International CLaDA-BG Conference 2021

Language Technologies and Digital Humanities

in Bulgaria (LTaDH-BG)

in conjunction with RANLP 2021

PROCEEDINGS

of International CLaDA-BG Conference 2021

Language Technologies and Digital Humanities

in Bulgaria

Edited by Petya Osenova and Kiril Simov

6 - 7 September 2021, Varna, Bulgaria

2023, Incoma Ltd., Bulgaria

EU Context and Financial Support

CLARIN
Common Language Resources and
Technology Infrastructure

DARIAH-EU
Digital Research Infrastructu
for the Arts and Humanit

REPUBLIC OF BULGARIA
MINISTRY OF EDUCATION
AND SCIENCE

Proceedings
of International CLaDA-BG Conference 2021
Language Technologies and Digital Humanities in Bulgaria
6 - 7 September 2021, Varna, Bulgaria

Online ISBN 978-954-452-094-6
acl-bg.org

Designed and Published by
INCOMA Ltd. Shoumen, BULGARIA

Organisation

Organizing Chairs

- **Kiril Simov,** Bulgarian Academy of Sciences

Programme Committee

- ***Alexandra Milanova***, Institute of Balkan Studies & Centre of Thracology, Bulgarian Academy of Sciences, Bulgaria
- ***António Branco***, University of Lisbon, Portugal
- ***Boyka Mircheva***, Cyrillo-Methodian Research Centre, Bulgarian Academy of Sciences, Bulgaria
- ***Darja Fišer***, University of Ljubljana, Slovenia
- ***Desislava Paneva-Marinova***, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
- ***Dimitar Iliev***, Sofia University “St Kliment Ohridski”, Bulgaria
- ***Dimitar Popov***, Shumen University, Bulgaria
- ***Eva Hajičová***, Institute of Formal and Applied Linguistics, Charles University, Czech Republic
- ***Ivan Georgiev***, IICT & IMI, Bulgarian Academy of Sciences, Bulgaria
- ***Ivan Kratchanov***, National Library "Ivan Vazov" – Plovdiv, Bulgaria
- ***Gennady Agre***, IICT, Bulgarian Academy of Sciences, Bulgaria
- ***Jurgita Vaičenonienė***, Vytautas Magnus University, Lithuania
- ***Karlheinz Mörth***, Austrian Academy of Sciences, Austria
- ***Kiril Simov***, IICT, Bulgarian Academy of Sciences, Bulgaria
- ***Koraljka Kuzman Šlogar***, Institute of Ethnology and Folklore Research, Croatia
- ***Inguna Skadiņa***, Institute of Mathematics and Computer Science, University of Latvia, Latvia
- ***Lars Borin***, Språkbanken, University of Gothenburg, Sweden
- ***Ludmila Dimitrova***, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
- ***Maciej Ogrodniczuk***, Polish Academy of Sciences, Poland
- ***Maria Stambolieva***, New Bulgarian University, Bulgaria
- ***Mila Maeva***, Institute of Ethnology and Folklore Studies with Ethnographic Museum, Bulgarian Academy of Sciences, Bulgaria
- ***Milena Dobreva***, Sofia University “St Kliment Ohridski”, Bulgaria
- ***Monica Monachini***, Institute for Computational Linguistics, Italy
- ***Nicolas Larrousse***, Huma-Num, Centre National de la Recherche Scientifique, France
- ***Nikola Ikonomov***, Bulgariana, Bulgaria
- ***Petya Osenova***, Sofia University “St Kliment Ohridski” and IICT, Bulgarian Academy of Sciences, Bulgaria

- ***Radoslav Pavlov***, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
- ***Roumiana Preshlenova***, Institute of Balkan Studies & Centre of Thracology, Bulgarian Academy of Sciences, Bulgaria
- ***Slavia Barlieva***, Cyrillo-Methodian Research Centre, Bulgarian Academy of Sciences, Bulgaria
- ***Snežana Petrović***, Institute for the Serbian Language, Serbian Academy for Sciences, Serbia
  Velka Popova, Shumen University, Bulgaria
- ***Veselka Zhelyazkova***, Cyrillo-Methodian Research Centre, Bulgarian Academy of Sciences, Bulgaria
- ***Vladimir Alexiev***, Ontotext, Bulgaria
- ***Yura Konstantinova***, Institute of Balkan Studies & Centre of Thracology, Bulgarian Academy of Sciences, Bulgaria

Table of Contents

Editorial Introduction to International CLaDA-BG 2021 Conference

Kiril Simov[1], Petya Osenova[1] and Zara Kancheva[1]

[1]*Institute of Information and Communication Technologies, Bulgarian Academy of Sciences,*

The papers in these proceedings provide solutions for the digital management and presentation of cultural heritage objects and knowledge, and present various language research and lexical resources: speech corpora, a Wordnet, a parliamentary data corpus and an event-annotated corpus.

An extensive review of the existing speech corpora in Bulgarian is done by [1]. The paper explores the history of Bulgarian speech corpora and every corpus is presented with its purpose and background of creation, scope, genre, speech spontaneity, presence of phonetic transcriptions and annotations, functionalities, format, availability, terms of use, etc.

The paper [2] presents the results obtained on a summer school for disinformation detection in Bulgarian social media content. They analyse Bulgarian tweets and combine three approaches: bag of words, concerning frequency, semantic and deep syntax analysis. The paper also aims at increasing the awareness in media literacy education.

The event annotation of historical documents for the aim of creating Bulgaria-centric knowledge graph is presented in [3]. The data is provided from different partners in the CLaDA-BG infrastructure, Bulgarian Wikipedia and several dictionaries, and the goal of the work is to build formalized annotation schemes for the all main areas in social sciences and humanities. The focus of the paper is on the challenges faced during the annotation process: (1) context-related (ambiguity, sign-referent-NE relation, missing context); (2) representational (two competing solutions available, predicate-Event relation) and (3) schema-based (the schema provides more than one interpretation is available or it does not cover some important facts; gaps in the schema).

The work of [4] presents the latest version of the BulTreeBank WordNet for Bulgarian (BTB-WN) – 4.0, and focuses on the consolidation of meanings in synsets, verification of the inherited from the OEW structure and relations, and addition of new interlingual relations and relations between BTB-WN synsets. The challenges of interlingual wordnet mapping are commented. The paper shows the approach and the evolution of decisions made for the creation and expansion of the BTB-WN, as well as the introduction of lemma markers which provide additional linguistic information for the members of the synsets.

The work of [5] presents the creation and implementation of The Humanities and Social Sciences Data Storage, Retrieval and Curation Environment. This is a web-based software environment, supporting a variety of digital cultural units and rich functionality for interaction, which has a focus on components providing storage, retrieval and management of data and metadata. The environment provides a metadata management and presentation functional module (incl. specific services), a metadata model management module, administrative services that are linked to a media repository and a user data repository.

The other paper focused on a corpus with audio data is [6]. The work seeks a solution for the problems of protecting multi-modal corpora with human speech, which are being developed in the Laboratory of Applied Linguistics (LabLing) at the University of Shumen. Two types of protection are used – cryptographic and steganographic.

Organizing Chair

- Kiril Simov, Bulgarian Academy of Sciences

Program Committee

- Alexandra Milanova, Institute of Balkan Studies and Centre of Thracology, Bulgarian Academy of Sciences, Bulgaria
- António Branco, University of Lisbon, Portugal
- Boyka Mircheva, Cyrillo-Methodian Research Centre, Bulgarian Academy of Sciences, Bulgaria
- Darja Fišer, University of Ljubljana, Slovenia
- Desislava Paneva-Marinova, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
- Dimitar Iliev, Sofia University “St Kliment Ohridski”, Bulgaria
- Dimitar Popov, Shumen University, Bulgaria
- Eva Hajičová, Institute of Formal and Applied Linguistics, Charles University, Czech Republic
- Ivan Georgiev, IICT & IMI, Bulgarian Academy of Sciences, Bulgaria
- Ivan Kratchanov, National Library "Ivan Vazov" – Plovdiv, Bulgaria

CLaDA-BG 2021 Conference: Language Technologies and Digital Humanities in Bulgaria 2021, September 06–07, 2021, Varna, Bulgaria

✉ kivs@bultreebank.org (K. Simov); petya@bultreebank.org (P. Osenova); zara@bultreebank.org (Z. Kancheva)

- Gennady Agre, IICT, Bulgarian Academy of Sciences, Bulgaria
- Jurgita Vaičenonienė, Vytautas Magnus University, Lithuania
- Karlheinz Mörth, Austrian Academy of Sciences, Austria
- Kiril Simov, IICT, Bulgarian Academy of Sciences, Bulgaria
- Koraljka Kuzman Šlogar, Institute of Ethnology and Folklore Research, Croatia
- Inguna Skadiņa, Institute of Mathematics and Computer Science, University of Latvia, Latvia
- Lars Borin, Språkbanken, University of Gothenburg, Sweden
- Ludmila Dimitrova, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
- Maciej Ogrodniczuk, Polish Academy of Sciences, Poland
- Maria Stambolieva, New Bulgarian University, Bulgaria
- Mila Maeva, Institute of Ethnology and Folklore Studies with Ethnographic Museum, Bulgarian Academy of Sciences, Bulgaria
- Milena Dobreva, Sofia University "St Kliment Ohridski", Bulgaria
- Monica Monachini, Institute for Computational Linguistics, Italy
- Nicolas Larrousse, Huma-Num, Centre National de la Recherche Scientifique, France
- Nikola Ikonomov, Bulgariana, Bulgaria
- Petya Osenova, Sofia University "St Kliment Ohridski" and IICT, Bulgarian Academy of Sciences, Bulgaria
- Radoslav Pavlov, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Bulgaria
- Roumiana Preshlenova, Institute of Balkan Studies and Centre of Thracology, Bulgarian Academy of Sciences, Bulgaria
- Slavia Barlieva, Cyrillo-Methodian Research Centre, Bulgarian Academy of Sciences, Bulgaria
- Snežana Petrović, Institute for the Serbian Language, Serbian Academy for Sciences, Serbia
- Velka Popova, Shumen University, Bulgaria
- Veselka Zhelyazkova, Cyrillo-Methodian Research Centre, Bulgarian Academy of Sciences, Bulgaria
- Vladimir Alexiev, Ontotext, Bulgaria
- Yura Konstantinova, Institute of Balkan Studies and Centre of Thracology, Bulgarian Academy of Sciences, Bulgaria

References

[1] D. Dimitrova, Bulgarian Speech Corpora: A Review, 2022.

[2] M. Dobreva, H. Krasteva, S. Gargova, Expanding the Boundaries of Disinformation Research on Bulgarian Social Media Content: The Experience from an Inspirational Summer School, 2022.

[3] P. Georgieva, I. Anastasova, P. Osenova, K. Simov, Challenges in Event Annotation across Linguistic Levels and Domains, 2022.

[4] Z. Kancheva, I. Radev, B. Miteva, G. Georgiev, Wordnet in a Contrastive Bulgarian-English Aspect: Hierarchies and Linguistics Idiosyncrasies, 2022.

[5] D. Paneva-Marinova, M. Goynov, L. Zlatkov, D. Luchev, R. Pavlov, L. Pavlova, Work-in-progress: Implementation of Humanities and Social Sciences Data Storage, Retrieval and Curation Environment for National Library "Ivan Vazov" – Plovdiv needs, 2022.

[6] D. Popov, V. Popova, K. Kordov, S. Zhelezov, R. Iglikova, Approaches to the Protection of Audio Files in BULGARIAN LABLING CORPUS, 2022.

Bulgarian Speech Corpora: A Review

Denitsa Dimitrova [1]

[1] *Sofia University "St. Kliment Ohridski", Tsar Osvoboditel 15, Sofia, Bulgaria*

Abstract

This paper aims to provide a summary of the existing corpora of spoken Bulgarian. The corpora are examined for their scope, genre, speech spontaneity, presence of phonetic transcriptions and annotations, as well for their availability. Some of the corpora (e.g., the BgSpeech corpus of colloquial Bulgarian, the Bulgarian children's language corpus, the Gewiss corpus of students' academic speech and others) are freely available and downloadable from online repositories and thus are reviewed in detail. All of the downloadable corpora include the text files with the transcriptions, but by far not all of them offer the audio files of the recordings. The description for the rest of the corpora is obtained from the respective publications of their authors. The corpora vary significantly in all of the above-mentioned criteria for evaluation and none of them are phonetically transcribed using (only) the IPA transcription conventions. The last section of the paper discusses the applicability of the corpora for further phonetic research.

Keywords

Bulgarian, speech corpora, phonetic transcriptions

1. Introduction

A speech corpus is "a collection of one or more digitized utterances usually containing acoustic data and often marked for annotations" [1]. Spoken corpora that consist both of transcriptions and the accompanying recordings thus are an invaluable resource for phonetic and phonological analysis, taking into account that the presence and availability of the audio data is of greater importance since "an auditory transcription is at best an essential initial hypothesis but never an objective measure" [1], whereas the sound recording does not change or vary.

Speech corpora have been compiled for Bulgarian since 1975 for a variety of purposes (like research, language preservation or software creation). There are collections of standards, colloquial and dialectal Bulgarian, of spontaneous and planned speech, such as radio or TV broadcasts or parliamentary speeches and debates, in all kinds of acoustic environments. Some of the corpora contain transcriptions, but many of them utilize a transcription system that uses symbols different from those in IPA and that is why for the most part these data sets are inaccessible for scholars outside Bulgaria. Most of the older corpora are not available for download, many of the still available do not offer the actual audio recordings. An overview of the existing spoken corpora for Bulgarian in a chronological order can be found in Table 1.

The dates assigned to the corpora are the dates of their first online release or the first publication if the corpus itself has not been made available online, however many of the corpora have been since then expanded and/or technically further developed, with the second year being the year of the last major change reported in a publication.

