs

Currently, we focus on a set of frames and their corresponding verb classes to build a uniform model for conceptual description that can be expanded both in size and in terms of description features later on.

The efforts to align different lexical-semantic resources aim at combining various information into an extensive complex representation of the lexical units (in our case verbs) and the description of the main participants in the corresponding conceptual frame.

We consider the WN synsets with their assigned FN frame and VN class. Each synset is characterised by a pair of a frame and a verb class. As shown in Example 1 (a-c), for different synsets a frame can be corresponding to a number of verb classes, e.g. the frame *Body\_movement* can corre-

<sup>9</sup><http://globalwordnet.org/resources/gwa-base-concepts/>

<sup>10</sup><https://concepticon.clld.org/>

spond to the verb class *crane-40.3.2* (with explicit body part participating in the movement), the verb class *curtsey-40.3.3* where the body part is incorporated in the verb’s meaning, or the verb class *modes\_of\_being\_with\_motion-47.3* where the movement concerns the whole body. Although in general verb classes are more concrete than frames, there are also cases where a number of frames are linked to a single verb class, as in Example 1 (c-d), hence the frame-to-verb class correspondence is ‘many-to-many’.

**Example 1.** Alignment between FrameNet frames and VerbNet classes.

(a) **WordNet synset:** *eng-30-00145902-v purse* ‘contract one’s lips into a rounded shape’

**FrameNet frame:** *Body\_movement*: Agent (Sentient); Body\_part (Body\_part)

**VerbNet class:** *crane-40.3.2*: Agent [+animate]; Patient [+body\_part]; Topic; Recipient [+animate]

(b) **WordNet synset:** *eng-30-02040549-v curtsey*; *curtsey* ‘bend the knees in a gesture of respectful greeting’

**FrameNet frame:** *Body\_movement*: Agent (Sentient); Body\_part (Body\_part)

**VerbNet class:** *curtsey-40.3.3*: Agent [+animate]; Topic; Recipient [+animate]

(c) **WordNet synset:** *eng-30-01865383-v bob* ‘move up and down repeatedly’

**FrameNet frame:** *Body\_movement*: Agent (Sentient); Body\_part (Body\_part)

**VerbNet class:** *modes\_of\_being\_with\_motion-47.3*: Agent [+int\_control]; Theme [+concrete]; Location [+location & -region]

(d) **WordNet synset:** *eng-30-01868258-v waver*, *wave* ‘sway to and fro’

**FrameNet frame:** *Self\_motion*: Self\_mover (Sentient); Area (Location) — Source (Source); Path (Path); Goal (Goal); Direction

**VerbNet class:** *modes\_of\_being\_with\_motion-47.3*: Agent [+int\_control]; Theme [+concrete]; Location [+location & -region]

We have identified 96 pairs of FN frame and VN verb class assigned to 2,016 verb synsets. Out of the pairs only 12 have identically named frame and verb class, which suggests close correspondence (e.g., *Escaping* – *escape-51.1*, *Filling* – *fill-9.7*, *Destroying* – *destroy-44*, etc.). There are 20 frames mapped to more than one verb class, out of which 11 frames are mapped to 3 or more verb classes

each.

### 3.3 Alignment between Semantic Roles and Frame Elements

The challenges to the mapping of frame elements and semantic roles stem from several sources: (i) differences in the conceptualisation of the situations between frames and verb classes; (ii) differences in the status of the frame elements and semantic roles (not all core elements necessarily have a semantic role counterpart and vice versa); (iii) differences in the syntactic description across the resources, etc.

**Example 2.** FrameNet frame *Escaping* aligned to VerbNet class *escape-51.1*.

**WordNet synset:** *eng-30-02074677-v escape*; *get away*; *break loose*

**FrameNet frame:** *Escaping*

**Core FN FEs:** *Escapee* (Semantic Type: Animate\_being); *Undesirable\_location* (Semantic Type: Source)

**VerbNet class:** *escape-51.1*

**VN roles:** *Theme*; *Initial\_location*; *Destination*; *Trajectory*

FN element and status	Semantic type	VN role	VN restriction
Escapee	Animate_being	Theme	[concrete +]
Goal	Goal	Destination	[concrete +]
Means	State_of_affairs		
Manner	Manner		
Undesirable_location	Source	Initial_location	[concrete +]
Speed	Speed		
Vehicle			
Time	Time		
Purpose	State_of_affairs		
Place	Locative_relation		
Depictive			
Path		Trajectory	[concrete +]
Degree	Degree		
Distance			
Explanation			

Consider Example 2, which represents the mapping of the FrameNet frame *Escaping* to the VerbNet class *escape-51.1*. Judging from their names, one expects the alignment to be very straightforward. However, *Escaping* has two core frame elements – *Escapee* and *Undesirable situation*, while *escape-51.1* is associated with four semantic roles: *Theme*, *Destination*, *Initial location* and *Trajectory*. Table 1 shows the mapping of the frame elements and the semantic roles: *Escapee* maps to *Theme*

and *Undesirable location* maps to *Initial Location*. In addition, the semantic role *Destination* corresponds to the peripheral frame element *Goal*, and *Path* aligns with *Trajectory*.

The table of Example 2 shows the alignment between the frame elements of the frame *Escaping* to the roles of the VerbNet class *escape-51.1*.

The judgment of which frame element corresponds to which semantic role is made by employing semantic information from the two resources, including comparison of definitions and similarity in the naming of the elements and roles (where possible) and inferred knowledge abstracted away from the structure of FrameNet where the frame elements are too specific. The latter case involves knowledge about the relations between more general and more concrete frame elements, which is obtained from a shallow hierarchy of frame elements based on the Inheritance relation between frames (Leseva et al., 2018a). For instance, the fact that *Text\_creation* inherits its properties from several frames (forming a chain of inheritance from a more specific to a more general frame) – *Text\_creation* > *Intentionally\_create* > *Creating* > *Transitive\_action* – allows us to identify a corresponding inheritance relation between relevant frame elements involved in these frames: *Author* > *Creator* > *Creator* > *Agent* and *Text* > *Created\_entity* > *Created\_entity* > *Patient*, that is the frame elements expressed as the subject and the direct object position in the frames under discussion. Having obtained this correspondence to more general frame elements, we try to map them to relevant roles in the semantic role set of the VerbNet verb class aligned with the respective frame. Thus, *Author* will be mapped to *Agent* in VerbNet.

Similarly, in Example 2 above *Escaping* inherits its properties from *Departing*, enabling us to infer the frame element Inheritance relations *Escapee* > *Theme* and *Undesirable\_location* > *Source*. The comparison between the most abstract frame elements and the semantic roles in the respective VerbNet class, leaves us with the straightforward alignment: *Theme* – *Theme* and the more or less transparent one: *Source* – *Initial\_location*.

### 3.4 Alignment between Syntactic Patterns

After aligning FN frames to VN verb classes (assigned to synsets or groups of synsets), and FN FEs to VN roles, we move towards mapping syntactic patterns from the resources aiming at providing a

new, syntactic layer to the conceptual description of general lexis verbs. The criteria for equivalence between two syntactic patterns obtained from the two resources include:

- correspondence in the number of elements or roles expressed in a syntactic pattern;
- correspondence between the frame element and the semantic role mapped to it as part of the previous task;
- correspondence in the syntactic restrictions (PP heads, clause types or subordinating elements) defined for the mapped frame elements and semantic roles;
- correspondence between the syntactic expression of each mapped frame element and semantic role – both in terms of the type of syntactic phrase by means of which they are expressed (NP, PP, etc.), and the syntactic position in which they are projected (e.g. subject, object).

**Example 3.** Aligned syntactic patterns for the FrameNet frame *Escaping* and the VerbNet class *escape-51.1*.

VN	NP(Theme)	V	
FN	NP.Ext(Escapee)	V	
VN	NP(Theme)	V	NP(Destination)
FN	NONE		
VN	NONE		
FN	NP.Ext(Escapee)	V	DNI.(Undesirable_location)
VN	NP(Theme)	V	NP(Initial_Location)
FN	NP.Ext(Escapee)	V	NP.Obj(Undesirable_location)
VN	NP(Theme)	V	NP(Trajectory)
FN	NONE		
VN	NP(Theme)	V	PP.destination(Destination)
FN	NP.Ext(Escapee)	V	PP[into].Dep(Goal)
VN	NP(Theme)	V	PP.initial_location (Initial_Location)
FN	NP.Ext(Escapee)	V	PP[from].Dep (Undesirable_location)
VN	NP(Theme)	V	PP.initial_location (Initial_Location)
FN	NP.Ext(Escapee)	V	PP[from].Dep (Undesirable_location)
VN	NP(Theme)	V	PP.initial_location (Initial_Location)
FN	NONE		PP.destination(Destination)
VN	NP(Theme)	V	PP.trajectory(Trajectory)
FN	NP.Ext(Escapee)	V	PP[from].Dep(Path)

The syntactic pattern alignment procedure is implemented as a set of mapping rules. As a result of their application we obtain a list of the equivalent syntactic models for a given FrameNet frame and

VerbNet class (Examples 3 and 4). Where no correspondence is discovered, the table cell is marked as NONE.