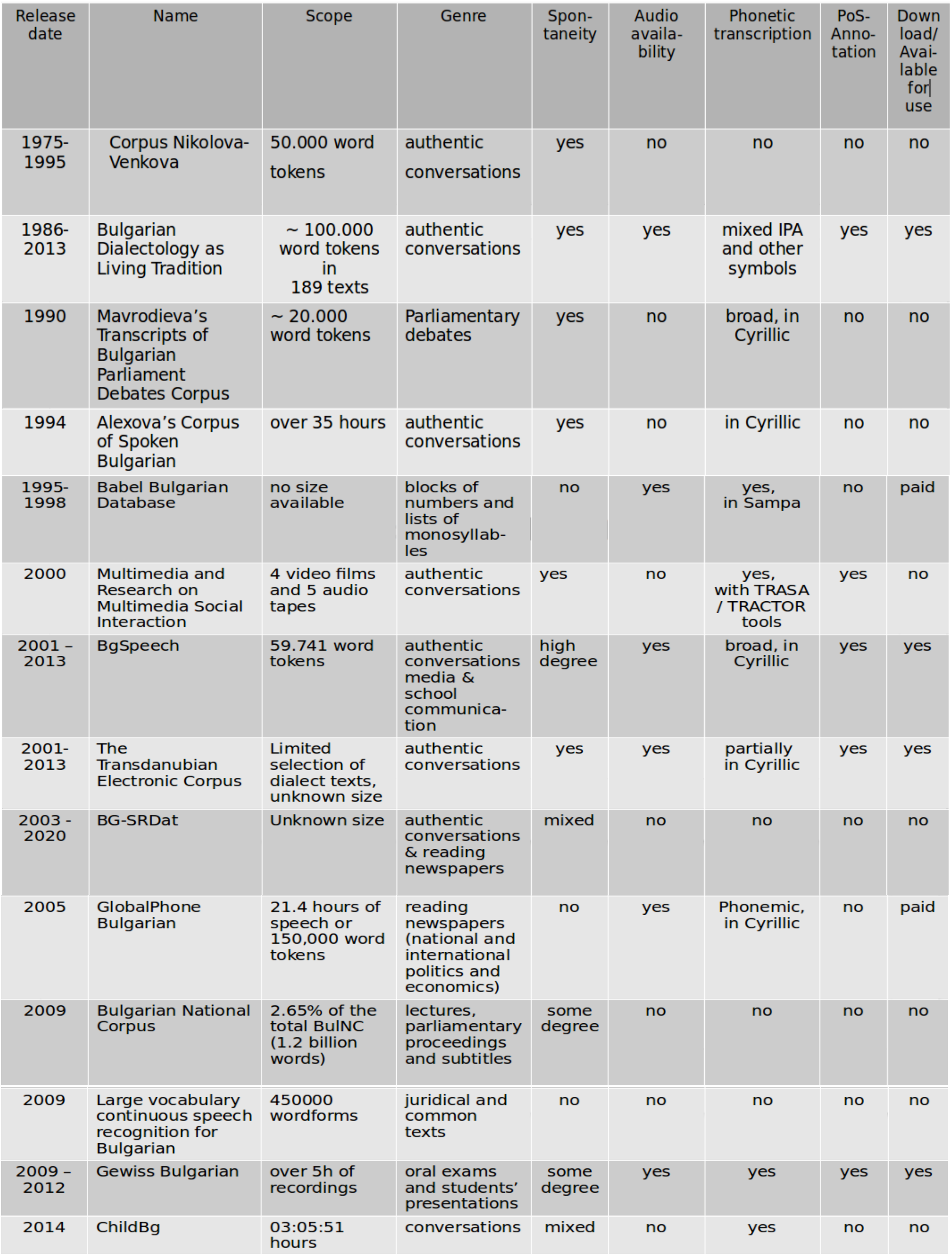
The paper is structured as follows: section 2 with its subsections presents a detailed review of the available speech corpora, their authors, their annotation methodology, the way of distribution, etc., and section 3 summarizes the paper.

2. A review of the speech corpora in Bulgarian in detail

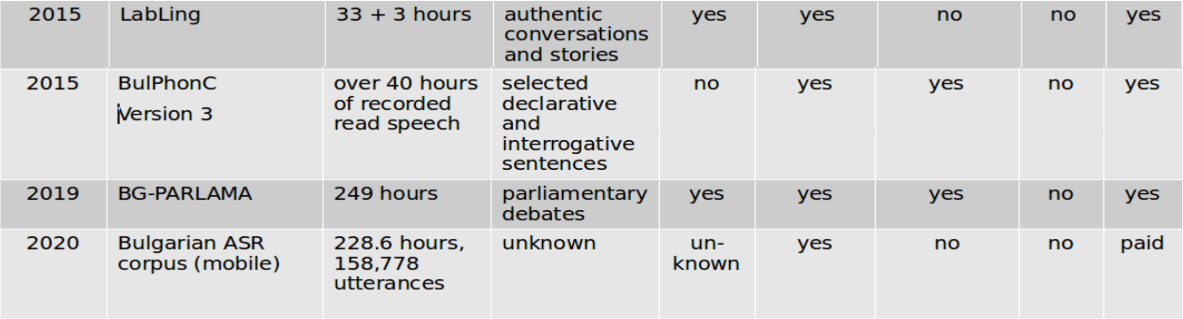
The review follows the chronological order of the overview (see Table 1 below). Its goal is to provide the background of the creation of every corpus in the list along with some details about it

EMAIL: d.dimitrova@uni-sofia.bg
ORCID: 0000-0002-6367-9167

like the genre (as the distinctive type of the texts constituting the corpus [2]), degree of spontaneity and convention of the phonetic transcriptions if present. Wherever possible, an excerpt of the respective corpus alongside its source is provided at the beginning of the section.



Release date	Name	Scope	Genre	Spon-taneity	Audio availa-bility	Phonetic transcription	PoS-Anno-tation	Down load/ Avai-lable for use
1975-1995	Corpus Nikolova-Venkova	50.000 word tokens	authentic conversations	yes	no	no	no	no
1986-2013	Bulgarian Dialectology as Living Tradition	~ 100.000 word tokens in 189 texts	authentic conversations	yes	yes	mixed IPA and other symbols	yes	yes
1990	Mavrodieva's Transcripts of Bulgarian Parliament Debates Corpus	~ 20.000 word tokens	Parliamentary debates	yes	no	broad, in Cyrillic	no	no
1994	Alexova's Corpus of Spoken Bulgarian	over 35 hours	authentic conversations	yes	no	in Cyrillic	no	no
1995-1998	Babel Bulgarian Database	no size available	blocks of numbers and lists of monosyllables	no	yes	yes, in Sampa	no	paid
2000	Multimedia and Research on Multimedia Social Interaction	4 video films and 5 audio tapes	authentic conversations	yes	no	yes, with TRASA / TRACTOR tools	yes	no
2001 – 2013	BgSpeech	59.741 word tokens	authentic conversations media & school communication	high degree	yes	broad, in Cyrillic	yes	yes
2001-2013	The Transdanubian Electronic Corpus	Limited selection of dialect texts, unknown size	authentic conversations	yes	yes	partially in Cyrillic	yes	yes
2003 - 2020	BG-SRDat	Unknown size	authentic conversations & reading newspapers	mixed	no	no	no	no
2005	GlobalPhone Bulgarian	21.4 hours of speech or 150,000 word tokens	reading newspapers (national and international politics and economics)	no	yes	Phonemic, in Cyrillic	no	paid
2009	Bulgarian National Corpus	2.65% of the total BulNC (1.2 billion words)	lectures, parliamentary proceedings and subtitles	some degree	no	no	no	no
2009	Large vocabulary continuous speech recognition for Bulgarian	450000 wordforms	juridical and common texts	no	no	no	no	no
2009 – 2012	Gewiss Bulgarian	over 5h of recordings	oral exams and students' presentations	some degree	yes	yes	yes	yes
2014	ChildBg	03:05:51 hours	conversations	mixed	no	yes	no	no



2015	LabLing	33 + 3 hours	authentic conversations and stories	yes	yes	no	no	yes
2015	BulPhonC Version 3	over 40 hours of recorded read speech	selected declarative and interrogative sentences	no	yes	yes	no	yes
2019	BG-PARLAMA	249 hours	parliamentary debates	yes	yes	yes	no	yes
2020	Bulgarian ASR corpus (mobile)	228.6 hours, 158,778 utterances	unknown	un-known	yes	no	no	paid

Table 1
Bulgarian speech corpora: An overview

2.1. Corpus Nikolova-Venkova

r05 (t33b,t38,t46,t71a)
t33b
Е, и какво... на сестра ми той внесе хиляда и осемстотин за апартамент, на другата ми сестра той купи телевизор и разни работи, на брат ми той внесе колата и плати три хиляди лева мито. Ний иначе, изобщо аз не мога да си представя... Ти какво си въобразяваш... Значи то е... Няма начин... Той работи, той е стар архитект предполагам и работи непрекъснато и печели. Той работи денонощно. Може да има и пет деца. Аз ако седя при него само да му помагам, ще изкарвам може би по двайсе хиляди на година. А пък аз тука, понеже с теми съм ангажирана, с други работи, не мога дори да му помогна на човека... Искам да кажа, че работа за архитекти има много. Може човек да работи и много, може хубаво да изкарва, но е мъчение направо. Ай да ходим да обядваме.

Figure 1: Text excerpt from the Corpus Nikolova-Venkova

This corpus comprises transcribed conversations from 1975 – 1977. It is part of a larger corpus of authentic conversations (100 000 word forms) which forms the basis of “The Frequency Dictionary of Colloquial Bulgarian” by Tsvetana Nikolova (Sofia University) [3]. The dictionary is the first of its kind and to this day a valuable resource for linguistic analyses. Freely available as a zip file on the website created in memory of the renowned Bulgarian linguist Prof. Miroslav Yanakiev [4]. The original recordings were made with a hidden portable tape recorder in randomly selected places (shops, streetcars, offices, homes). The recordings were then manually transcribed by Tsvetana Nikolova (with the active support and help from Prof. Yanakiev) on paper data sheets, and were then 1993 – 1995 entered by Tsvetomira Venkova into 25 files (chunks) of 2,000-word tokens each, which result in approximately 50,000-word tokens [5]. Most informants are from Sofia, while 3 recordings were made in Samokov and two in Plovdiv. None of the informants were aware of being recorded at the time. The texts are written in Cyrillic script within the standard orthography and there is no annotation of the structural organization of the real speech communication (dialogue, speaker turn, speaker overlaps etc.) and also no phonetic phenomena were registered. The corpus is included in several catalogues of linguistic corpora under the catalogue reference ELRA[2]-SD154 and was made freely available for download for research purposes on the unfortunately no longer running personal website of Tsvetomira Venkova and was also hosted on the Oslo Department for East European and Oriental Studies page, maintained by the renown Norwegian expert for Bulgarian language and culture Kjetil Rå Hauge**.** Unfortunately, this website is also down with no further information about a possible relocation of the resources.

[2] ELRA – European Language Resources Association

2.2. Mavrodieva’s Transcripts of Bulgarian Parliament Debates Corpus

Георги Пирински днес ми се струва / че в днешните разискванийъ по тази точка от дневния рет / всеки ясно трябва да поеме своята отговоронс // И / би било престъпление / ако не кажа няколко думи тази вечер прет вас сега // убеден съм / и искам да ми повярвате / че днес тези дни / прет нас е / последния ни шанс // да направим един управляем прехот / да направиме една управляема адаптация / на нашата икономика / към новите условия / ф които ще трябва да живеем да се развива нашта страна / оттукнататък // проблемите които имаме да решаваме / са свързани със три главни фактора // едният е дълбоките структурни промени и

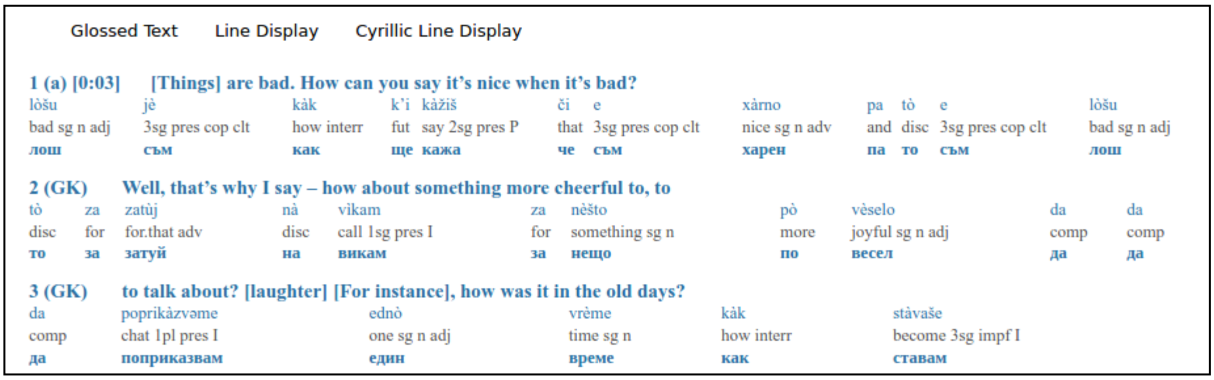
равновесието в нашта икономика и наследеното прес последните десетилетия // и тук има въпрос затова / плот на какво са решенията / които съ довели до тези структурни нарушения в равновесието / така или иначе тва е единия фактор // другия фактор / това са новите външноикономически условия / за които тук нееднократно бе говорено //

Figure 2: Text excerpt from Mavrodieva’s corpus

This corpus is built up with transcripts made by Ivanka Mavrodieva (Sofia University) from recordings of the debates of the 7th Great National Assembly on 31 October, 1990 for her dissertation about the parliamentary rhetoric in Bulgaria from that period. The broadcasts were registered at the Sociolinguistics Laboratory at Sofia University. The texts amount to a total of approximately 20.000 words and are transcribed phonetically according to the Bulgarian transcription convention in Cyrillic, the transcriptions however are, according to Tisheva in her paper about the electronic language resources for Bulgarian [6], characterized by the lack of a systemic and consequent approach to the representation of the phonetic and morphological specifics of the parts of the speech. As stated above, the corpus is no longer available for download.