Example 3 shows the alignment of the syntactic patterns between the frame *Escaping* and the class *escape-51.1* following the mapping between the FEs and VN semantic roles (Theme – Escapee, Initial\_location – Undesirable\_location, Destination – Goal and Path – Trajectory). Misalignment occurs in the cases of additional semantic roles that are not considered core FEs (e.g., Trajectory).

**Example 4.** Aligned syntactic patterns for the FrameNet frame *Killing* and the VerbNet class *murder-42.1* (e.g., *kill, slay, annihilate, assassinate*, etc.).

VN	NP(Agent)	V	NP(Patient)	
FN	NP.Ext(Killer)	V	NP.Obj(Victim)	
VN	NP(Agent)	V	NP(Patient)	{with} PP.instrument (Instrument)
FN	NP.Ext(Killer)	V	NP.Obj(Victim)	PP[with].Dep (Instrument)
VN	NP.instrument (Instrument)	V	NP(Patient)	
FN	NP.Ext (Instrument)	V	NP.Obj(Victim)	
VN	NONE			
FN	NP.Ext(Cause)	V	NP.Obj(Victim)	

**Example 5.** Aligned syntactic patterns for the FrameNet frame *Killing* and the VerbNet class *suffocate-40.7* (e.g., *asphyxiate, choke, suffocate*, etc.).

VN	NP(Agent)	V	NP(Patient)	
FN	NP.Ext(Killer)	V	NP.Obj(Victim)	
VN	NP(Agent)	V	NP(Patient)	{with} PP.instrument (Instrument)
FN	NP.Ext(Killer)	V	NP.Obj(Victim)	PP[with].Dep (Instrument)
VN	NONE			
FN	NP.Ext (Instrument)	V	NP.Obj(Victim)	
VN	NONE			
FN	NP.Ext(Cause)	V	NP.Obj(Victim)	
VN	NP(Agent)	V	NP(Patient)	{to, into} PP.result(Result)
FN	NONE			

Examples 4 and 5 show different degree of misalignment between the syntactic patterns of the corresponding frames and verb classes. The frame *Killing* allows for the Instrument to appear as an external NP which matches a syntactic pattern within the verb class *murder-42.1* but not the verb class *suffocate-40.7*. Further, while the verbs under the frame *Killing* incorporate the result (the death of the Patient / Victim), the verb class *suffocate-40.7* also allows for a different Result as shown in the

last row of the table in Example 5 (e.g., *suffocate to/into unconsciousness*).

The asymmetries in the syntactic patterns covered by matched FN frames and VN verb classes for particular WN synsets are indicative of the need for more detailed syntactic analysis and the study of both the alignment between the FEs and the semantic roles and their syntactic realisation.

#### 4 Discussion of Results

The task of aligning FrameNet and VerbNet poses a number of challenges.

(1) Aligning frame elements and semantic roles at a different level of granularity.

This task is approached by employing (i) the semantic alignment of the fine-grained FrameNet frame elements to the more generalised VerbNet semantic roles using straightforward correspondences and the frame element hierarchy discussed in 3.3; (ii) the syntactic mapping – correspondences in terms of syntactic categories, prepositions, subordinating conjunctions, types of clauses, etc. – between frame elements and VerbNet roles with similar semantics and place in the conceptual description of particular verbs.

(2) Aligning the syntactic patterns for frames and verb classes with a different number of components or ones that allow alternative syntactic realisation. For instance, the syntactic description of the frame *Statement* includes the pattern:

##### NP.Ext(Medium) V Sfin

*The sign announced that the bar was closed.*

while no syntactic patterns with a finite clause are found in the description of the corresponding VerbNet class *talk-37.5*.

The semantic and syntactic information coming from different resources can serve for the validation of the linguistic generalisations captured in each of them. Thus, discrepancies across resources may be a sign of missing information in one of them and can be used for the enhancement of the poorer description.

Alternatively, the lack of correspondence may also be a red flag of the lack of semantic correspondence between seemingly identical or similar senses and hence should be studied with caution.

(3) Taking care of alterations such as passives (which are defined in FrameNet as separate syntactic patterns but are not represented in VerbNet).

Our approach would be to use the more comprehensive and explicit description in order to validate



the various alternations both within a language and cross-linguistically.

(4) Adapting syntactic patterns across languages and capturing significant parallels and differences in the syntactic projection cross-linguistically.

Both semantic and syntactic patterns may be adopted and possibly adapted cross-linguistically. Using an already available predefined set of patterns and refining or modifying them where needed, allows for a uniform representation of the data across languages and may be used to obtain a more complete description in the cases where corpora are not large enough to yield examples for all possible syntactic frames. Even so expert validation is indispensable.

With respect to the two research questions we have obtained the following results.

#### **How can we integrate semantic and syntactic information to enhance the conceptual description of WordNet synsets and literals?**

Such an integration may be implemented by employing semantic correspondences and syntactic patterns which apply to all (or most of) the synonyms in a given synset. In addition, more specific syntactic frames are needed in many cases to fine-tune these patterns and to cater for the syntactic realisation of individual literals, e.g. specific prepositions introducing prepositional phrases for different synonyms.

For example, the Bulgarian correspondence of the synset eng-30-00811375-v *avoid* includes, among others, the verbs *izbyagvam* and *stranya*. The former is associated with patterns corresponding to the ones defined in FrameNet and VerbNet for English:

**NP.Ext(Agent) V NP.Obj(Undesirable\_situation)**

EN: *Her friends now avoided her.*

BG: *Priyatelite i sega ya izbyagvaha.*

The latter, *stranya*, however, requires its *Undesirable\_situation* element to be realised as a PP headed by the preposition *ot* (from), which is not the case in English:

**NP.Ext(Agent) V PP[from](Undesirable\_situation).**

EN: *Her friends now avoided her.*

BG: *Priyatelite i sega stranyaha ot neya.*

This necessitates the definition of language-specific syntactic frames on the basis of evidence from the language under study.

**To what extent is syntactic information lan-**

#### **guage independent and can it be transferred from English to less-resourced languages such as Bulgarian?**

Although by no means identical, semantic descriptions are largely applicable across languages as far as senses are defined in a similar manner and should be largely uniform within a given synset. Syntactic frames are much more divergent cross-linguistically, yet there are major trends and similarities that may be transferred with caution across languages and resources.

With respect to general lexis verbs, our expectations are that they are realised by means of more common and well-established syntactic patterns with less specific features. Many of them are similar between languages.

However, an extensive analysis of syntactic structures should be carried out in order to determine the degree to which syntactic patterns defined for English can be adapted automatically to serve Bulgarian. To this end there are various corpora that can be used to extract occurrences of certain verbs, to study their context, combinations with prepositions, etc.

## **5 Conclusions and Future Work**

The research presented in this paper aims at providing a reliable alignment between: (a) FrameNet frame elements and VerbNet semantic roles on the basis of mapped FN frame – VN verb class pairs assigned to a number of WordNet synsets; (b) FrameNet lexical units' syntactic patterns and VerbNet syntactic frames. These combined allows for expanding the conceptual description of verbs with information about their syntactic realisation. Further, the data offer extensive opportunities to investigate to what extent the conceptual and the syntactic information can be transferred between languages, especially languages from one language family. These observations can play a crucial role in expanding semantic and syntactic description of Bulgarian verbs and thus, boost the development of new NLP applications.

## **Acknowledgments**

This paper is carried out as part of the scientific programme under the project *Enriching the Semantic Network Wordnet with Conceptual Frames* funded by the Bulgarian National Science Fund (Grant Agreement No. KP-06-N50/1 of 2020).