2.3. Bulgarian Dialectology as Living Tradition

Bulgarian Dialectology as Living Tradition is "a searchable and interactive database of oral speech representing a broad range of Bulgarian dialects” [7]. It took a collective Bulgarian-American effort to construct this dialectal database and make it available to a wide spectrum of users. The corpus consists of 189 texts transcribed from the recordings collected in the course of several field trips in 71 different Bulgarian villages. Two types of transcriptions are provided – one using the Cyrillic set of symbols accepted by Bulgarian dialectologists, and another using a modified international transcription. "The primary transcription system is a combination of symbols adapted to the specific requirements of Bulgarian dialectal data. Certain symbols are taken from the International Phonetic Alphabet (IPA) and others from the academic transliteration that is the norm among Slavists. Where phonetic precision is needed in order to render important dialectal distinctions, IPA symbols are used. Elsewhere, simplified forms are used in order to make the transcription more accessible to non-phoneticians” [8]. The texts are then translated into English and annotated for relevant linguistic features so that data can be searched at many different levels. Five searches are possible at the website – word form search, lexeme search, linguistic trait search, thematic content search and phrase search.



1 (a) [0:03] **[Things] are bad. How can you say it’s nice when it’s bad?**

lòšu	jè	kàk	k’i	kàžiš	či	e	xàrno	pa	tò	e	lòšu
bad sg n adj	3sg pres cop clt	how interr	fut	say 2sg pres P	that	3sg pres cop clt	nice sg n adv	and	disc	3sg pres cop clt	bad sg n adj
лош	съм	как	ще	кажа	че	съм	харен	па	то	съм	лош

2 (GK) **Well, that’s why I say – how about something more cheerful to, to**

tò	za	zatùj	nà	vìkam	za	nèšto	pò	vèselo	da	da
disc	for	for.that adv	disc	call 1sg pres I	for	something sg n	more	joyful sg n adj	comp	comp
то	за	затуй	на	викам	за	нещо	по	весел	да	да

3 (GK) **to talk about? [laughter] [For instance], how was it in the old days?**

da	poprikàzvəme	ednò	vrème	kàk	stàvaše
comp	chat 1pl pres I	one sg n adj	time sg n	how interr	become 3sg impf I
да	поприказвам	един	време	как	ставам

Figure 3: Text excerpt from the corpus Bulgarian Dialectology as a Living Tradition [7]

2.4. Alexova’s Corpus of Spoken Bulgarian

(Съпрузите спорят за желанието на съпругата и децата да вземат в апартамента си куче.)

Съпругата: //днеска тр 'аа да хддиме да го зеем и ас се вьртъ като: побъркана/'/

Записващата: Виж го Веско!

Съпругът: //като почни да си върши естественити нужди/ ги върши как попадни/ кат е малко/ с'ако бебе// и тр'аа да растелиш ц 'алата стайа весници//

Записващата: Как?

Съпругът: //и постепенно да махаш идин по идин весник и то свиква да сире на весницити само/7 и накрайа оставаш само идин весник ф ъгълъ/ и накрайа като свикни ф него ъгъл да сир 'е/ тва до четвъртийъ месец/докато свикне ф него ъгъл да сир 'е/ махаш и него весник//

Figure 4: Text excerpt from Alexova’s Corpus of Spoken Bulgarian

The corpus consists of transcribed conversations in family contexts and was collected and processed by Krasimira Aleksova (Sofia University) for her dissertation “Language processes in the family (on material from the capital city)”, defended in 1994. On the whole there are 35 spontaneous family conversations, most of which were collected with a hidden tape recorder between mid-80’s and mid-90’s, from 65 working-age and 6 underage informants from a total of 28 families [9]. The corpus is compiled to conform specific research tasks like tracking the language processes in Sofia families, extracting sociolinguistic variables that correlate with the social characteristics of the speakers etc. This is the reason why metadata include detailed information about the social status of the speaker. The used system for phonetic transcription reflects a very high degree of individual phonetic and morphological specifics of the speech of every member of the studied families and uses Stoyko Stoykov’s system for phonetic transcription [10] modified with some additional signs and symbols. The corpus is no longer available online.

2.5. Babel Bulgarian Database

The BABEL Database is a speech database that was produced by a research consortium funded by the European Union under the COPERNICUS programme [11]. The task of the project was to create databases for Eastern and Central European languages comparable with the corpora system of the European Union languages created by the SAM (Speech Assessment Methods) project [12]. Its goal was to provide a common European resource for spoken language engineering and research. The creation of the BABEL database of spoken Bulgarian, Estonian, Hungarian, Polish and Romanian followed the protocols established by the SAM project and the transcriptions were carried out within the new transcription convention SAM developed – SAMPA, the machine-readable phonetic/ phonemic alphabet. In a joint pilot project to create a small corpus of spoken Bulgarian by the Universities of Sofia (Bulgaria) and Reading (U.K.) the authors implemented a variety of transcription symbol sets since there was no extension of SAMPA developed for Bulgarian [13]. During the course of the work on the BABEL project the Bulgarian version of SAMPA was established and the recordings of the read speech were transcribed according to this standard. Like every other language in the BABEL database the Bulgarian corpus consists of the basic "common" set which is:

- Many Talker Set: 30 males, 30 females; each to read twice the five blocks of numbers (each of which contains 10 numbers), 3 connected passages and one «filler» passage.

- Few Talker Set: 5 males, 5 females, selected from the above group, each to read 5 times the blocks of numbers, 15 connected passages and 2 «filler» passages, and 5 repetitions of the lists of monosyllables.

- Very Few Talker Set: 1 male, 1 female, selected from Few Talker set: each to read blocks of monosyllables in carrier sentences and five repetitions of the context words.

- Extension part: semi-spontaneous answers to questions, the answers were recorded by the 10 Few Talker Set speakers. [14]

The BABEL database was made available as CD-ROM through the ELRA website in 2000 and can still be purchased there for 300 Euro for members of ELRA or 600 Euro for non-members of the association.

2.6. Multimedia and Research on Multimodal Social Interaction

Multimedia and Multimodal Spoken Language Corpora Analysis, Stage 1 was a bilateral project between the University of Götheborg (GU), Department for Linguistics and Sofia University

"St. Kliment Ohridski" (SU), Faculty of Slavic Studies funded by Open society foundation – Bulgaria. The aim of the project was the creation of a standardized multimodal annotated and available over the internet corpus of Bulgarian that represents different areas of social life and can be used for future sociolinguistic and interlingual research projects [15].

$A: И а{с:з} ставам пре{д} стави си1 / отивам / тихо / отварям вратата /

Веса въ{ф:в} това време се съблича / нош{т}ницата си облича и като се

стресн{А:а} < тая пуста жена > ...

$D: Тя ш{т}е си2 и{с:з} кара акъла ма

$B: и тя ф{в}се по{т:д}скача Веса / а ти к{ак}во си каз{А:а}ла на майка си2

$C:тя съ{ф:в}сем така

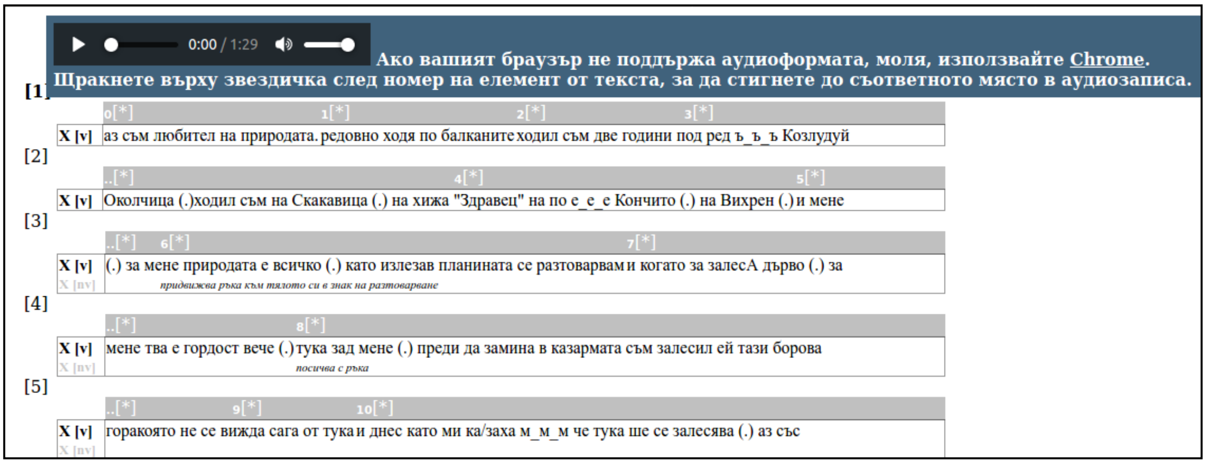
Figure 5: Text excerpt from the Multimedia and Research on Multimodal Social Interaction corpus.

Four video films and five audio tapes with authentic conversations were recorded and digitalized (no information is given about the speakers number, age, social status and gender). The authors adopted and modified the Swedish standards for transcription of the spoken language texts TRACTOR that combines both types of putting spoken language into written – transcription and orthography and even formulated a transcription standard for Bulgarian, but according to Tisheva and Dzhonova "the efforts of the project team did not attain the desired result, because the annotation used for this transcripts was applicable only for specific computer tools – TRASA and TRACTOR" [16]. The project was not continued and the little transcription that was made was stored as an archive on the website of the Department of Linguistics at the Göteborg University. The section with the Bulgarian corpus is no longer maintained.

2.7. BgSpeech

The Corpus of Spoken Bulgarian (CSBg) was developed at the Faculty of Slavic Studies at the Sofia University by researchers from the Department of Bulgarian language [17]. It consists of six hours recordings of informal communication and 1:51 hours of formal with a transcribed total of 59471 words and contains audio recordings, video recordings and transcripts. The creation of BgSpeech was carried out in two stages. In the first part, 2001 – 2004, non-formal conversations between friends, associates, and relatives were recorded with a hidden tape recorder. The transcripts are freely available in text format online at [18]. In order to be easy to use for a broader range of researchers the transcription is orthographic with some additional symbols for the notation of prosodic and paralinguistic information.

Later on, for the creation of the multimedia corpus, an XML standard for annotation of spoken corpora was adopted. The texts in the new transcripts were aligned with the sound track of the corresponding audio or video recording. The software used was the freely available EXMARaLDA Partitur-Editor from the EXMARaLDA package which is a system for working with oral corpora on a computer [19]. Formal communication from media and school with a high degree of spontaneity was recorded on video and audio. The transcripts follow the orthography, are exportable to XML format, contain non-verbal information, such as pauses, gestures, and mimics and include meta-information about the recording and the speakers (their education, age, occupation, etc.). The metadata provides information about the communication channel (radio, television, face-to-face, internet), the degree of preparedness of the speech (spontaneous, prepared), the domain (informal, formal, mass media, politics, business, academic), and the recording (audio, video; if broadcast: name; date and hour of recording/broadcast) [17]. Finally, a parallel corpus was created, where a normalized transcript and the original one are presented in a two-column view with the corrected deviations being highlighted in red. The transcripts along with the corresponding audio and video files are available at [18].



Ако вашият браузър не поддържа аудиоформата, моля, използвайте Chrome.
Щракнете върху звездичка след номер на елемент от текста, за да стигнете до съответното място в аудиозаписа.

[1]
0[*] 1[*] 2[*] 3[*]
X [v] аз съм любител на природата. редовно ходя по балканите ходил съм две години под ред ъ_ъ_ъ Козлудуй

[2]
..[*] 4[*] 5[*]
X [v] Околчица (.)ходил съм на Скакавица (.) на хижа "Здравец" на по е_е_е Кончито (.) на Вихрен (.) и мене

[3]
..[*] 6[*] 7[*]
X [v] (.) за мене природата е всичко (.) като излезав планината се разтоварвам и когато за залесА дърво (.) за
X [nv] *придвижва ръка към тялото си в знак на разтоварване*

[4]
..[*] 8[*]
X [v] мене тва е гордост вече (.) тука зад мене (.) преди да замина в казармата съм залесил ей тази борова
X [nv] *посичва с ръка*

[5]
..[*] 9[*] 10[*]
X [v] горакоято не се вижда сага от тука и днес като ми ка/заха м_м_м че тука ше се залесява (.) аз със
X [nv]

Figure 6: Text excerpt from the BgSpeech Corpus of Spoken Bulgarian [18]

2.8. The Transdanubian Electronic Corpus

1. човѐк когѝ съ докàра дъ мрѐ| съ збѐрът сѐ тъкà рòднините| го обикàл’ът| го почѝтът| пѝтъд го||
2. го плàчът?||
3. за плàчене нỳ мàй е вòрба||
4. Да.
5. го мѝйът
6. го къ̀път| го рѐшът| го рѐдът| го облечъ̀т||
7. ѐ| ако бъ̀де къ̀ е нъ зòр| вѝкът пòпъ| дъ го кòнкъ| и ако фàне| ‹...›| и ако фàне да е жѝф|
8. òн прирѝта штò съ вѝкъ| съ вàйке| съ онòдвъ| съ| бѝе със òдъръ| зъ мрек’ѐ||
9. по‿штò òн е умрѐн каквò да му мàй прàвът?||
10. ‹...› съ и пòпъ е напрàвил каквò| ако е фàнъл пòп и ако ли нѐ àлилуй||

Figure 7: Text excerpt from the Transdanubian Electronic Corpus [20]

The Transdanubian Electronic Corpus: Supplement to The Bulgarian Dialects in Romania by Maxim Mladenov currently places at users’ disposition a limited selection of transcripts of Bulgarian dialect texts spoken in Romania [20]. The authentic conversations on various topics included in the corpus were collected by the Bulgarian dialectologist prof. Maxim Mladenov during his fieldwork in 1962–1975 and show the Transdanubian dialects as they were at a specific point of their history; namely in the 1960s and 1970s. The transcriptions were made manually by himself according to Stoykov’s phonetic transcription convention for Bulgarian [21]. The creation of the electronic corpus was initiated by Olga Mladenova in 2001. This project comprises two stages: in the first phase, 2001 – 2010, the recordings were digitalized, the transcripts were typed to MS Word files, then compared with the original transcriptions and with the recordings and were finally thematically indexed [22].