## References

- Collin F. Baker. 2008. FrameNet, present and future. In *The First International Conference on Global Interoperability for Language Resources*, Hong Kong, City University, City University.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference. Montreal, Canada*, pages 86–90.
- Aljoscha Burchardt, Katrin Erk, and Anette Frank. 2005. A WordNet detour to FrameNet. In *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen*, volume 8 of *Computer Studies in Language and Speech*. Lang, Frankfurt, Germany.
- Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. Verbatlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, November 3 – 7, 2019*, page 627 – 637. Association for Computational Linguistics.
- Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press, Cambridge, MA.
- Óscar Ferrández, Michael Ellsworth, Rafael Muñoz, and Collin F. Baker. 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.
- Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. Language resources and evaluation. *Communications. ACM*, 42(1):21–40.
- Karin Kipper-Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon. PhD Thesis*. Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA.
- Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova, and Ekaterina Tarpomanova. 2012. *The Bulgarian National Corpus: theory and practice in corpus design*. *Journal of Language Modelling*, 0(1):65–110.
- Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44:978–990.
- Egoitz Laparra and German Rigau. 2010. eXtended WordFrameNet. In *Proceedings of LREC 2010*, pages 1214–1219.
- Svetlozara Leseva and Ivelina Stoyanova. 2019. Enhancing conceptual description through resource linking and exploration of semantic relations. In *Proceedings of 10th Global WordNet Conference, 23 – 27 July 2019, Wroclaw, Poland*, pages 229–238.
- Svetlozara Leseva and Ivelina Stoyanova. 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*. Prof. Marin Drinov Academic Publishing House of the Bulgarian Academy of Sciences.
- Svetlozara Leseva, Ivelina Stoyanova, Hristina Kukova, and Maria Todorova. 2018a. *Integration of subcategorisation information in WordNet’s relational structure*. *Balgarski ezik*, 65(2):11–40.
- Svetlozara Leseva, Ivelina Stoyanova, and Maria Todorova. 2018b. Classifying verbs in WordNet by harnessing semantic resources. In *Proceedings of CLIB 2018, Sofia, Bulgaria*.
- Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago and London: The University of Chicago Press.
- Johann-Mattis List, Michael Cysouw, and Robert Forkel. 2016. Concepticon: A Resource for the Linking of Concept Lists. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2393–2400. European Language Resources Association (ELRA).
- George A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41.
- Martha Palmer. 2009. Semlink: linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*. 9–15.
- Martha Palmer, Claire Bonial, and Diana McCarthy. 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.
- Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher. R. Johnson, Collin. F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: extended theory and practice*. International Computer Science Institute, Berkeley, California.
- Lei Shi and Rada Mihalcea. 2005. Putting pieces together: combining FrameNet, VerbNet and WordNet for robust semantic parsing. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science*, volume 3406. Springer, Berlin, Heidelberg.
- Kevin Stowe, Jenette Preciado, Kathryn Conger, Susan Brown, Ghazaleh Kazeminejad, James Gung, and Martha Palmer. Semlink 2.0: chasing lexical resources. In *Proceedings of the 14th International Conference on Computational Semantics, pages 222–227 June 17–18, 2021*, pages 222–227. Association for Computational Linguistics.



Sara Tonelli and Daniele Pighin. 2009. New features for framenet – wordnet mapping. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL'09)*, Boulder, USA.

Zdenka Urešová, Eva Fucíková, Eva Hajičová, and Jan Hajič. 2020a. Synsemclass linked lexicon: Mapping synonymy between languages. In *Proceedings of the Globalex Workshop on Linked Lexicography, Language Resources and Evaluation Conference (LREC 2020)*, Marseille, 11–16 May 2020, pages 10 – 19.

Zdenka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2020b. Syntactic-semantic classes of context-sensitive synonyms based on a bilingual corpus. In *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 242–255. Springer International Publishing.

Piek J. T. M. Vossen, Laura Bloksma, and Rodriquez Horacio. 1998. The EuroWordNet base concepts and top ontology. Workingpaper, Vrije Universiteit.

# Croatian repository for the argument/adjunct distinction – SARGADA

**Matea Birtić**

Institute of Croatian Language  
and Linguistics  
mbirtic@ihjj.hr

**Ivana Brač**

Institute of Croatian Language  
and Linguistics  
ibrac@ihjj.hr

**Siniša Runjaić**

Institute of Croatian Language  
and Linguistics  
srunjaic@ihjj.hr

## Abstract

The distinction between arguments and adjuncts is a relevant topic in many linguistic theories (Tesnière, 1959; Chomsky, 1981; Langacker, 1987; Van Valin, 2001; Herbst, 2014, etc.). Even though theories provide similar definitions of arguments and adjuncts, sometimes it is difficult to draw a clear line between them. In order to determine ambiguous syntactic parts as arguments or adjuncts, various tests have been proposed, but they often give contradictory results and are not fully reliable. Nevertheless, they can be used as an auxiliary tool. The project *Syntactic and Semantic Analysis of Arguments and Adjuncts in Croatian – SARGADA* was launched with the aim of thoroughly investigating the distinction between arguments and adjuncts in Croatian, and to apply the theoretical results in a syntactic repository which would be a valuable resource for improving NLP tools and for researching and teaching Croatian.

In this paper, we will present diagnostic tests chosen as a tool to distinguish between arguments and adjuncts in the Croatian language. The repository containing sentences with ambiguous syntactic phrases and our workflow will also be described.

**Keywords:** Croatian language, syntax, argument/adjunct distinction, diagnostic tests, digital repository.

## 1 Introduction

Many linguistic theories (Tesnière, 1959; Bresnan, 1982; Chomsky, 1981; Langacker, 1987, Van Valin, 2001, etc.) distinguish between arguments (complements) and adjuncts as two separate grammatical categories, defining arguments as (semantically) obligatory, selected by a specific verb, and necessary to understand the event expressed by the verb (*Peter fixes the car.*) and adjuncts as optional, not selected by a specific verb, and not necessary for understanding the event expressed by the verb (*Peter fixes the car in the yard.*). Although the opposition is sometimes considered to be binary, most theories nowadays differentiate between obligatory and optional arguments, therefore operating with three distinct categories of non-predicate elements (obligatory arguments, optional arguments, and adjuncts).

Although a classification into arguments (complements) and adjuncts is made in almost all grammatical theories, it is rather difficult to draw a clear line between them (e.g., Vater, 1978; Schütze, 1995; Müller, 1996; Koenig, Mauner, and Bienvenue, 2003). Furthermore, quite a large number of various tests has been proposed to distinguish between arguments and adjuncts, which often yields controversial results. The project *Syntactic and Semantic Analysis of Arguments and Adjuncts in Croatian – SARGADA*, financed by the Croatian Science Foundation, was launched with the aim of clearly and precisely investigating the criteria for the definition and delimitation of arguments and adjuncts in the Croatian language, and to apply the theoretical results in a syntactic repository serving as a valuable resource for improving

natural language processing tools and for researching and teaching the Croatian language.

Since many theoretical approaches deal with distinguishing between argument and adjunct, we decided to conduct a thorough analysis of arguments and adjuncts and the criteria for their delimitation from the viewpoint of traditional Croatian grammars and three contemporary linguistic theories: valency theory and dependency grammar, generative grammar, and cognitive grammar. Combining three different linguistic theories is methodologically justified by the theoretical demands this project seeks to answer: (1) which criteria and tests are suitable to define and extract arguments and adjuncts in Croatian; (2) is the established distinction between arguments and adjuncts grammatically tenable; (3) could the distinction between arguments and adjuncts be defined independently of theory?

In this paper, in Section 3, we offer an answer to the first question by presenting diagnostic tests chosen to distinguish between argument and adjunct. In Section 4, we present the repository that contains sentences with ambiguous syntactic parts regarding the distinction of argument and adjunct. Section 5 concludes the paper.

## 2 Diagnostic tests

As has already been stated, there is no consensus on which tests should be used to distinguish arguments and adjuncts. In this paper, we will present tests chosen as a tool for distinguishing arguments and adjuncts in the repository. Dependency grammar uses, for example, the omission test, the implication test, the *do so* test, the paraphrase with dependent clause, and the *this happened* test. Generative grammar uses structure preservation/changeability after operation, the *do so* test, extraction from *wh*-islands, iterativity, etc. Cognitive grammar uses the methodological principle of conceptual (in)dependence. In the repository, roughly speaking, the omission test, the implication test, the *this happened* test, and the substitution test are taken from dependency grammar; the *do so* test and extraction from *wh*-islands are taken from generative grammar; and the dialogue and iterativity test come from functional generative description. A few other tests were considered, but it was decided not to include them because they are not applicable to

Croatian or not relevant (the dialogue test, paraphrase with a dependent clause, etc.).

### 2.1 Omission test

The omission test, also called the optionality test (Needham and Toivonen, 2011), the *Eliminierungs* test (Helbig and Schenkel, 1978), *Reduktionstest* (Engel, 2009<sup>4</sup>), etc., is a standard test to separate obligatory elements in a sentence from non-obligatory elements, i.e., optional arguments and adjuncts. If a syntactic phrase can be omitted, and the sentence remains grammatical, the omitted part is not an obligatory argument, but either an optional argument (1) or an adjunct (2). The problem is that some arguments can be omitted (e.g., with the verbs *eat*, *read*, *sing*) and some adjuncts are obligatory (e.g., some phrases in passive constructions). According to dependency grammar models, every obligatory phrase co-occurring with a specific verb is an argument.

- (1) *Ivan jede pizzu.*  
Ivan is-eating pizza.ACC.SG.  
'Ivan is eating (pizza).'
- (2) *On ide u crkvu (nedjeljom).*  
he goes to church Sunday.INST.SG.  
'He goes to church (on Sunday).'