In the second phase, after 2011, the corpus was transformed into an electronic one, the audio recordings were enhanced for a better quality of the audio, metadata and information about the dialects were added to the transcripts. The work on the electronic corpus was sponsored by a research grant from the Social Sciences and Humanities Research Council of Canada. The transcription in this part of the project is based entirely on the symbols prescribed by the IPA and is a narrow one, which means that all perceivable phonetic variation is taken into account, except for purely physiological and accidental processes, such as vowel nasalization between nasal consonants or word-finally. Other phonetic details, frequently deemed automatic and hence of no linguistic value, including vowel reduction, consonant palatalization before front vowels, voicing assimilation and devoicing of final obstruents are represented [20]. These transcriptions are unfortunately not present on the website, however the corresponding audio with a lot of supplemental interactive metadata is available online.

2.9. BG-SRDat: A Corpus in Bulgarian Language for Speaker Recognition over Telephone Channels

In its first phase the creation of the BG-SRDat Corpus by Atanas Ouzunov (Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences) included the recording of two different speech data – reading a text from a newspaper with an average length of about 40 seconds, and uttering a short phrase with the length of about 2 seconds. The two speech datasets were uttered in various sessions by different number of male speakers – 26 and 13, respectively. To achieve more realistic real-world conditions the speech data was collected by different types of telephone calls (internal-routing, local and long-distance) and various acoustical environments (noisy offices, halls and streets). The main purpose of the BG-SRDat was to provide data for evaluation of various speaker recognition techniques with noisy telephone speech in Bulgarian language [23].

For his PhD thesis defended in 2020 the author extended the corpus with three more types of speech data – reading a 80 seconds long passage from a newspaper, spontaneous conversations about a random topic of an approximately 7 minutes length with occasional speaker overlaps and a short phrase in English. For most of the data two sessions were recorded with microphone (26 speakers and 28 recordings) and over a telephone call (30 speakers with a total of 60 recordings).

The corpus itself is not published and not available online. Due to the non-linguistic purpose of the corpus the recordings are not transcribed, the audio files are provided with metadata files with information about the recording environment, speaker ID and speech data type.

2.10. GlobalPhone Bulgarian

The Bulgarian part of GlobalPhone[3] (ELRA-S0319) was collected by Anelia Mircheva for her student research project Bulgarian Speech Recognition and Multilingual Language Modeling at the Institute for Theoretical Informatics, University Karlsruhe in 2005 in the cities of Sofia and Pazardzhik, Bulgaria [26]. Data was collected from 77 speakers in total, all of whom Bulgarian native speakers from the west and central part of Bulgaria, aged from 18 to 65 years, 45 female and 32 male ones. The majority of speakers are well educated – graduated students, construction engineers, and teachers. Each speaker read on average about 112 utterances from newspaper articles (mainly from the online editions of three national Bulgarian newspaper websites *Banker*, *Sega* and *Cash*), corresponding to roughly 16.6 minutes of speech or 1940 words per person, in total 8674 utterances were recorded in small-sized rooms with low background noise using a close-talking microphone Sennheiser HM420 at 16kHz and 16bit resolution in PCM format. For each speaker there is a separate speaker session file with metadata about recording place and environmental noise conditions provided. More than 10,000 sentences were downloaded from the respective websites and processed (manually edited to normalize and clean the text, resolve abbreviations and numbers). Following the standard GlobalPhone protocols the authors focused on national and international politics and economics news [27]. The total sum of spoken utterances was 8674, corresponding to 21.4 hours of speech or 150,000 spoken words in total, covering a vocabulary of 23,000 words. Phonemic transcriptions in Cyrillic in UTF-8 encoding are available. The corpus can be purchased from the ELRA website for academic purposes for 600 Euro with an active ELRA-membership or 700 Euro for non-members of the association.

2.11. Bulgarian National Corpus

The Bulgarian National corpus (BulNC) was created at the Institute for Bulgarian Language "Prof. L. Andreychin” by researchers from the Department of Computational Linguistics and the Department of Bulgarian Lexicology and Lexicography [28]. It comprises several electronic corpora with a large variety of texts of different size, media type (written and spoken), style, and period (synchronic and diachronic), developed in the period 2001-2009 for the purposes of the two departments, and has been substantially expanded in the following years. With 979.6 million tokens from all Bulgarian texts the corpus reflects the state of (mainly the

[3] The GlobalPhone corpus developed in collaboration with the Karlsruhe Institute of Technology (KIT) was designed to provide read speech data for the development and evaluation of large continuous speech recognition systems in the most widespread languages of the world, and to provide a uniform, multilingual speech and text database for language independent and language adaptive speech recognition as well as for language identification tasks. The entire GlobalPhone corpus enables the acquisition of acoustic-phonetic knowledge of overall 22 spoken languages, amongst which Arabic, Chinese (Mandarin), French, German, Russian and Spanish [25].

written form of) the Bulgarian language from the middle of XX c. (1945) until the present. The original Bulgarian texts constitute 37.1% of the corpus, the translated texts – 40.5%, and the for the remaining 22.4% the source and the direction of translation is not given. There are also texts of different modality: predominantly written (91.11%) with spoken texts (8.89%) of limited types – lectures, parliamentary proceedings and subtitles [29]. The majority of the texts (97.5%) are obtained from the internet either through automatic crawling or manual downloading, while the remaining 2.5% are provided by the authors or publishers. The written texts are detailedly annotated, including morphosyntactic tagging and lemmatization and also word senses, synonyms, hypernyms and similar_to adjectives, noun phrases, and named entities. BulNC is supplied with a web interface for searching the corpus, as well as for building concordances and extracting collocations. The search system allows complex linguistic queries involving different levels of annotation combined in various ways [29]. The spoken corpus of the BulNC however is nowhere to be found on the website and there is no further information about the existence of transcriptions. Furthermore, it is not among the subcorpora available for downloading at the website. The benefits of including the BgSpeech corpus into the larger collection of the BulNC which would enable the user to perform more complex analyses on the spoken corpus' texts with BulNC's searching and extracting tools are discussed in [30].

2.12. Large vocabulary continuous speech recognition for Bulgarian

In a project for realizing a large vocabulary continuous speech recognition for Bulgarian the authors Mitankin (Institute for Parallel Processing, BAS), Mihov (Institute for Parallel Processing, BAS) and Tinchev (Faculty for Mathematics and Informatics, Sofia University) [31] used a corpus of legal texts that consists of $200x10^6$ words, the total number of words (monograms) in their dictionary is 442, 501. Their acoustic model was trained by adapting a speaker independent model with one hour of speaker's speech data. No information about the age and gender of the speaker(s) is given. The test set consists of 63 utterances from juridical texts and 1276 utterances from common texts. The authors have built a vocabulary of phonetized word forms using the phonetization rules for Bulgarian developed in the framework of the SpeechLab project [32]. Amongst the most common errors in the recognition were wrong word boundary segmentation, confused full form of the definite article inflection in masculine nouns with their non-full form, as well as the prepositions *в*, 'in' and *с*, 'with' which in their unvoiced phonetizations are very easily confused with any noise like speaker's aspiration especially at the utterance start. A user feasibility test showed that the system is already in a state permitting daily use for text dictation. Since the purpose of the creation of this corpus is not a strictly linguistic one there are no annotations and transcriptions available. The corpus is not available online.

2.13. GeWiss Bulgarian – Project "Gesprochene Wissenschaftssprache kontrastiv"

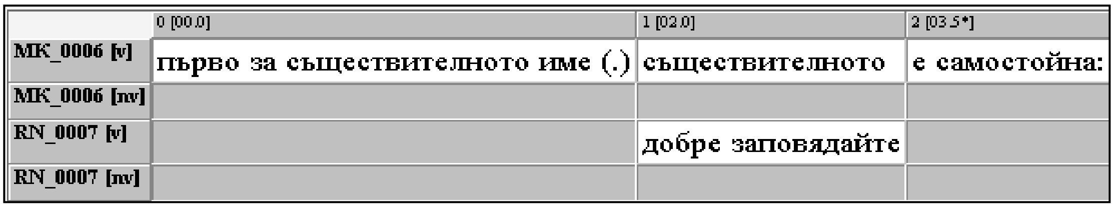
The GeWiss project is a product of the cooperation of researcher teams from the Leipzig University in Germany, Aston University in Birmingham, UK, and the Wroclaw University in Poland. The goal of the initial project is to create an online available contrastive corpus of academic speech in German, English and Polish [33]. Later on, researchers from the Sofia University joined with two subcorpora – Spoken Academic German by Bulgarian students and Spoken Academic Bulgarian again by students of the Sofia university.

The Bulgarian subcorpus [34] contains academic interaction – students' presentations and exams, as well as university teachers' talks and lectures. The corpus is orthographically transcribed and annotated on the following levels: literal and POS-tagging, lemmatization, time-alignment, orthographical transcription, annotation of discourse phenomena and language mixing.

The transcriptions in GeWiss are done manually with the software EXMARaLDA Partitur-Editor [19] on the basic transcription level of the adapted for the needs of the Bulgarian team transcription convention GAT@, that was developed specifically for the conversational analysis of spoken German, and the Bulgarian transcription convention developed by the project BgSpeech. In order for the transcript to be readable by non-linguists and non-phoneticians it is an orthographic representation of the utterances with no special symbols like IPA letters, punctuation marks and capital letters, only deviations from the codified pronunciation norm were registered.

GeWiss's metadata are represented in the XML format of the EXMARaLDA software and can be accessed and administrated with its metadata administration application, EXMARaLDA Corpus Manager (COMA). Metadata include social-demographic data, education information and information about language skills.

The corpus is part of the CLARIN infrastructure under the CLARIN RES License and is available after free registration in the Database of Spoken German (Datenbank für Gesprochenes Deutsch, DGD) for download and online browsing via the DGD (AGD @ IDS Mannheim) [35].



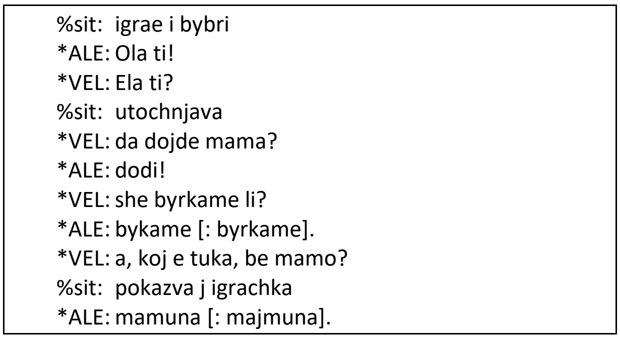
	0 [00.0]	1 [02.0]	2 [03.5*]
MK_0006 [v]	първо за съществителното име (.)	съществителното	е самостойна:
MK_0006 [nv]			
RN_0007 [v]		добре заповядайте	
RN_0007 [nv]			

Figure 8: Example of a transcribed in the Partitur-Editor software text from the GeWiss corpus [34]

2.14. ChildBG, Interactive multimedia corpus of children's speech in Bulgarian

For her dissertation "Recognition of children's speech in Bulgarian" [36] the author Radoslava Kraleva (South-Western University, Blagoevgrad) designed a software for building a corpus of children's speech in Bulgarian and compiled a test corpus with recordings of speech of children between 3 and 8 years old and also of several adult speakers. The corpus comprises a total of 03:05:51 hours, of which 01:14:12 hours recordings of boys, 01:30:39 hours of girls. There are 563 recordings available in the ChildBG corpus of which 493 or 02:44:51 hours children's speech and 70 or 00:21:00 hours of adult speakers. [37] All recordings are made with the same Philips SBM MDI50 microphone which facilitates the further analysis in different non-soundproof environments (university auditorium, a hall in the kinder garden or in the school, at home). The interactive corpus contains images of the words to be spoken and thus the produced speech can be considered as spontaneous. The recordings are phonetically transcribed in the machine-readable X-SAMPA format and are also POS-tagged. The corpus provides also detailed metadata about the speaker and their psychological condition during the recording and the environment [38]. The corpus is not available online.

2.15. Labling, Multimodal Bulgarian Corpus of Child Speech Data



%sit: igrae i bybri
*ALE: Ola ti!
*VEL: Ela ti?
%sit: utochnjava
*VEL: da dojde mama?
*ALE: dodi!
*VEL: she byrkame li?
*ALE: bykame [: byrkame].
*VEL: a, koj e tuka, be mamo?
%sit: pokazva j igrachka
*ALE: mamuna [: majmuna].

Figure 9: Text excerpt from the LabLing corpus [41]

The Multimodal Bulgarian Corpus of Child Speech Data Labling is developed by Dimitar Popov and Velka Popova at the Applied Linguistics Laboratory, Konstantin Preslavsky University of Shumen [39, 40, 41]. The database [40] is based on an array of longitudinal data from Popova's personal archive comprises two types of speech resources: the longitudinal corpus with spontaneous speech of five children at their early age (from 1 to 3 years old) and the narrative corpus with stories based on a series of pictures with 90 children at pre-school age (from 3 to 6 years old).

In the longitudinal corpus data collection and processing was supported by the Center for General Linguistics, ZAS to the projects "Erwerb sprachlicher Markierungen zur Differenzierung von ±Begrenztheit" (2003 – 2005) and "Syntaktische Konsequenzen des Morphologieerwerbs" (2000 – 2002). The participating children from the northeastern part of Bulgaria were recorded in the process of their daily interaction surrounded by their relatives. Three of the recordings were made by the researchers' team of LabLing and the rest by the mother of one of the children. The recordings are transcribed orthographically in Latin in CHAT format. The text files are browsable and downloadable on the LabLing website, for some of the recorded children audio is also available.

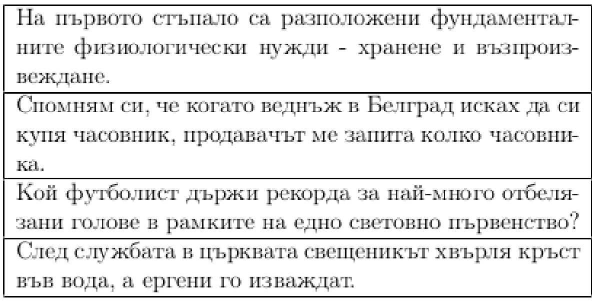
The narrative corpus was developed within the project "Erwerb und Disambiguierung intersententialer pronominaler Referenz" (2006 – 2007) and contains 91 transcripts of children`s narratives extracted from 50 monolingual children. Most of the recording were made using a recorder in several kindergartens in north-eastern Bulgaria, with only a few separate cases recorded at home or in the street. The corpus comprises the speech of children divided in 3 groups:

1. The first group includes 21 children aged 3-4 years – 36 narratives (21 of which without audio, 15 with both audio and transcripts)
2. The second group includes 23 children aged 4-5 years – 43 narratives (10 of which without audio, 33 with both audio and transcripts);
3. The third group includes 6 children aged 5-6 years – 12 narratives (with both audio and transcripts). [40]

The LabLing participates in the CHILDES talkbank and is part of the consortium of the Bulgarian national research infrastructure for resources and technologies for linguistic, cultural and historical heritage, integrated within CLARIN EU and DARIAH EDU, CLaDA-BG [42].

2.16. BulPhonC Version 3

The Bulgarian Phonetic Corpus BulPhonC [43] comprises over 40 hours of recorded read speech annotated automatically on phoneme level. The creation of the BulPhonC corpus has been supported by the project AComIn: Advanced Computing for Innovation funded by the FP7 Capacity Programme, Research Potential of Convergence Regions at the Bulgarian Academy of Sciences (BAS).



На първото стъпало са разположени фундаменталните физиологически нужди - хранене и възпроизвеждане.
Спомням си, че когато веднъж в Белград исках да си купя часовник, продавачът ме запита колко часовника.
Кой футболист държи рекорда за най-много отбелязани голове в рамките на едно световно първенство?
След службата в църквата свещеникът хвърля кръст във вода, а ергени го изваждат.

Figure 10: Example of the sentences in BulPhonC.

The corpus has been compiled at the Department of Linguistic Modeling and Knowledge Processing of the Institute of Information and Communication Technologies (IICT) by Dimitar Hristov, Ivan Zamanov, Ivana Yovcheva, Marina Kraeva, Nelly Hateva, Petar Mitankin and Stoyan Mihov. A total of 147 speakers were recorded, of them 62 male and 85 female Bulgarian speakers with an average speaker's age – 37 years [44]. The recording environment is a sound-proof studio with Sennheiser MK 4 microphone at 16 kHz sampling rate. The corpus contains 319 phonetically rich sentences divided into two parts. Part 1 contains 148 sentences and Part 2 contains the remaining 171 sentences. Most of the speakers have read only Part 1, on the whole there are 32 hours of 170 recordings with a total of 21891 pronounced sentences. Each utterance has a corresponding annotation on phoneme level in a format supported by Praat. The recorded signals were automatically segmented into utterances. All automatically segmented utterances were manually verified and the incorrectly segmented utterances were removed from the corpus. The remaining utterances were automatically annotated on phoneme level [44].

The corpus can be downloaded in a compressed tar.gz format containing 16-bit PCM wave files with 16 kHz sampling frequency (one file for each utterance) along with the text of the utterance (orthography) and their corresponding phoneme annotation files in Praat TextGrid format. It is free for scientific usage on demand provided that no parts of the corpus will be used for the development of commercial products of any kind.

2.17. Bulgarian Parliament ASR (BG-PARLAMA) corpus

Diana Geneva, Georgi Shopov, and Stoyan Mihov (IICT, BAS) have developed a new corpus of Bulgarian speech suitable for training and evaluating modern speech recognition systems. The data in the Bulgarian Parliament ASR (BG-PARLAMA) corpus is collected from the recordings of the plenary sessions of the Bulgarian Parliament. The texts are manually transcribed and then processed automatically together with the audio data of the speeches to build an aligned and segmented corpus using NLP tools and resources for Bulgarian. The resulting corpus consists of 249 hours of speech from 572 speakers [45]. The BG-PARLAMA corpus is the largest speech corpus currently available for Bulgarian.

One of the few sources of audio and transcriptions of spoken Bulgarian is the Bulgarian Parliament, where all speeches from the plenary sessions are transcribed orthographically manually and recorded on video. The texts and videos are published on the parliament's web page. The website of the Bulgarian Parliament provides mp4 video files for all plenary sessions from 2010 and transcripts from 1991 up till now (the older transcripts scanned and stored in pdf format), for the building of the corpus were used all sessions from 2010 to June 2018 and all transcripts from 1991. The texts have a different degree of spontaneity, as some of the speeches are prepared, while others are not. Transcriptions were normalized by converting them to lower-case and then tokenized.

The corpus can be downloaded in a compressed tar.gz format containing 148607 16-bit PCM wave files with 16 kHz sampling frequency along with the text of the utterance (orthography) at [45]. It is free for scientific usage on demand provided that no parts of the corpus will be used for the development of commercial products of any kind.

2.18. King-ASR-705 Speechocean

A commercial product developed by the Chinese AI data resource provider Speechocean is also available online – the Bulgarian Speech Recognition Corpus (Mobile). According to the website [47] there are 228.6 hours of recorded speech from 200 speakers (90 male and 110 female) or 158,778 utterances spoken on mobile platforms in a quiet (office/home) environment. There is no information about genre and degree of spontaneity, no transcriptions are available. The corpus can be purchased from the website or exchanged through a complicated system of membership discounts and bonus points on the website.

3. Conclusion

The review showed that on the one hand there are already many spoken resources in Bulgarian available for further processing and scientific usage. On the other hand, many of the existing are either not available or they do not offer the audio data. As stated in the introduction, phoneticians need mainly the audio to conduct a proper phonetic analysis. Audio data is provided in the following corpora: [7], [11], [18], [20], [25], [35], [38], [43], [46] and [47], with the last one, [11] and [25] being paid. Most of the older corpora (up until around 2000 – 2001) were uploaded to no longer running websites but maybe can be requested from their authors. There is also a possible legal issue with the absence of speakers' statements of agreement in the older corpora as they were created in a time when this was not a requirement. I see this as an opportunity for CLaDA-BG as an infrastructure for language resources for Bulgarian after clarifying the legal status of the recordings that were made without consent agreement to make a repository for spoken corpora where they can be stored and organized in a database. Noticeable is also the prevalence of the read speech over the spoken one in terms of overall duration, since the large corpora created for ASR contain mostly read speech. The development of a large corpus of spoken spontaneous Bulgarian with speakers from different age groups and regions could give the foundation for a thorough linguistic analysis of the state of various levels of modern Bulgarian.

4. Acknowledgements

I would like to express my gratitude to the team of CLaDA-BG and the anonymous reviewers for giving me the opportunity to present my research and for providing me with constructive and helpful remarks at the conference in Varna (2021) and later on with technical assistance during the formatting of the paper.

5. References

[1] J. Harrington. The Phonetic Analysis of Speech Corpora. Wiley-Blackwell, Malden, MA, 2010.

[2] P. Baker, A. Hardie, T. McEnery, A Glossary of Corpus Linguistics. Edinburgh University Press Ltd, 2006.

[3] Ts. Nikolova. Frequency dictionary of Colloquial Bulgarian. Nauka i izkustvo, Sofia, 1987 (in Bulgarian).

[4] Frequency dictionary of Colloquial Bulgarian. URL: https://miryan.org/glotta/cvetanka_nikolova_rechnik.html.

[5] Tsv. Venkova, An Electronic Corpus of Colloquial Bulgarian looking backward and forward], Problemi na sociolingvistikata 11 (2014), 430-440 (in Bulgarian).

[6] Y. Tisheva, M. Dzhonova, Electronic resources for the colloquial Bulgarian speech (the initiative BgSpeech). Littera et Lingua 2 (2010). URL: https://naum.slav.uni-sofia.bg/node/1735 (in Bulgarian).

[7] R. Alexander, V. Zhobov. Bulgarian Dialectology as Living Tradition (2016). URL: http://www.bulgariandialectology.org

[8] R. Alexander. Bulgarian Dialectology as Living Tradition: A Digital Resource of Dialectal Speech, Balkanistica 28 (2015), 1-13.

[9] K. Alexova, Language and family. Towards a methodology of a research of the spoken language in the microcommunities]. Interview press, Sofia, 2000 (in Bulgarian).

[10] S. Stoykov, Introduction to Bulgarian phonetics. Nauka i izkustvo, Sofia, 1966 (in Bulgarian).

[11] Babel Bulgarian Database, 2000. http://catalogue.elra.info/en-us/repository/browse/ELRA-S0085/

[12] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld & J. Zeiliger, "EUROM – A Spoken Language Resource for the EU", in Eurospeech'95, Proceedings of the 4th European Conference on Speech Communication and Speech Technology, 1995, Vol 1, pp. 867-870

[13] A. Misheva, S. Dimitrova, V. Filipov, E. Grigorova, M. Nikov, P. Roach, S. Arnfield, Bulgarian speech database: a pilot study. Proc. 4th European Conference on Speech Communication and Technology, 1995, 859-863

[14] P. Roach, S. Arnfield, W. Barry, J. Baltova, M. Boldea, A. Fourcin, Gonet, W., R. Gubrynowicz, E. Hallum, L. Lamel, K. Marasek, A. Marchal, E. Meistar, K. Vicsi (n.d.), BABEL: an Eastern European multi-language database. Proceeding of Fourth International Conference on Spoken Language Processing, 1996. doi:10.1109/icslp.1996.608002

[15] K. Petrova, K. Aleksova, Adaptation of Swedish Transcription System for Spoken Language Analysis for Bulgarian, 2002. https://www.fi.muni.cz/tsd2002/papers/60_Krasimira_Petrova.pdf

[16] Y. Tisheva, M. Dzhonova, Colloquial Bulgarian on the Web, Computer Applications in Slavic Studies: Proceedings of Azbuky.Net International conference and workshop, 24-27 October 2005. "Boyan Penev” Publishing Center, Sofia, 2006, p. 217-232.

[17] Y. Tisheva, M. Dzhonova, R. K. Hauge, The Corpus of Spoken Bulgarian, Papers of BAS. Humanities and Social Sciences, volume 5 (1), 2018, p. 20-28. https://www.papersofbas.eu/images/papers/papers-1-2018/Tisheva.pdf

[18] BgSpeech – the Corpus of Spoken Bulgarian on the web, 2003. URL: http://www.bgspeech.net/bg/resources.html.

[19] Exmaralda Partitur Editor, a free tool for transcribing and annotating of digital audio and video files, 2009. URL: https://exmaralda.org/en/partitur-editor-en/

[20] The Transdunabian Electronic Corpus of Bulgarian Dialects Spoken in Bulgaria, 2001. URL: http://www.corpusbdr.info/

[21] S. Stoykov, Bulgarian Dialectology. Izdatelstvo na BAN, Sofia 1993 (in Bulgarian).

[22] O. Mladenova, Bulgarian dialects in Romania: an experiment for the creation of an electronic corpus of the dialectal speech of the bilingual people], Ezik i literatura, volume 3-4, 2012, p. 115-122 (in Bulgarian).

[23] A. Ouzounov, BG-SRDat: A Corpus in Bulgarian Language for Speaker Recognition over Telephone Channels, Cybernetics and Information Technologies, volume 3 (2), 2003, p. 101-108.

[24] A. Ouzunov, Detection of Speech in Speaker Recognition Systems, PhD Thesis, Institute for Information and Communication Technologies, Bulgarian Academy for Sciences, 2020 (in Bulgarian).

[25] GlobalPhone Bulgarian, 2005. URL: http://catalog.elra.info/en-us/repository/browse/ELRA-S0319/

[26] A. Mircheva, Bulgarian Speech Recognition and Multilingual Language Modeling, Project Term (Studienarbeit), Institute for Theoretical Informatics, University Karlsruhe. 2006.

[27] T. Schultz, GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University, Proceedings of the International Conference of Spoken Language Processing, ICSLP 2002, Denver, CO, 2002, p. 345-348.

[28] Bulgarian National Corpus, 2009. URL: http://dcl.bas.bg/bulnc/

[29] S. Koeva, I. Stoyanova, S. Leseva, R. Dekova, T. Dimitrova, E. Tarpomanova, The Bulgarian National Corpus: Theory and Practice in Corpus Design, Journal of Language Modelling, volume 1, 2012, p. 65-110. https://doi.org/10.15398/jlm.v0i1.33

[30] Y. Tisheva, M. Dzhonova, Kh. R. Hauge, BGSPEECH and the representation of oral speech in the Bulgarian National Corpus. Problemi na ustnata komunikaciya 10(2) (2016), 175-186 (in Bulgarian).