### 2.2 Implication test

The implication test or *Folgerungs* test (Engel, 2009<sup>4</sup>) is also known as the *Core Participant Test* (Needham and Toivonen, 2011). The test relies on the semantics of verbs. According to this test, if a verb presupposes the appearance of an entity, then we are dealing with an argument.<sup>1</sup> The presence of a participant in the semantic structure of a verb can be signaled by a pronoun or an adverb (3) and the pronoun or adverb cannot be negated (4). The Croatian verb *boraviti* 'stay' always presupposes that there is a place where someone is staying. The verb's meaning cannot be realized without a "place".

- (3) *On boravi negdje.*  
he is-staying somewhere  
'He is staying somewhere.'
- (4) \**On boravi negdje, ali*

<sup>1</sup> One of the reviewers observed that by implication test adjuncts would qualify as arguments since most concrete acts would imply a place which is commonly assumed to be an adjunct. What matters here is that we are talking about what the verb presupposes, not the action in general.

he is\_staying somewhere but  
*negdje ne postoji.*  
 somewhere NEG exists

‘\*He is staying somewhere, but somewhere does not exist.’

In dependency grammars, this procedure is called anaphorisation. The application of this test makes sense for the optional arguments, while it is not needed for the obligatory arguments since they are already indicated by the omission test.

### 2.3 Do so test

In order to prove that Chomsky's claim (1965: 95–106) that place and time adverbials are sister constituents of VP and can occur freely with any VP, while direction, duration, place, frequency, and some manner adverbials subcategorize the verb, Lakoff and Ross (1976) introduced the *do so* test. According to the *do so* test, a non-stative verb and its arguments may be substituted with *do so*, while elements that occur after *do so* are outside the nuclear VP and are adjuncts.<sup>2</sup> Thus, the direct object, indirect object, directional adverbs, and affected locations are inside the verb phrase, while other adverbials are outside the nuclear verb phrase. In example (5), a *trip* is an argument and *last Tuesday* is an adjunct.

(5) John took a trip last Tuesday, and I'm going to do so tomorrow.

In many studies (e.g. Przepiórkowski, 2016), it is shown that the test is not reliable, especially for instruments and some *with* phrases that are, according to this test, always adjuncts. The problem that we would like to point out lies in the translation, i.e., choosing the Croatian equivalent of the verb *do*. *Do so* can be translated into Croatian as ‘*činiti isto*’, ‘*postupiti isto*’, etc. If we apply this test to three-place verbs that originally take accusative and dative arguments, such as the verb *pružati* ‘bring, give’, and we replace it with the verb *činiti* that has the same valency pattern as the original verb *pružati* ‘bring, give’, it follows that the dative complement is an adjunct since it occurs after the pro-verb (6). But if we replace the verb *pružati* ‘give’ with the verb *postupiti*, which in this case has the prepositional phrase *s* ‘with’ + the instrumental as its argument, it follows that the dative is an argument (7). So, the results depend on the distributional properties of a pro-verb or its subcategorization.

(6) *Djeca pružaju utjehu*

children give comfort.ACC.SG  
*odraslima, a odrasli*  
 adults.DAT.PL and adults.NOM.SG  
*to čine djeci.*  
 it do children.DAT.SG

‘Children give comfort to adults, and adults do so to children.’

(7) *Djeca pružaju utjehu*  
 children give comfort.ACC.SG  
*odraslima, a odrasli*  
 adults.DAT.PL and adults.NOM.SG  
 \**tako postupaju djeci.*  
 so do children.DAT.SG

‘Children give comfort to adults, and adults do so to children.’

### 2.4 This happened test

According to the *this happened* test (Brown and Miller, 1991: 90), if a sentence can be paraphrased by two sentences, one contains a nuclear predication and the other an adverbial. Example (8) can be paraphrased by two sentences; therefore, *in the kitchen* is an adjunct, while *on the table* in (9) is an argument.

(8) Ivan se popeo na stol. To se dogodilo u kuhinji.  
 ‘John stood on the table. This happened in the kitchen.’

(9) \*Ivan se popeo. To se dogodilo na stol.  
 ‘\*John stood. This happened on the table.’

### 2.5 Replacement test

The replacement test, as we call it in our repository, or *Ersatzprobe* (Ágel, 2000: 180), targets the syntactic level and should differentiate arguments from adjuncts. It is connected with the assumption that the morphological form of an argument is dictated by a verb (10), while the morphological form of an adjunct is not (11).

(10) *On piše zadaću / \*zadaći*  
 he is-writing homework.ACC homework.Dat  
 / \**na zadaći.*  
 on homework.LOC

‘He is writing homework / \*to homework / \*on homework.’

(11) *On piše zadaću na stolu*  
 he is-writing homework.ACC on table.LOC  
 / *u kuhinji / jučer.*  
 in kitchen.LOC yesterday

‘He is writing homework on the table / in the kitchen / yesterday.’

<sup>2</sup> Although adjuncts can be included in *do so* repetition.

## 2.6 Substitution test

The substitution test or *Supklassentest* (Engel, 2009) examines verb specificity. If the verb can be replaced with another verb or verb form in the environment of the same syntactic phrase, then the phrase next to it is an adjunct (Ágel, 2000; Šojat, 2008). Authors have noted that the same syntactic phrases can be arguments in one case, but not in another. The given example shows that the examined verbs can be replaced by verbs from the same or related semantic class and they require the same arguments (12).

- (12)
- |                      |           |              |   |                 |              |                 |
|----------------------|-----------|--------------|---|-----------------|--------------|-----------------|
| <i>Brat</i>          | <i>je</i> | <i>bacio</i> | / | <i>gurnuo</i>   | /            | <i>zavitlao</i> |
| brother              | AUX       | threw        |   | pushed          |              | swirled         |
| /* <i>razveselio</i> | <i>se</i> |              | / | /* <i>pojeo</i> | <i>kamen</i> | <i>u</i>        |
| cheered              | REFL      |              |   | ate             | stone        | into            |
| <i>vodu.</i>         |           |              |   |                 |              |                 |
| water                |           |              |   |                 |              |                 |
- 'The brother threw / pushed / swirled /\*cheered / \*ate a stone into the water.'

This is closely connected with the notion of subcategorization in generative grammar or what is called *Subklassenspezifisk* in the German dependency tradition, but is also widely used in traditional grammars. In dependency grammar, it is said that arguments are specific for a subclass of verbs, and therefore they are subclass specific. In the generative tradition, it is said that the verb is subcategorized for its arguments. The test is not reliable for adverbial arguments since they are not uniform in their morphological form, but are still obligatory.

## 2.7 Extraction from *wh*-islands

According to generative grammar, islands are parts of sentences from which it is difficult to extract phrases. There are strong islands, from which nothing can be extracted, and weak islands, from which some phrases can be extracted. The traditional assumption, going back to Huang (1982) and Chomsky (1986), is that arguments can be extracted from weak islands, but adjuncts and subjects cannot. The extraction of arguments from weak islands is better than the extraction of adjuncts and subjects, but is still not considered completely acceptable.

- (13a) Marko piše zadaću na stolu.  
'Marko is writing homework on the table.'

(13b) \*Gdje se Marija pita piše li Marko zadaću?  
'Where does Mary wonder if he writes homework?'

- (14a) Marko popravlja auto.

'Marko fixes the car.'

- (14b) ?Što se Marija pita poravlja li Marko?

'What does Mary wonder if Marko has fixed?'

According to Chomsky (1986) and Huang (1982), (14b) is supposed to be better than the example in (13b), which is true according to our intuition. In (13b) the adjunct phrase is extracted from a *wh*-island (indirect question with *li*-particle), and in (14b) the argument phrase is extracted from a weak island.

A legitimate question in this context, which has to be further investigated, is what counts as a weak island in Croatian. For instance, are indirect questions with the particle *li* really weak islands in Croatian, or do we have to find another context that will be a better context for sorting out arguments? There is also a long-standing question in linguistic literature about whether extraction from a weak island is truly sensitive to the argument/adjunct distinction or to some other linguistic property (Miliorini 2019).

## 2.8 Iterativity test

According to the iterativity test, adjuncts can be iterated freely, while arguments cannot (15) (Bresnan, 1982; Forker, 2014). However, on closer inspection, adjuncts can be iterated only if they refer to the same phenomenon or entity with a different degree of precision (Verspoor, 1997: 66; Brunson, 1993: 14), as shown in the example from Verspoor (1997: 66) (16) and its translation into Croatian (16b). The problem is that iteration is often possible for arguments as well (17).

- (15) \*John escaped from prison with dynamite with a machine gun.

(16a) Sam kicked a ball in the morning at 10 o'clock.

(16b) Sam je udario loptu ujutro u 10 sati.

- (17) On se žalio na susjedu, na njezino ponašanje.

'He complained about the neighbor, about her behavior.'

According to Przepiorkowski (2016) and Bresnan (1982), time, location, and manner can occur with any verb and can be iterated, but instruments cannot.