[31] P. Mitankin, S. Mihov, T. Tinchev, Large vocabulary continuous speech recognition for Bulgarian, Proceedings of the RANLP 2009, Borovets, September 2009. http://lml.bas.bg/~stoyan/bgsrranlp.pdf

[32] M. Andreeva, I. Marinov, and S. Mihov, Speechlab 2.0 # a high-quality text-to-speech system for bulgarian, Proceedings of the RANLP 2005, Borovets, 2005, p. 52-58. http://dx.doi.org/10.13140/2.1.4801.4721

[33] K. Alexova, L. Laskova, Y. Velkova, A Corpus of students' academic speech. Balgarski ezik 2011 (3), p. 72–88 (in Bulgarian).

[34] A. Slavcheva, K. Alexova, L. Laskova, Y. Velkova, The Comparative Corpus of Academic Speech and the Bulgarian Data Inside, Littera et Lingua, 2012 (1). https://naum.slav.uni-sofia.bg/lilijournal/2012/1/slavchevaa (in Bulgarian).

[35] DGD, Datenbank für Gesprochenes Deutsch. URL: https://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome

[36] R. Kraleva, Speech Recognition: A spoken children's speech in Bulgarian]. Universitetsko izdatelstvo "Neofit Rilski", Blagoevgrad, 2019 (in Bulgarian).

[37] R. Kraleva, Acoustic-phonetic modeling for spoken children's Bulgarian, PhD Thesis abstract, Southwestern University "Neofit Rilski", Blagoevgrad, 2014 (in Bulgarian).

[38] R. Kraleva, A research of the ways to collect data for an interactive multimedia corpus of spoken children's Bulgarian, "Electronic forms of training in university education", 2014, p. 67-78 (in Bulgarian).

[39] D. Popov, V. Popova, Multimodal Presentation of Bulgarian Child Language, in: A., R. Potapova, N. Fakotakis (Eds.). Speech and Computer. 17th International Conference, SPECOM 2015, Athens, Greece, September 20-24, 2015, Proceedings. Springer International Publishing Switzerland 2015, 293-300.

[40] V. Popova, D. Popov, 2015, Bulgarian Labling Corpus. URL: https://childes.talkbank.org/access/Slavic/Bulgarian/LabLing.html, doi: 10.21415 / PHWH-J834

[41] V. Popova, D. Popov, Bulgarian corpus with children's speech on the CHILDES platform, Proceedings from the XIII Internarional scientific-methodical conference, Ufimsk. Gos. Avits. Techn. Universitet, 2021, p. 136–147 (in Bulgarian).

[42] CLaDA BG, E-Infrastructure for Bulgarian Language and Cultural Heritage. URL: https://CLaDA-BG.eu/en

[43] BulPhonC, the Bulgarian Phonetic Corpus. URL: http://lml.bas.bg/BulPhonC/

[44] N. Hateva, P. Mitankin, S. Mihov, BulPhonC: Bulgarian Speech Corpus for the Development of ASR Technology, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), 2016, p. 771-774.

[45] Bulgarian ASR Corpus BG-PARLAMA, 2019. URL: http://lml.bas.bg/BG-PARLAMA/

[46] D. Geneva, G. Shopov, S. Mihov: Building an ASR Corpus Based on Bulgarian Parliament Speeches, Proceedings of the 7th International Conference Statistical Language and Speech Processing (SLSP 2019), 2019, p. 188-197

[47] Speechocean, a Chinese AI data resource provider, 2020. URL: https://en.speechocean.com/datacenter/details/1862.html

Expanding the Boundaries of Misinformation Research on Bulgarian Social Media Content: The Experience from an Inspirational Summer School

Milena Dobreva[1,*,†], Hristiana Krasteva[1,2,†] and Silvia Gargova[1,†]

[1]*GATE Institute, Sofia University St Kliment Ohridski, Sofia, Bulgaria*
[2]*Plovdiv University Paisii Hilendarski, Plovdiv, Bulgaria*

Abstract
In July 2021, Gate Institute delivered the first summer school on discovering disinformation in social media content in Bulgarian. The school had several major building blocks: introduction to misinformation as a phenomenon, exploring the Bulgarian social media landscape, Bulgarian language phenomena that help identify suspicious content, natural language processing and machine learning techniques and tools combating misinformation. While there are numerous training activities on discovering disinformation on the global scale, this school was unique, focusing on Bulgarian language specifics. It also increased awareness in media literacy education among the participants and identified areas that can be developed further in research and in provision of education.

Keywords
Disinformation, social media, fake news research, language markers, BOW, deep syntax, semantic analysis, media literacy, fact-checking

1. Introduction

Misinformation is not a new phenomenon, but with the growth of social media use, it has become one of the most severe problems in contemporary societies. In this section we provide overview of the social media use in Bulgaria and outline some current research questions related to the spread of disinformation on social media. The importance of education in this domain and the provision of general information are also discussed during a summer school, specifically designed to explore the challenges in discovering disinformation in Bulgarian. In Section 2, we present some of the innovative content developed for the summer school. In Section 3 we provide conclusions and ideas for future research.

1.1. Social Media Use in Bulgaria

In recent years, social media has been gaining more popularity in Bulgaria. According to statistics from datareportal.com [1], 62% of Bulgarians were social media users in 2020. The most active users of social media platforms are people between the ages of 18 and 44. According to the same source, the most popular social media in the country is YouTube, followed by Facebook. Twitter platform is less popular among Bulgarians compared to some other countries. The exact number of users and preferred channels is difficult to establish – estimates of other sources may differ from those provided by [1], but we chose it due to the regularity with which their estimates are made and published. The data illustrate generic trends of spread of social media and preferences to specific channels among the population. Fig. 1 captures the most popular social media channels in Bulgaria, and Fig. 2 illustrates a cross-cultural difference in the preferred social media platforms in the UK and Bulgaria.

The awareness of most popular channels is essential for constructing studies which would explore the circulation of various messages on the social media and is also helpful when cross-cultural studies are being designed.

1.2. Social Media and Disinformation

Social media are the ideal environment for quick circulation of all sorts of messages [2]. There are several interdisciplinary research questions which do not have trivial answers: Is our society and the different demographic groups in it prepared for the digital age and the massive use of social media? Most people use social media daily, but how many of them know how social media works? While social media are platforms that allow the creation, sharing and discussion of information, interests, ideas, etc. and they make sharing and connecting people easier, they also have a dark side. The large popularity, the huge number of users and the speed at which information is spread are starting to attract more and more attention. The trust in friends, celebrities (but also businessmen, politicians, etc.) helps the acceptance of messages which come through these channels and the development of

milena.dobreva@gate-ai.eu (M. Dobreva); hnikolaeva@uni-plovdiv.bg (H. Krasteva); svgargova@gmail.com (S. Gargova)
0000-0002-2579-7541 (M. Dobreva)

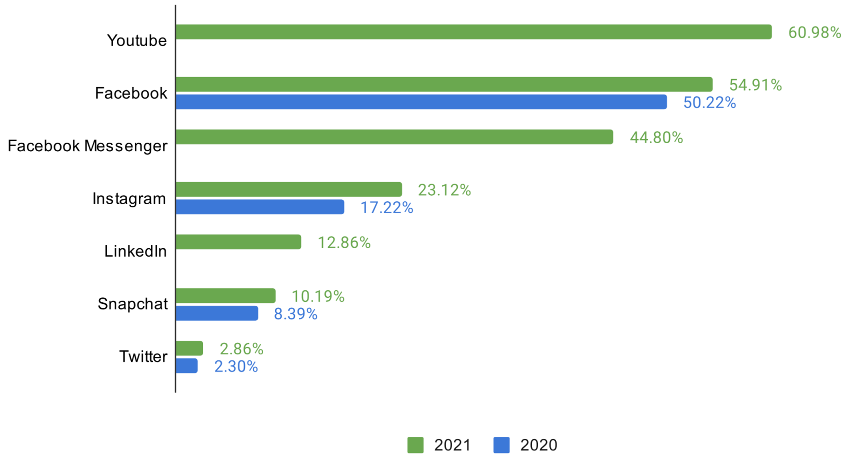


Figure 1: Most popular social media in Bulgaria in 2020 and 2021 according to [1]

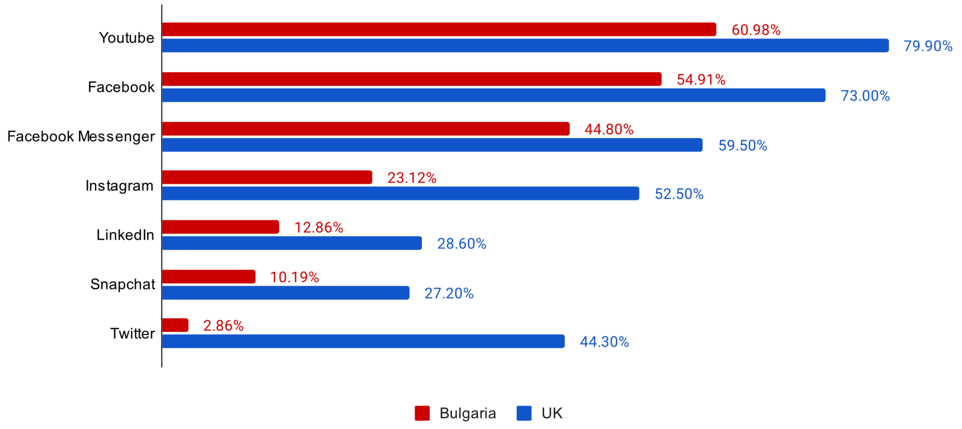


Figure 2: Comparison of the most popular social media platforms in Bulgaria and in the UK in 2021 according to [1]

skills to discern and check in this environment is one of the growing needs. Until there are people who engage with incorrect information and spread it further, there will be also plenty of supply. A study by Vosoughi shows that fake news spread faster than real news, and that humans, not bots, are responsible for this. [3] In this paper we do not have sufficient space to discuss the multiple terms around incorrect messages which circulate on social media. We will be using misinformation as a most generic term which shows that there is something incorrect in a piece of information and does not suggest this is due to a malicious intent. The term disinformation suggests an intent to confuse the recipient. Other terms in this domain include fake news – which applies to incorrect messages distributed as news items. More related terms are mapped according to the domain of use and the intent in Table 1.

Combating disinformation is a complex and time-consuming activity. Currently there are several major avenues of work to reduce the spread and the impact of disinformation. Fact-checking organizations, mostly originating from the media and non-governmental organisations (NGO) sectors, are involved in detailed checks of thousands of statements appearing in the media and then spreading through social media. There are also aggregators of checked statements, e.g. the Poynter Institute[1]

[1] Poynter Institute (n.d.). Institutional website. https://www.poynter.org/

Table 1
Types of misinformation

Intent	General	Politics	Journalism
The authors do not have an intent to harm but there could be technical or factual errors or missing details in their message.	Misinformation		
The authors are aware they are distributing incorrect information.	Disinformation	Propaganda	Fake news
The authors aim to harm a person or an institution.	Mal-information, Satire, Parody, Slander		
The intent can be of any of the above listed.	Rumors		

currently provides access to over 7,000 fact checks from over 70 countries. Only six of them are related to Bulgaria; the first Bulgarian fact checker is still not visible in this database. Research institutions and technology companies also contribute to the fight against disinformation. There were several recent EC-supported research projects which advanced the methods of automated verification of information. Media literacy is the educational domain most engaged with disinformation. The media literacy index ranks European countries [4]; Bulgaria was 30th out of 35 countries. Numerous educational institutions are providing programs to improve citizens' ability to assess information, discover relevant sources, and apply critical thinking and rational judgment. A comprehensive approach is needed to tackle the problem of disinformation. We cannot rely on fact-checking organizations and technological solutions alone. Data-checking organizations cannot verify all claims. This process is labor-intensive and expensive. There is only one such organization in Bulgaria, which makes the process even more time-consuming. Tech companies develop tools to help journalists and fact-checkers in the difficult task of claims checking, but there is a gap in having suitable tools for analysis of texts in Bulgarian. Finland is one of the few countries that successfully fights disinformation. One of the key factors is the implementation of media literacy as part of the education. This motivated us to put a summer school focusing on language aspects which contribute to deceive the readers. The work on the summer school was inspired by the example of Finland and the need to do more research and development of technologies for our native language which will help us fight disinformation.

1.3. General Information on the Summer School

On 26-27 July 2021, GATE Institute prepared and delivered the first summer school on disinformation in social media in Plovdiv, Bulgaria. The school had several major building blocks: introduction to disinformation as a social and media phenomenon; exploring the Bulgarian social media landscape; language features in Bulgarian that help identify suspicious content; natural language processing (NLP) and machine learning (ML) techniques and tools combating disinformation. In an international landscape where there are numerous training activities on discovering disinformation, the uniqueness of this school was in its focus on Bulgarian language specificities. It also contributed to the awareness in media literacy education and identified areas that can be developed further in research and educational provision. The school was organized in collaboration with the Association for the development of information society and was attended by 25 undergraduate students, postgraduate students, early-stage researchers, and librarians.

2. Novel content

The summer school aimed at presenting state-of-the-art knowledge on the ways automated tools are built to combat disinformation. This is an exciting domain on the intersection of machine learning and natural learning processing. The school focused on the Bulgarian language content, which was already explored in several publications; however, there is plenty of work to be done to build tools for analyzing and verifying content in Bulgarian. The focus of the school meant that there are two major challenges in preparing the educational sequence and content: Social media content is challenging due to the brevity of the style and the combination of different media (text, image, video). Tools for analysis in Bulgarian still need to be developed. In order to answer these challenges, the school team worked together to build the content, and some of this work was actual new research. The school also accommodated modules that explored the language markers specific for Bulgarian before discussing the natural language processing methods, approaches, and tools. A first experiment was also done with the participation of the attendees who took part in an annotation exercise which was based on authentic anonymized social media texts extracted from Facebook with CrowdTangle.