## 3 Repository

The applied part of the project includes the gathering of data, a corpus search, and creating a database for the description of ambiguous phrases regarding argument/adjunct status. In this paper, the current state of the repository's development

after two years of the four-year project is presented.

During the planning of the development of the relational database structure for the SARGADA repository, we consulted online resources in which conceptual solutions for the repository could be found. Linguistic information resources, in which the syntactic and semantic level of sentence parts are processed, can be roughly divided into several categories based on selected linguistic methodologies and schools:

- a) Syntactically parsed and morphosyntactically marked parts of general or specialized corpora of texts; e.g. numerous corpora via the *Sketch engine* platform (Kilgarriff et al., 2014).
- b) Dependency treebanks, as exclusively syntactic resources in the narrowest sense; e.g. *The Hamburg Dependency Treebank* (Foth et al., 2014), *Dependency Treebank for Czech* (Hajič et al., 2018).
- c) Valency lexicons, i.e. syntactic resources in a broader sense, created as the result of general linguistic or national projects; e.g. *ValPal – Leipzig Valency Classes Project* (Hartmann, Haspelmath, and Taylor, 2013), *T-PAS – Typed Predicate Argument Structure for Italian* (Jezek et al., 2014).
- d) Lexical databases with elaborated systems for marking semantic frames; e.g. *Framenet* (Fillmore and Baker, 2010), *Verbnet* (Kipper Schuler, 2005).

The SARGADA repository with its conceptual basis and as a digital resource of a specific part that directly arises as a by-product of syntactic research of ambiguous syntactic parts does not belong to those categories and therefore does not have a specific model. Another important distinguishing feature of the SARGADA repository concerning the studied resources is that the goal of its development is not to include already prepared linguistic data according to an unambiguous theoretical idea, but quite the reverse. This repository should examine new linguistic data about less researched syntactic categories of arguments and adjuncts for the Croatian language.

When compiling the model, we mostly followed dependency grammar due to the notion of the non-binary determination of the distinction between arguments and adjuncts. Notions about arguments and adjuncts from generative grammar

will serve as an additional control during the process of examining individual examples. In parallel with the study of these linguistic theories, the traditional grammar of Croatian, Serbian and Bosnian was consulted, as well as the works of prominent South Slavic syntacticians who, directly or indirectly, touch on the topic of arguments and adjuncts.

### 3.1 Workflow

Following the previously mentioned theories and analyzed data in the literature, in the first phase of preparation for the repository, a list of verbs was compiled. The list includes 111 Croatian verbs which are accompanied by ambiguous sentence parts that can be either arguments or adjuncts. After deeper analysis, we found that some of these verbs have different meanings that involve various valency patterns, so we are actually operating with 111 lemmas. Therefore, we decided to classify the lemmas into separate groups according to the ambiguous sentence part that appears in their valency patterns. For the purpose of creating the repository, these groups of syntactically ambiguous parts that occur with certain verbs have defined so-called “macrogroups” (groups of verbs that co-occur with the same ambiguous part). The verbs in the repository are classified according to these macrogroups, and we have singled out 12 groups.<sup>3</sup>

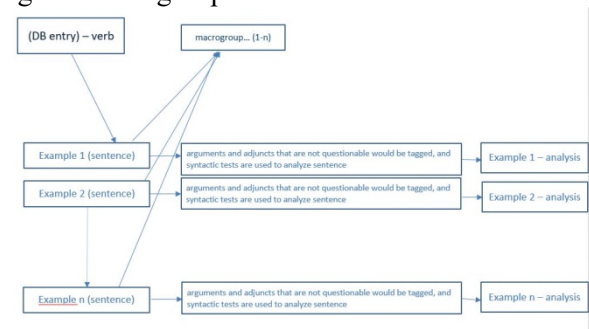


Figure 1. Schematic representation of the database organization.

<sup>3</sup> 1. verbs with place adverbials (e.g. *živjeti* ‘live’); 2. verbs with goal adverbials (e.g. *baciti* ‘throw’); 3. verbs with source adverbials (e.g. *dolaziti* ‘come from’); 4. verbs with time adverbials (e.g. *trajati* ‘last’); 5. verbs with quantity adverbials (verbs of exchange of goods and money, e.g. *stajati* ‘cost’); 6. verbs with manner adverbials (e.g. *ponašati se* ‘behave’); 7. verbs with cause adverbials (e.g. *proizlaziti* ‘result’); 8. verbs with purpose adverbials (e.g. *koristiti se* ‘use’); 9. verbs with instrumental case (e.g. *mirisati* ‘smell’); 10. verbs with benefactive dative case (e.g. *ispeći* ‘bake’), 11. verbs with inner objects (e.g. *sanjati* ‘dream’); 12. sport verbs (e.g. *trčati* ‘run’).



The workflow can be divided into a few steps, also shown in Figure 1:

1. Lemma input.
2. Selected sentence as an example for syntactic testing.<sup>4</sup>
3. Linking the selected example to a macrogroup.<sup>5</sup>
4. Tagging sentence parts in the example that are not ambiguous in terms of distinguishing arguments and adjuncts<sup>6</sup> (manual parsing).
5. Determining the sentence part that will be analyzed in the example by syntactic tests for argument/adjunct distinction.

When the sentence part for which the argument/adjunct distinction has to be tested is determined, the selected tests are performed outside the information system, and the outcomes of the tests are recorded in the database for each of them. Every test can give three possible results: 'Arg' (argument), 'Adj' (adjunct), or 'Not' ('test not used'). In this way, we seek to present results that are not binary but scalar.<sup>7</sup> In theory, it is possible that, for a particular example, all tests will give

---

<sup>4</sup> Examples are collected from Croatian linguistic literature, by translating cognate examples from international linguistic literature, and some of the examples were gathered by researchers during their investigations.

<sup>5</sup> In theory, the number of sentences or examples for one meaning of a verb and for one macrogroup is unlimited ( $n$ ), but it was decided in advance that one or two illustrative examples would be processed for each macrogroup for one meaning of a verb. On the other hand, we have already emphasized that due to polysemy, several macrogroups can be processed for each lemma.

<sup>6</sup> These are the following sentence parts: *Argument\_S* (argument\_subject), *Verb* (verb), *Argument\_DO* (argument\_direct object), *Argument\_IO* (argument\_indirect object), *Argument\_PP* (argument\_prepositional phrase), *Adjunct* (adjunct), *Aux* (auxiliary verb), *Reflex* (reflexive), *Conj* (conjunction) and *TEST* (ambiguous part for testing). We decided to tag unquestionable arguments and adjuncts during the parsing process and test only ambiguous sentence parts.

<sup>7</sup> One of the reviewers has brought to our attention that the scalar approach is not actually useful for the organization of the lexical or grammatical database. In a lexical or grammatical database, it is essential to define each element by some binary feature. Thinking exhaustively about the problem, we understand the reviewer's point of view, and we will reconsider our approach by giving a definite opinion on the status of some complements as arguments or adjuncts for further manipulation of the data. However, we think that the scalar approach is adequate as an illustration of our research, and it will give appropriate insights for teachers, students and researchers of the Croatian language.

the answer 'Arg', and then the system will show that the particular ambiguous sentence part is, without any doubt, an argument. The same, of course, applies to the adjunct. However, for most examples, different tests are expected to show different results that can be expressed as a scalar value and then graphically displayed after final processing. This would fulfil the applied part of the project in accordance with the work plan, and we believe that in this way the completed analysis would be more potent for the further development of research on the distinction between arguments and adjuncts in the Croatian language.

### 3.2 Technical information on the current stage of the development of the repository

The server infrastructure has been set up, i.e., an *Ubuntu 18* server operating system with LAMP architecture (Linux, Apache, MySQL and PHP) has been configured and installed, and a subdomain <http://sargada.jezik.hr> has been opened. The first (changeable) version of the database for the needs of the SARGADA repository was created and structured, and all the necessary programs for the development of basic models were installed. The presented linguistic model has been translated into a graphical interface using the *Javascript* language, i.e., the *Vue.js framework*, which enables flexible editing of the logical structure. The mark-up language HTML was used to structure and display the data according to the design, and the visual user interface was described and set up using *Cascading Style Sheets* (CSS) according to the instructions of the project members (as shown in Figure 2).

The screenshot shows the SARGADA repository interface. At the top left is the logo and the word 'Repozitorij'. Below it, there are three example forms for adding linguistic data:

- Skupina Glagoli s adverbijalnom dopunom mjesta**: Example 1 shows fields for 'Argument\_S' (kuća), 'Verb' (stoji), and 'Test' (stvorenja).
- Skupina Glagoli s adverbijalnom dopunom cilja**: Example 1 shows a dropdown for 'Odaberi vrstu', and fields for 'Argument\_S' (kuća), 'Verb' (je), and 'Test' (bitava).

At the bottom, there are two buttons: 'Dodaj primjer' (Add example) and 'Završi skupinu' (Finish group).

Figure 2. The current version of the SARGADA repository user interface.

The development of a central data management system (CMS) for users (project members) continues, so they will soon enter, edit and control linguistic data through this user interface. Currently, the PHP code is being developed and, through it, this input system will communicate with the configured database and save the structured data according to linguistic settings. When this code is completed, a stable (full-length version) will be prepared for entering data. Simultaneously, the database, back-end system and central management system will be tested based on these user actions. After all the data has been entered and harmonized, a graphic template will be designed for interaction with external users. This will allow for the creation of a visible system (front end) for online publishing and searching on the Internet, which would fulfil the work plan on the applied part of the SARGADA project.<sup>8</sup>

## 4 Conclusion

The paper presents the theoretical and applied part of the SARGADA project. The approach to distinguishing between argument and adjunct is presented in the first part of the paper. Arguments

are separated from adjuncts based on eight tests mostly taken over from dependency grammar and to a lesser degree from generative grammar. The tests are applied to sentence examples in the repository. The sentence examples are sorted according to their characteristic ambiguous part into 12 macrogroups. Since the ambiguous sentence parts examined in our project are “in-between arguments and adjuncts”, we decided to employ a gradual approach to distinguishing between argument and adjunct and to present scalar data.<sup>9</sup> The current state of the infrastructure of the digital repository SARGADA, which emerges as a product of work on the distinctions between arguments and adjuncts in these sentences, is also presented. The biggest gain of the parallel working process is that the need to create an applied digital resource prompted the creation of a methodology by which the tested results of theoretical research should be expressed at a scalar rather than a binary level. However, even greater added value is the fact that the process of transposing the linguistic model into the structure of the database and user interface spurred additional project tasks and produced results that were not even conceived at the initial stage of the project.

This project is important for a better understanding of the argument/adjunct distinction both cross-linguistically and with regard to Croatian and cognate languages. In addition, our research is also important for Croatian studies since the examined syntactic phrases had not previously been exhaustively described and their status was not unambiguously solved within Croatian linguistic literature. The repository of sentences that is freely available online will be of use in several segments of society (a tool for teaching and studying Croatian, or for improving natural language processing tools).

## Acknowledgments

This work has been fully supported by the Croatian Science Foundation under the project *Syntactic and Semantic Analysis of Arguments and Adjuncts in Croatian* – SARGADA (2019–04–7896).

<sup>8</sup> Online publishing on the Internet would be the minimum goal of creating a repository, and the added value would be, for example, the development of an application programming interface (API) of the SARGADA repository with other linguistic resources of the Institute of Croatian Language and Linguistics (or other research groups).

<sup>9</sup> See footnote 7.

## References

- Vilmos Ágel 2000. *Valenztheorie*. Gunther Narr Verlag, Tübingen.
- Joan Bresnan. 1982. Polyadicity. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*. MIT Press, Cambridge, Massachusetts, pages 149–172.
- Keith Brown and Jim Miller. 1991. *Syntax: A Linguistic Introduction to Sentence Structure*. Routledge, London.
- Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The M.I.T. Press, Cambridge, Massachusetts.
- Noam Chomsky. 1981. *Lectures on Government and Binding: The Pisa Lectures*. Foris Publications, Dordrecht.
- Noam Chomsky. 1986. *Barriers*. The MIT Press, Cambridge – London.
- Ulrich Engel. 2009. *Syntax der deutschen Gegenwartssprache*. Erich Schmidt Verlag, Berlin.
- Charles J. Fillmore and Collin F. Baker. 2010. A Frames Approach to Semantic Analysis. In *The Oxford Handbook of Linguistic Analysis*. Oxford University Press, Oxford, UK/New York, New York.
- Diana Forker. 2014. A Canonical Approach to the Argument/Adjunct Distinction. *Linguistic Discovery*, 12: 27–40.
- Kilian Foth, Arne Köhn, Niels Beuck, and Wolfgang Menzel. 2014. Because Size Does Matter: The Hamburg Dependency Treebank. In *Proceedings of the Language Resources and Evaluation Conference 2014 / European Language Resources Association (ELRA) (2014)*, eBook.
- Jan Hajič et al. 2018. *Prague Dependency Treebank 3.5*. Institute of Formal and Applied Linguistics, LINDAT/CLARIN, Charles University, LINDAT/CLARIN. PID: <http://hdl.handle.net/11234/1-2621>.
- Iren Hartmann, Martin Haspelmath, and Bradley Taylor, editors. 2013. *Valency Patterns Leipzig*. Max Planck Institute for Evolutionary Anthropology, Leipzig. <http://valpal.info>.
- Gerhard Helbig and Wolfgang Schenkel. 1983. *Wörterbuch zur Valenz und Distribution deutscher Verben*. 7. Aufl. Niemeyer, Tübingen.
- Thomas Herbst. 2014. The Valency Approach to Argument Constructions. In Thomas Herbst, Hans-Jörg, and Susen Falhaber, editors, *Constructions. Collocations. Patterns*. De Gruyter Mouton, Berlin – Boston, pages 167–216.
- Cheng-Teh James Huang. 1982. *Logical Relations in Chinese and the Theory of Grammar*. Doctoral dissertation, MIT.
- Elisabetta Jezek, Bernardo Magnini, Anna Feltracco, Alessia Bianchini, and Octavian Popescu. 2014. T-PAS; A Resource of Typed Predicate Argument Structures for Linguistic Analysis and Semantic Processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 890–895, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten Years On. *Lexicography*, 1: 7–36.
- Karin Kipper Schuler. 2005. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Dissertations available from ProQuest. AAI3179808.
- Jean-Pierre Koenig, Gail Mauner, and Breton Bienvenue. 2003. Arguments for Adjuncts. *Cognition* 89, 67–103.
- George Lakoff and John Robert Ross. 1976. Why You Can't Do So into the Sink. In James D. McCawley, editor, *Syntax and Semantics, Volume 7: Notes from the Linguistic Underground*. Academic Press, New York, 101–131.
- Ronald W. Langacker. 1987. *Foundations of Cognitive Grammar. Volume I: Theoretical Prerequisites*. Stanford University Press, Stanford.
- Rafaella Miliorini. 2019. Extraction from Weak Islands: Alternative to the Argument/Adjunct Distinction. *ReVEL* 17/16: 37–58.
- Stefan Müller. 1996. *Complement Extraction Lexical Rules and Argument Attraction*. [https://hpsg.fu-berlin.de/~stefan/Pub/case\\_celr.html](https://hpsg.fu-berlin.de/~stefan/Pub/case_celr.html).
- Stephanie Needham and Ida Toivonen. 2011. *Derived Arguments*. [web.stanford.edu/group/cslipublications/cslipublications/LFG/16/papers/lfg11needhamtoivonen.pdf](http://web.stanford.edu/group/cslipublications/cslipublications/LFG/16/papers/lfg11needhamtoivonen.pdf).
- Adam Przepiórkowski. 2016. How Not to Distinguish Arguments from Adjuncts in LFG. In *Proceedings of the Joint Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar*, pages 560–580.

Carson T. Schütze. 1995. PP Attachment and Argumenthood. In Carson T. Schütze, Jennifer Ganger, and Kevin Broihier, editors, *Papers on Language Processing and Acquisition*. MIT Working Papers in Linguistics 26. MIT, Cambridge, pages 95–151.

Krešimir Šojat. 2008. *Sintaktički i semantički opis glagolskih valencija u hrvatskom*. PhD dissertation, University of Zagreb, Zagreb.

Robert D. Jr. Van Valin. 2001. *An Introduction to Syntax*. Cambridge University Press, Cambridge.

Heinz Vater. 1978. On the Possibility of Distinguishing between Complements and Adjuncts. In Werner Abraham, editor, *Valence, Semantic Case and Grammatical Relations*. John Benjamins B. V., Amsterdam, pages 21–45.

Cornelia Maria Verspoor. 1997. *Contextually-Dependent Lexical Semantics*. PhD dissertation, University of Edinburgh, Edinburgh.



# Towards Dynamic Wordnet: Time Flow Hydra

**Borislav Rizov**

Sofia University "St. Kliment Ohridski" Sofia University "St. Kliment Ohridski"

bobyrizov@gmail.com

**Tinko Tinchev**

Sofia University "St. Kliment Ohridski"

tinko@fmi.uni-sofia.bg

## Abstract

Hydra is a Wordnet management system where the Synsets from different languages live in a common relational structure (Kripke frame) with a user-friendly GUI for searching, editing and alignment of the objects from the different languages. The data is retrieved by means of a modal logic query language. Despite its many merits the system stores only the current state of the wordnet data. Wordnet editing and development opens questions for wordnet data, structure and its consistency over time. The new Time Flow Hydra uses a Dynamic wordnet model with a discrete time embeded where all the states of all the objects are stored and accessed simultaneously. This provides the ability to track the changes, to detect the desired and undesired results of the data evolution. For example, we can ask which objects 10 days ago had 2 hyponyms, and 5 days later have 3.