2.1. Linguistic Aspects of Researching the Language of Fake News in Bulgarian

Linguistic analysis composes substantial part of the effort to find, explore, research, and predict fake news and disinformation, since the type of data is usually textual. A wide variety of models and approaches has been implemented in structuring and analyzing the texts of disinformation. One way to classify these models and approaches would be to define them according to what they aim for. In this case they can be divided into lexical and syntactic. The lexical approach is more engaged with semantics (semantic analysis, for example based on the use of Bag of words – BoW model), whereas the syntactic is engaged with structure (deep syntax). However, language specificity and typology impose the need for personalisation of the methods or checking for language-specific characteristics at the least. Led by this idea during the summer school organized by the GATE Institute we focused on the specificity of Bulgarian misinformation and fake news from the perspective of different linguistic characteristics, as the starting stages of the project include gathering of tweets for annotation and linguistic analysis.

2.1.1. Methods

The linguistic approach to disinformation is focused on so called "language markers" of fake information. We thus research the grammar and syntax of these texts [5]. We focus here on three major approaches that we discussed during the summer school:

1. BoW: analysis of the frequency of appearance of certain words. The good sides of this model include the prospect possibility to identify semantically fake information. Setbacks, however, may include too sparse vectors due to the large count of words and tweets, not considering phrasal potency in Bulgarian (defined by the grammar of constituents [6]), and poor or no word order modeling or analysis.
2. Semantic analysis: this approach engages analysis of lexical meaning of utterances or written texts and its comparison to modeled profile on the topic. [7] Poor modeling and need for active human interaction may be considered as downsides of this approach.
3. Deep syntax: by using the tools of sentence structure graphic analysis (a method in the paradigm of generative syntax) tree-like images of sentences are created. The aim is to improve visualization and fine analysis, as well as to facilitate the comparison with the structures of already proven fake pieces of information [5, 8].

2.1.2. Working Definitions

For our experiments, we used fake news examples since the literature exploring misinformation in English social media content most frequently focuses on identifying this kind of misinformation. We use the following definition for fake news: intentionally submitted in the public space or media (be it digital or other) disinformation (e.g., Obama is a Muslim) [9] Among previous research we can differentiate several directions of linguistic interest. According to the classification of Horne and Adali (2017) [**?**] these are type of style, complexity of language and style, which is said to be explicated in both greater word count and number of sentences, as well as psychology i.e., emotional speech. Horne and Adali further conclude that fake news language is substantially different from that of real news in terms of the following characteristics:

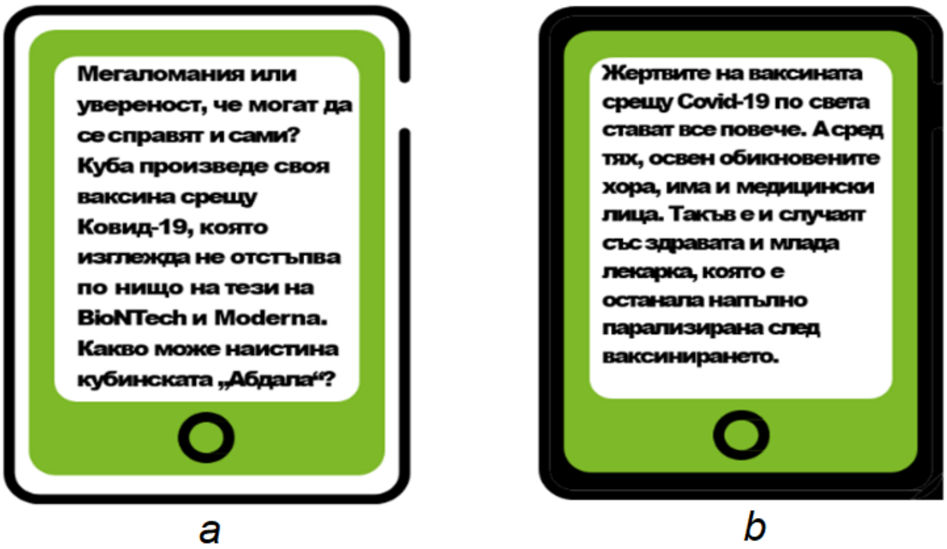
- Fake news has shorter texts.
- Less terminology is used.
- Shorter words are used.
- Less punctuation is used.
- Less nouns and more adverbs are used.

Our working definition for language markers is as follows: the specific changes in language with which the text of disinformation was created. These changes are a direct consequence of the intention of the creator to deceive the perceiver of the text. There are three different groups of markers – lexical (specific phrasing that is more biased and emotional), grammatical (evidentiality, reported speech and second and third person pronouns as tools for distancing from the fake information) and graphic (all-caps headings, many exclamation marks).

2.1.3. Novel Approaches to Language-Specific Characteristics of Fake News in Bulgarian

Previous research evidence suggests that a part of the language markers for detecting fake news are language-specific [10]. Wah Chu et. al. who researches Chinese and English conclude that the more complex the language base, the more different the language markers. We thus consider systematic research of Bulgarian language markers obligatory for final conclusions on this language. We hereby present limited evidence, based on Deep Syntax of two short paragraphs from a real and fake piece of information. The main goal was to achieve primary visualization of the trends in our initial observations on the corpus of the GATE Institute of Sofia for the purposes of the summer school.

Before the summer school, we did research on a sample of Bulgarian language material in a demo experiment. This included extracted Facebook posts from groups discussing issues around the global Covid-19 pandemic. The data were extracted using CrowdTangle and since the



Мегаломания или увереност, че могат да се справят и сами? Куба произведе своя ваксина срещу Ковид-19, която изглежда не отстъпва по нищо на тези на BioNTech и Moderna. Какво може наистина кубинската „Абдала“?

Жертвите на ваксината срещу Covid-19 по света стават все повече. А сред тях, освен обикновените хора, има и медицински лица. Такъв е и случаят със здравата и млада лекарка, която е останала напълно парализирана след ваксинирането.

Figure 3: Examples of fake news messages in Bulgarian

volume is big (300.000 posts in general), we did a subset which is extracted posts which we manually annotated using the annotation schema used in the report on Covid-19 misinformation by FirstDraft. [11] We will briefly discuss the experimental protocol. It includes the following steps: two paragraphs with relatively similar length are taken – one from the fake news corpus, the other from legitimate news website. Both types of texts are primarily sourced from social media. We compare their language characteristics, considering markers such as emotional language, distancing from the text, and graphic markers, and then we focus on Deep Syntax.

Example a. Megalomania or certainty that they can do it on their own? Cuba produced its own vaccine against COVID-19, that seems to be as good as BioNTech and Moderna's preparations. What can the Cuban Abdala really do? (Real news, Fig. 3a.)

Example b. The victims of the COVID-19 vaccine increase even more across the world. And amongst them, except ordinary people, there is also medical personnel. That is the case of the healthy and young female medical doctor, left completely paralyzed after having been vaccinated. (Fake news, Fig. 3b) The following features can be observed in the first paragraph:

- Emotional language – few adjectives are used.
- Distancing of the speaker – the text is in present simple tense.
- Cognitive complexity – one elliptical question with the verb to be, one complex sentence and one wh-question. Average depth of the sentence structure of the sentences is 4 levels.

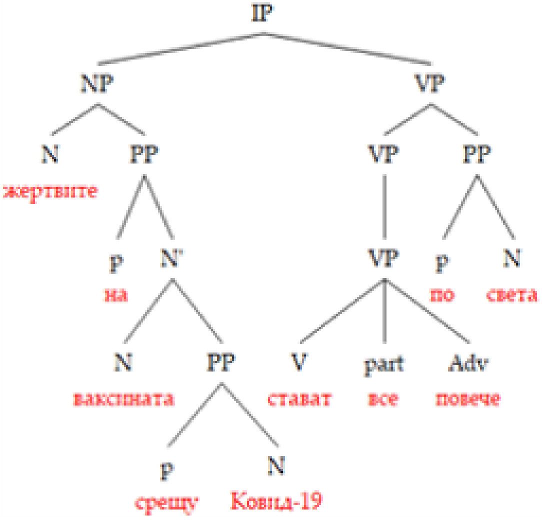


Figure 4: Tree-like sentence structure of the first sentence in the first (example from Fig. 3b) paragraph from the experiment. This visualization facilitates fine syntactic analysis and depth levels count.

The second paragraph shows deviations from the trends in the first one. In terms of emotional language – there are many adjectives and nouns with negative connotation. The distancing of the speaker is achieved by combining present tense and third person verbs. The cognitive complexity is realized by using two simple sentences with an inserted part, and one complex sentence. The average sentence structure depth is 4,33 levels, which marks a small increase in the depth in comparison with the first paragraph. The hypothesis of measuring scope and depth of the sentence phrase, that Horne and Adali [12] use in their research, and we also employ here, can be explained with the following clarification made by the authors: "The complexity characteristic is based on deep calculations for natural language processing so that we can detect the complexity of a title or article." Here we differentiate between two levels of complexity: of the sentence and of the word. To calculate the level of complexity of the sentence, we calculate number of words per sentence, as well as depth of the syntax tree, depth of the syntax tree of the noun phrase and of the verb phrase. The expectation is that the greater word count in a sentence and the bigger depth of the syntax trees, will result in bigger average complexity of the syntax structure. We thus measure for scope value number of words in the linear structure and for depth value – end number of constituents up to the IP (inflectional phrase).

2.1.4. Summary and Results

Our limited data shows that the language characteristics of fake news such as emotional expressions and distancing of the speaker, that are achieved by linguistic markers such as shocking titles, piling up of emotional words, adjectives, and adverbs, are visible. On the other hand, however, the complexity variable shows tendencies for systematic differences with English. The scope and depth parameters of the two paragraphs show significant differences that need further research until formation of plausible hypothesis for Bulgarian. It seems here, however, that fake information uses more complex structures with greater word count and more depth. This is due to fact that there are more simple sentences in the fake excerpt, and that in simple sentences depth is defined according to the levels of phrases, whereas in complex sentences the depth is defined according to levels of subordinate clauses.

The following trend for Bulgarian can be summarized: pilling up of modificators in the simple sentence implies complexity, but it's connected to the emotional influence of the text. For future work we plan to apply to the corpus of fake news tweets of the GATE Institute of Sofia improved BoW, semantic analysis and deep syntax models, aimed at systematically researching Bulgarian language of fake news (and more general of misinformation).

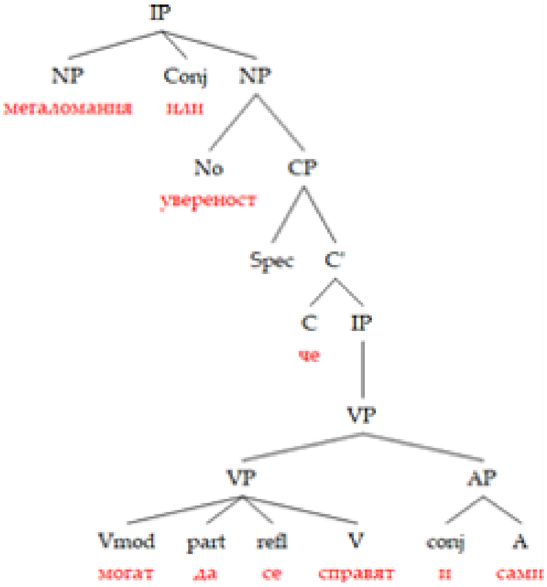


Figure 5: Tree-like sentence structure of the first sentence in the second (example from Fig. 3a) paragraph from the experiment. This visualization facilitates fine syntactic analysis and depth levels count.

3. Conclusions and Future Work

One of the achievements of the work and the delivery of the summer school was the discovery of new research and development areas. The feedback from the participants clearly indicates the usefulness and the desire to continue their involvement in this area. Here are some of the written feedback statements which illustrate the success of the school:

"At first, I started reading the news more often and 'testing' the new skills I learned during the school. My main goal has been achieved - the inspiration is there and although there is nothing visible and real at this stage, plans for their future development are all around me. The wonderful thing is that I have made some interesting contacts, and the potential future collaborations with speakers and participants are something particularly exciting."

"I shared with the audience that I was going to develop a lecture course, invite lecturers from Sofia University and participants from the school to meet with students at my university to share experiences, collaborate and partner in the future."

"I now report any fake news that comes my way, not pass it by as I used to."

"I will apply what I learned at the summer school to my work with library professionals"

The team preparing the school will continue its research to develop a tool identifying disinformation, with an initial case study of Covid-19 disinformation. We will also focus on education. The example of Finland, which has the highest media literacy index in Europe, is inspirational. More work is needed to deliver educational content for primary and secondary schools and teach students about disinformation and how to check claims. In order to stimulate critical thinking in students, we can teach them how to write research papers based on books in the library and other reliable sources. The librarians' community can be a valuable ally in this, and it was excellent to see that the school attracted the interest and the active engagement of library professionals. One particularly exciting aspect is the fact that identifying potential misinformation requires better knowledge of our own language and a critical eye for the way messages are written.

Acknowledgments

Funding: This research work has been supported by GATE project, funded by Operational Programme Science and Education for Smart Growth under Grant Agreement No. BG05M2OP001-1.003-0002-C01.

Author contributions: Milena Dobreva, Hristiana Krasteva and Silvia Gargova participated in the organization of the Summer School and wrote the paper. **Other participants** in the content creation for the school: Prof. Kalina Bontcheva (University of Sheffield), Prof. Petya Barkalova (University of Plovdiv), Irina Temnikova (PhD, computational linguist) and Ruslana Margova (PhD student and a practicing journalist).

References

[1] Digital 2021: Bulgaria, DataReportal – Global Digital Insights, 2021. URL: https://datareportal.com/reports/digital-2021-bulgaria.