**Keywords:** wordnet, modal logic language, Kripke frame, Hydra

## 1 Introduction

The wordnets in the world are evolving and growing in number. A lot of applications for development and visualization of such databases were developed in the last decades and one of them was the system Hydra whose main advantage was the modal logic query language for wordnet. Several years ago we introduced a new web version of the system with a simple, fast and comfortable interface. It allows the visualization and editing of wordnets for several languages simultaneously and concurrently by many users by means of a mobile first web interface. The system also has the ability to clone / replicate data from other languages in the database, which facilitates and accelerates the development of new synonymous sets. This can also be used for linguistic comparisons of language features. One of the challenges in the wordnet

world was the alignment of the databases developed for different languages. The concurrent work of different teams in different languages in a single environment could greatly facilitate this task and the overall the wordnet development. In this paper we are presenting a new dynamic model for wordnet, which guarantees the integrity of the data and all intermediate stages of its development. It also implies timely detection of data and structure inconsistency and this saves the very expensive human resources. The user has access to the data states in all the moments of its evolution at the same time. The user is also provided with a powerful modal query language with a new temporal modalities. Hydra prevents the loss of data even in the case of malicious user behaviour. The system has much better database model and the queries are processed much faster than in the previous versions, some of them are in orders of magnitudes faster.

## 2 Wordnet

Wordnet is a relational model (Koeva et al., 2004) of the language where the language concepts are represented as synonymous sets related to each other with over 20 semantic and lexical binary relations like hyperonymy, meronymy, antonymy and others. The main one is the super-subordinate relation hyperonymy (AKA is-a). It links the more general concepts like animal with its more specific ones like horse and bear. This creates a hierarchical structures of concepts (noun and verb) in the language.

## 3 Wordnet for many languages

Wordnet development started for English in Princeton (Miller et al., 1990) and then this idea was taken up for more than 40 languages. Most of these are developed or are still in development using the so called synchronous model where the hyperonymy

structure follows this of the Princeton WordNet. Using common identifier or alignment mappings the synsets encoding similar concepts in the different languages are linked to each other. Such relation or identifier is called ILI - Inter lingual index. These large wordnet databases with this relational model proved to be very useful for many linguistic tasks but experience several important problems. Being developed by different teams using different software platforms, file formats, databases, etc. the Wordnet databases are stored and maintained separately. The alignment (ILI maintenance) is made periodically usually for particular language pairs and particular version of these wordnet databases. Collaborative Interlingual index was developed to help reduce the sparse ILI mapping problem, but it did not succeed much. Some of the main problems in the past are the language database separation and the inconsistent synset identifiers in the central Princeton WordNet database.

#### 4 Static model for wordnet

In a fixed moment of time we have a family of synonymous sets (**synsets**) - the concepts in the language - interconnected by semantic relations like hyperonymy and meronymy. Diving in them we find some associated data like part of speech and the words that they are comprised of. We call a word in a particular synset **literal**. Keep in mind that a single word can be found in several synsets while the literals are unique (can be thought as  $\langle \text{synset}, \text{word}/\text{compound} \rangle$  pairs). These literals are connected by lexical relations. In a wordnet database we also have some text data like sample usage, notes about some particular synset or literal features, etc. We call this data **notes**. The notes can be thought as  $\langle \text{synset}/\text{literal}, \text{text} \rangle$  pairs.

#### 5 Wordnet as a Kripke frame

Let's consider 3 types of objects - Synset, Literal, Note for the objects in wordnet databases. We define special binary relations to encode the relationships between them. In this way Literal relation connects a particular literal to its parent synset. Usage relation connects a note object with the synset that it is usage example of. We also have the usual relations such as hyperonymy, meronymy, antonymy, etc. We obtain a Kripke frame  $\langle W, R \rangle$ , where  $W$  is a three-sort universe and  $R$  is a set of binary relations between the objects in it. Such a frame naturally introduces a modal logic language, which

we'll present in the next sections. Each object in wordnet can be considered as a feature structure with a fixed set of features depending on its type. Thus the synsets are provided with these features: pos (part of speech), lang (language code), ili (common identifier for the same concept in the different languages). Literals have word and lemma features, while Notes have note (the text they represent). All of the 3 types have some common features like id (unique identifier for the static model), userId - the identifier of the user that made this object, etc.

#### 6 Hydra

The wordnet database management system Hydra (Rizov, 2008) was created in 2006 in order to address the problems found in the development of BulNet - Bulgarian Wordnet. A new model for wordnet as a Kripke frame was introduced where all the linguistic data for the various languages (most importantly Princeton WN and BulNet) live in a single relational structure. Several years later, the system was developed as a modern SPA web application (Rizov and Dimitrova, 2016). The system is in production <http://dcl.bas.bg/bulnet> with wordnets for 22 languages. The data searching is made by means of a modal logic query language.

#### 7 Dynamic model for wordnet

An ordinary static wordnet database is an incomplete instantaneous description of the language. There are synonymous sets in the languages that are not defined, some of the relations are not fully instantiated and the wordnet databases for the different languages are in different stages of their development. There are also some specific concepts for particular languages. Over time, both the language and its wordnet representation change and evolve. During this evolution, we have a different state of wordnet at any given time. This raises questions about the consistency of the data and its structure defined by means of the binary relations in this time flow.

If we take the snapshots of wordnet in the static model, we get a set of Kripke frames. Let's supply each object in each frame with the timestamp of its frame. Now let's take the union of the resulting set of disjoint frames. We get a single Kripke frame with all the manifestations of all of the objects in the wordnet. Formally it is the set:

$$\{\langle W_t, R_t \rangle\}_{t \in T}$$



where  $T$  is the time model. We implement a discrete time model with the assumption that at most one object or relational pair can be changed in a single moment of time. We guarantee this in our implementation and we are making the assumption even stronger. Every change in the data causes the creation of a new moment in this discrete time model. In this way the points of the time are those moments in which a single object or relational instance is changed, created or deleted. Regarding the physical time, the state of the object in a moment in it is the state of this object in the nearest moment in model time preceding the physical moment. In a fixed moment of the model time we can collect all the objects from this and the previous moments taking only the nearest (last version) state of each object. In this way we obtain the static Kripke frame for this particular moment. The collection of all the versions of all the objects we call Dynamic wordnet model.

## 8 Query language

The construction of wordnet and its editing opens questions about the change of data and their structure (evolution) over time. One may be interested in the availability of certain properties of the data and the relations. He would like to easily detect problems when they occur, to easily correct them without returning to a state with many changes made by many users, as is the case with the use of a backup. For example, object 1 has changed in the past. Meanwhile 2 and 3 have been changed (corrected), 4 has been created. We detect a problem with object 1 and its relations - there is some inconsistency in the structure. With the dynamic model, you can trace the whole process and find out exactly when and why the problem occurred. We do it by means of a modal logic language for wordnet which was created for the early implementations of Hydra and further developed with addition of temporal modalities. The system works with so called model checking - for a given modal formula, the set of the objects in which the formula is true is returned.

### 8.1 Dynamic wordnet language

We define the modal formulae syntax and the corresponding semantics inductively.

#### 8.1.1 Syntax

In our language we have:

- $\mathbf{N}$  - a set of individual constants (nominals)
  - in the system we use decimal numbers for them.
- $\mathbf{O}$  - a set of constants for the features in the objects and their values. They use the schema  $type('value')$ . For instance  $pos('n')$  is such constant.
- $\mathbf{R}$  - a set of relation symbols
- $\mathbf{TM}$  - a set of time modifiers

We have 4 types of temporal modifiers - for a fixed timestamp (real time moment), fixed operation moment (model time moment), relative future and relative past like this:

- t159737980000;
- o1235;
- f5;
- p3;

#### Atomic Formulae: AtomicFor

- $\perp$
- $\top$
- $\mathbf{N} \subseteq \mathbf{AtomicFor}$
- $\mathbf{O} \subseteq \mathbf{AtomicFor}$

#### Formulae: For

- $\mathbf{AtomicFor} \subseteq \mathbf{For}$ .

Let  $q$  and  $r$  be fomulae (queries),  $R \in \mathbf{R}$ ,  $t \in \mathbf{TM}$ , then the following are formulae:

- $!q$
- $q \ \& \ r$
- $q \ | \ r$
- $q \Rightarrow r$
- $q \Leftrightarrow r$
- $\langle R \rangle q$
- $[R]q$
- $\ll t \gg q$

We also use some relation modifiers, namely:

- $\sim R$  - the reverse relation of  $R$
- $R+$  - the transitive closure of  $R$
- $R^*$  - the reflexive and transitive closure of  $R$

## 8.2 Semantics

- A Time structure is  $\langle T, t_c, < \rangle$ , where  $T \neq \emptyset$  is a finite set,  $<$  is a linear ordering,  $t_c$  is  $\max_{<} T$  (the current moment)
- A Model of time is  $\langle \langle T, t_c, < \rangle, m \rangle$ , where  $m : \mathbf{TM} \times T \rightarrow T$
- A static model (Kripke frame for a given moment  $t$ ) is  $\mathfrak{M}_t = \langle W_t, \mathcal{R}_t, V \rangle$ , where  $W_t \neq \emptyset$ ,  $\mathcal{R}_t : \mathbf{R} \rightarrow \mathcal{P}(W_t \times W_t)$ ,  $V : \mathbf{N} \cup \mathbf{O} \rightarrow \mathcal{P}(W)$  and for  $c \in \mathbf{N}$   $V(c)$  has at most 1 element.
- A dynamic model is  $\mathcal{D} = \langle \{\mathfrak{M}_t\}_{t \in T}, \mathcal{T} \rangle$ , where  $\mathcal{T} = \langle \langle T, t_c, < \rangle, m \rangle$  is a model of time.

We define the **truth** of a formula in a object  $x$  in the Dynamic model  $\mathcal{D}$  by induction on the formula construction:

- $\mathcal{D}, t, x \not\models \perp$
- $\mathcal{D}, t, x \models \top$
- $\mathcal{D}, t, x \models c$  for  $c \in \mathbf{N} \cup \mathbf{O}$  iff  $x \in V_t(c)$

Each object in the database has an identifier and it is a nominal (constant) in our language. A synset identifier is encoded so as to be portable and it depends only on ili (identifier coming from PWN), pos (part of speech code) and the language (code) of the synset. In the implemented system this semantic more concretely is:

- $\mathcal{D}, t, x \models \$s$  iff  $x$  is a Synset
- $\mathcal{D}, t, x \models \$l$  iff  $x$  is a Literal
- $\mathcal{D}, t, x \models \$n$  iff  $x$  is a Note
- $\mathcal{D}, t, x \models \text{type}(\text{'value'})$  iff  $x.\text{type} = \text{value}$  (for instance  $x.\text{pos}=\text{n}$ , so  $x$  is a noun synset)
- $\mathcal{D}, t, x \models !q$  iff  $\mathcal{D}, t, x \not\models q$
- $\mathcal{D}, t, x \models q \ \& \ r$  iff  $\mathcal{D}, t, x \models q$  and  $\mathcal{D}, t, x \models r$
- $\mathcal{D}, t, x \models \langle R \rangle q$  iff  $\exists y(x \mathcal{R}_t(R) y \ \& \ \mathcal{D}, t, y \models q)$
- $\mathcal{D}, t, x \models \ll t \gg q$  iff  $\mathcal{D}, m(t, t), x \models q$
- We say that a formula is true in dynamic model at point  $x$ , denoted  $\mathcal{D}, x \models q$  iff  $\mathcal{D}, t_c, x \models q$

For the sake of an example we'll use concrete natural numbers in the following:

- $\mathcal{D}, t, x \models \ll o1235 \gg q$  iff  $\mathcal{D}, t_0, x \models q$  where  $m(o1235, t) = t_0$ .

As mentioned before, every data modification creates a model time moment which is referred as an operation id and  $t_0.\text{id}=1235$ .

- $\mathcal{D}, t, x \models \ll t159737980000 \gg q$  iff  $\mathcal{D}, t_0, x \models q$  where  $t_0$  is the nearest previous model moment to this timestamp
- $\mathcal{D}, t, x \models \ll p3 \gg q$  iff  $\mathcal{D}, t_0, x \models q$  where  $t_0$  is the nearest previous model moment to the moment  $t - 3$  days
- $\mathcal{D}, t, x \models \ll f5 \gg q$  iff  $\mathcal{D}, t_0, x \models q$  where  $t_0$  is the nearest previous model moment to the moment  $t + 5$  days

## 8.3 Query answering

A formula in the defined modal language is a query in Hydra. The result of such query  $q$  at a given time moment  $t$  is the set of the unique objects with respect to their ids such that their time is the most recent one which is prior to the time  $t$ . By default the time  $t$  is the current moment  $t_c$  when the query is executed. This moment  $t$  can be fixed to be some arbitrary moment by means of the GUI, we call this feature *Time Machine*.

## 9 Example queries

Let's see some useful queries.

- Find the noun synsets that are on top of hyperonymy hierarchy in English:
 

```
pos('n') & [hypernym]⊥ & lang('en')
```
- Find the synsets that are exactly two levels below the top in the hyperonymy hierarchy:
 

```
[hypernym][hypernym][hypernym]⊥ & <hypernym><hypernym>⊤
```
- Find inconsistency between Bulgarian and English:
 

```
<ili>(lang('en') & pos('n')) & [hypernym][hypernym]⊥ & <hypernym>⊤ & lang('bg') & [hypernym]⊥
```
- Find the literals that before 3 days were presenting the word 'test' and 2 days later are not:
 

```
<p3>(word('test')) & !<f2>word('test')
```

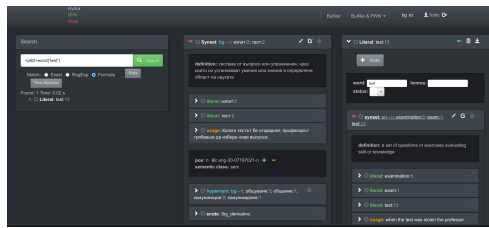


Figure 1: Hydra

## 10 Graphical user interface and implementation

Hydra is implemented in Javascript. It consists of a modern SPA (single page application) web app and a REST API service. The Wordnet data is stored in a Postgres database and the queries are translated to SQL queries. There is a preprocessing step where for each subformula its model time is determined. While most of the relations are stored in the database, some are implemented directly in the translations of the formula - such are the universal relation  $U$  and the transitive closures and reflexive and transitive closures of the other relations. Important feature is that no data can be lost during wordnet development even if there is some hostile user. Every change in the data creates a new copy of the object or relational instance touched with the same id but having the new data. For instance, when an object is deleted, a new record for it is created (operation record) where it is marked as 'deleted'. At any point the user can see the data as it used to be in the past with the so-called *Time Machine* - the user opens a dialog and selects a moment in the past and the data is as it used to be at this given time. The GUI is very simple and powerful. It has a Search Panel that has 3 modes for searching - using word, regular expression and a formula. The first 2 options find the synsets that have literals that match the provided word/regular expression. The latter is using the defined modal query language and it's much more powerful. There are 2 modes to visualize the data found. The first one (called 'Single') is visualizing the object selected from the list of the found items. The second one is aligning a pair of language wordnets that are present in the system. When the user selects a particular item from the list of the found items, the corresponding copies in the aligned languages are visualized. In this way the user can search some spanish word and see the aligned corresponding entries in french and English for example.

The visualization of an object consists of its

static data (like pos and language for the synsets) and the relations - all the connected objects by all the relations. The view is recursive and the data for the related objects is visualized on demand. Hydra is also a fully-fledged editor for this wordnet data. A user with sufficient rights can put an object into edit mode, the representational controls are replaced with edit controls and he/she can edit and save the data. Relational pairs are added by means of a wizard. These changes are sent to all the other users by means of notifications.

## 11 Conclusion and future work

The main achievement of our work is the expansion of Hydra's capabilities with time operators and the most valuable among them is the feature that when something is wrong and damaged in wordnet, we can repair it easily, as well as to understand the cause and user responsible for this error. One weakness is that the use of the system in this form requires competence in logical languages. To overcome this we are developing an GUI assistant to help the linguists with predefined queries and schema queries. For more complex queries some skills would remain required of course.

## 12 Acknowledgments

We thank the anonymous reviewers for their careful reading and valuable suggestions which made our paper better and more comprehensible.

## References

- Svetla Koeva, Stoyan Mihov, and Tinko Tinchev. 2004. Bulgarian wordnet – structure and validation. *ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY*, 7(1-2).
- George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to Wordnet: an on-line lexical database. *International journal of lexicography*, 3(4):235–244.
- Borislav Rizov. 2008. Hydra: a modal logic tool for wordnet development, validation and exploration. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco.
- Borislav Rizov and Tsvetana Dimitrova. 2016. Hydra for web: A browser for easy access to wordnets. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 342–346, Bucharest, Romania. Global Wordnet Association.





Organised by:



Department of Computational Linguistics  
Institute for Bulgarian Language  
Institute for Information and Communication  
Technologies  
**Bulgarian Academy of Sciences**

---

The Fifth International Conference *Computational Linguistics in Bulgaria* (CLIB 2022) is organised with the support of the National Science Fund of the Republic of Bulgaria under Grant Agreement No. КП-06-МНФ/7 of 20.07.2022.



---

ISSN: 2367-5675