[2] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, B. S. Silvestre, Social Media? Get Serious! Understanding the Functional Building Blocks of Social Media, Business Horizons 54 (2021) 241–251. doi:10.1016/j.bushor.2011.01.005.

[3] S. Vosoughi, D. Roy, S. Aral, The Spread of True and False News Online, Science 359 (2018) 1146–1151. doi:10.1126/science.aap9559.

[4] M. Lessenski, Media Literacy Index 2021. Double Trouble: Resilience to Fake News at the Time of Covid-19 Infodemic, Open Society Institute – Sofia, 2021. URL: https://osis.bg/wp-content/uploads/2021/03/MediaLiteracyIndex2021_ENG.pdf.

[5] J. Burkhardt, History of Fake News, Library Technology Reports 53 (2017) 5–9.

[6] J. Penchev, Bulgarian Syntax Government and Bindidng Theory, UPH University of Plovdiv "Paisii Hilendarski", 1993.

[7] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, B. Stein, A Stylometric Inquiry into Hyperpartisan and Fake News, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, volume 1, Association for Computational Linguistics, 2017, p. 231–240.

[8] P. Barkalova, Bulgarian Syntax. Configurational Analysis of Sentences, UPH University of Plovdiv "Paisii Hilendarski", 2019.

[9] P. Osenova, Linguistic Specificities of Rumours, in: Proc. Social Processes and Their Reflection in Language. Problems of Sociolinguistics 13, International Sociolinguistic Society in Sofia, 2018, pp. 225–231.

[10] S. K. W. Chu, R. Xie, Y. Wang, Cross-Language Fake News Detection, Data and Information Management 5 (2021) 100–109. URL: https://www.sciencedirect.com/science/article/pii/S2543925122000250. doi:https://doi.org/10.2478/dim-2020-0025.

[11] R. Smith, S. Cubbon, C. A. Wardle, Under the Surface: Covid-19 Vaccine Narratives, Misinformation and Data Deficits on Social Media – First Draft, 2020. URL: https://firstdraftnews.org/wp-content/uploads/2020/11/FirstDraft_Underthesurface_Fullreport_Final.pdf?x76851.

[12] B. Horne, S. Adali, This Just in: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire Than Real News, Proceedings of the International AAAI Conference on Web and Social Media 11 (2017) 759–766. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14976. doi:10.1609/icwsm.v11i1.14976.

Challenges in Event Annotation across Linguistic Levels

Preslava Georgieva, Iva Anastasova, Petya Osenova and Kiril Simov

Institute of Information and Communication Technologies - Bulgarian Academy of Sciences, Sofia, Bulgaria

Abstract

It is well known that the quality of the annotation process depends on the synchronization of various factors - the annotation scheme (**AS**), the type of the processed files, the way the information was presented, the goals of the workflow, the preparation of the annotators, among others. In the work process we faced a number of challenges that are often common for different types of annotation (morphosyntactic, semantic, etc.) and thus lead to annotation errors. These challenges are predominantly the following ones: ambiguity, uncertainty, lack of context, domain specifics. In this paper we present the problems during the event semantic annotation of mainly historical texts through the INCEpTION system. In addition to the general challenges during annotation, we also refer to metaphorical and methonymical uses in scientific texts as well as the gaps in the **AS**. By presenting every individual problem we try to reveal the causes of their occurrence and the possible solutions that are either already used in the process of work, or are going to be implemented in order to reach a solution for the problem.

Keywords

challenges, semantic annotation, INCEpTION, Knowledge Graph, Named Entities, Events

1. Introduction

In this paper we present the results from working on event annotation, the main aim of which is the creation of a Bulgaria-centric Knowledge Graph (see [1]). Although our ultimate goal is to create formalized annotation schemes for the all main areas in Social Sciences and Humanities, in the present we concentrate on historical documents. The data include research papers, biographical descriptions, history of geopolitical entities, big historical events like wars and rebellions, history of organisations, archive documents, descriptions of icons, etc. We included such a wide range of topics in order to verify the applicability of the current scheme and to identify the main event types within these documents. We envisage to test our current findings also in other areas of humanities such as literature, philosophy, anthropology, etc.

There are other considerations to be taken into account when an annotation scheme is designed. For example, which facts from the texts are considered to be important for the goals of the annotation; how easy it is for the annotators to recognise the instances of these facts given the variety of lexicalizations in the natural language; whether there exists an appropriate software system to well support the annotation process.

The annotation scheme consists of two parts: (1) Named Entity (NE) classes — the Entities that participate in a given Event, and (2) Event description — the main predicate expressing the Event as well the the roles of the participants within the Event.

The structure of the paper is as follows: In Section 2 we present some background information about the first steps in creating the annotation scheme. Then, in section 3, we present the work of other researchers comparing what we have in common and where we differ in our approach towards solving the annotation problems. In Section 4 the challenges are discussed in more detail that we faced in the work process with respect to both — NEs and Events. The final section concludes the paper and gives our insights for future work.

2. Background Information

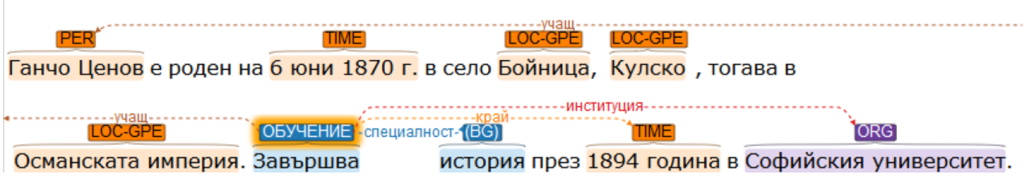
The software system we used for performing the semantic annotation, is INCEpTION[1], because it offers the possibility to model in connection various levels of linguistic annotation with open data like Wikipedia. Thus, INCEpTION perfectly corresponds to our goals of creating linked resources for Bulgarian. Moreover, in comparison with the BRAT annotation tool[2] and based on some previous experience of annotating with WebAnno[3], we decided that being a contemporary combination of both — BRAT and WebAnno — INCEpTION has more sophisticated architecture and is easy to work with by non-programmers. The current annotation scheme is an extension of the one reported in Laskova et al., 2020 [2]. The changed annotation scheme differs from the previous one in the following aspects: (1) the style of annotation was changed from arrow-based to span-based; (2) we have added some new kinds of Named Entities, new types of Events and roles. At the beginning of our

preslava@bultreebank.org (P. Georgieva); iva@bultreebank.org (I. Anastasova); petya@bultreebank.org (P. Osenova); kivs@bultreebank.org (K. Simov)

[1] https://inception-project.github.io/

[2] https://brat.nlplab.org/introduction.html

[3] https://webanno.github.io/webanno/



Ганчо Ценов е роден на 6 юни 1870 г. в село Бойница, Кулско , тогава в Османската империя. Завършва история през 1894 година в Софийския университет.

Figure 1: Arrow-based annotation style.

work, the annotation style was based on connecting the verb, which acts as a trigger (lexical stimulus unit) for the Event type — **arrow-based annotation style** (see **Fig. 1**). Arrows are pointing to the elements of the text which correspond to the roles of the Event. This annotation type (arrow-based) also allows one to link text segments that need to be annotated under a common role/Event, but are made discontinuous by other text.

Very identical to this style of annotation is the one presented by Mostafazadeh et al. [3], as it is applied through the BRAT annotation tool. The visual front end of BRAT was used as a basis for the creation of WebAnno [4], on which the INCEpTION platform is based. This determines the proximity of the view and the style of the annotation applied at the beginning of our work to that presented in [3], which is focused on the semantic annotation of "comprehensive set of temporal and causal relations between events" through CaTeRS (Causal and Temporal Relation Scheme). However the disadvantage of this annotation style (arrow-based) is that a verb is not always presented in the annotated segments and therefore no trigger exists for the Event type. It is the inability to cover non-verbal events that has been noted as well in the article of Mostafazadeh et al. [3] as one of the shortcomings that emerged during the annotation process.

Another problem is that sometimes the verb reflecting the event and the text segments that explicitly represent the roles to it are located at a great distance, which complicates their connection by the annotator, and in rare cases makes tracing the arrows a difficult task. These are some of the main reasons why the annotation style was changed to span-based. With this annotation type, the roles for the event no longer need to be marked only on their explicit manifestations in the text, but it is also possible to place them on pronouns, which through the field Refers-to in the annotation panel, refer to what they replace.

As Laskova, Osenova and Simov [2] point out, the **AS** has two levels — Named Entities (NE) level and the level of semantic roles and Events. At the beginning, the scheme, based on CIDOC-CRM [4] ontology and correlated with the entries of roles and Events in FrameNet [5], had 11 types of NEs and 9 types of Events. We currently have 16 types of NEs and 39 types of Events. The group of NEs has expanded in three directions — (1) creation of a more generalized type to an already existing but very specific one, (2) deepening of the hierarchy for existing general NEs and (3) creation of a separate category of NE, unrelated to the others. Some of the basic types of Named Entities are presented within Table 1. In the first case, in the original version of the **AS**, the Named Entity **JUR** was introduced to mark legal texts (e.g. Berlin Treaty). In the work process it came clear that **JUR**-role is very specific and does not cover the different types of texts and publications that are found in the annotated materials — books, collections, poems, novels, short stories and more. This showed the need to introduce also a more generalized type of Named Entity — **DOC**.

Table 1
Some of the Named Entity labels in the annotation scheme.

Label	Description
DOC	for various texts, including documents (except for juridical documents, annotated with — they take the label **JUR**)
JUR	for juridical documents: laws, regulations, or parts of them like articles, etc.
LOC	for locations/places — natural or man-made, which are not geopolitical units like mountains, lakes, streams, etc.
LOC-GPE	for geopolitical units (countries, regions, cities, cantons, etc.)
PER	for people (existing in reality or fictional ones)
PER-GPE	for nationalities (Bulgarian) or the birth place, the place where people live (Bulgarian (citizen), etc.)
PER-GRP	for groups of people that cannot be described as **PER-GPE** or **PER-LOC** (Slavs, etc.)
. . .	. . .

In the second case we have the opposite process —

[4] http://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_7.2.pdf

[5] https://framenet.icsi.berkeley.edu/fndrupal/

hyponymic variants are created to an already existing entity with more general semantics, as the processed texts require a higher level of specificity in these cases. An example of this is the Named Entity **PER**, which generally refers to individuals. Its existing hyponym is **PER-GPE** — a person associated with a geopolitical entity (e.g. Italian, Bavarian, Roman, etc.) We added **PER-LOC** — a person associated with a location that is not a geopolitical entity (e.g. mountaineer), and **PER-GRP** — names of groups of people who cannot be defined by the previous two tags (e.g. Slavs, Christians). Formally they are Named Entities, but actually present expressions for groups of people, united by a common feature (religion, political ideas, etc.) and are often used metonymically.

In the third case, we introduced new NEs, which do not belong to any of the already existing categories, but are needed for the annotation. These are **MSR** (for quantities) and **MSC** (for objects that do not fall into any of the other categories).

New labels have also been introduced at the Event level to cover events, which are specific for biographical descriptions and historical texts, as well as to avoid the use of too general Events. The type of Event level changes is similar to that of the NEs level. For example, at the earlier stage of the annotation process, in cases involving the exchange of funds (money and goods) between two or more parties, there was only a *Charity* Event (non general case). In the annotation process on biographical texts, many examples showed that the exchange of funds is not always for the purpose of helping people in need and an act of good will, but often refers to the support of a cause and is perceived as part of (official) obligation of the one giving the money. This led to the introduction of a new type of Event — *Funding*, which would cover these cases as well, as they have important informative value. Table 2 lists some of the most common events with their roles. You can see, for instance, the description of *Giving Birth* in FrameNet - "A Mother and Father produce a Child or an Egg" [6] whereas in our sheme it is simply "The birth of a human". Most of the descriptions and/or roles have been adapted to our **AS**.

3. Related Work

This section is not meant as an exhaustive review of the previous works related to NE and event annotation. It rather gives some examples of how other NLP groups approached these types of annotations: considering the basic list with respect to NEs, and considering the time localization of the events.

Similarity between annotation schemes is mostly found at the level of NEs, where the goals of annotators can generally be reduced to "the idea of categorizing the world into semantic classes" [5]. S. Kübler and H. Zinsmeister discuss the semantic annotation of corpora, presenting more broadly the work on two specific ones — the corpus of TüBa-D/Z treebank of written German (built on the basis of Verbmobil treebank of spoken German) and OntoNotes corpus of English, Chinese, and Arabic. The core of their NEs list consists of the labels PER, ORG, GPE, LOC, TIME. These labels are also part of our **AS**. Thus any additional labels can be added depending on the specific needs and objectives of each project. As part of the complementary NEs in our **AS** we can mention DOC, JUR, PER-GRP, SUM. Their appearance is mainly due to the type of the processed texts - articles and other materials related to historical events (battles, wars, peace treaties, alliances, etc.) and personalities (rulers, poets and writers, revolutionaries, etc.) that are important to Bulgarians.

Another part of the annotation process are the challenges during the annotation itself and the ways to deal with them. Common features here can be seen even when different language levels are annotated. For example, problems such as ambiguity, the presence of more than one possible solution (uncertainty), lack of sufficient context are also noted by Beck, Booth, El-Assady and Butt, see [6]. Although the problems described by them are manifested in POS tagging, the fact that they are related to the peculiarities of the processed texts, causes them to be found at different language levels. The possible solutions to the problems are also almost identical - annotation with the most probable option, marking with a label "miscellaneous"/"others" (in our work this type of solution refers to annotation with a more general Event, for example) or omission (leaving without annotation) in case the given text fragment has no information value.

Since we work with historical texts, one of the most important characteristics is the orientation of events in exact temporal lines, their marking with exact time characteristics. There is a lot of experience in this field. [7] present TimeBank-Dense — a much improved version of TimeBank. Most of the previous time annotating systems rely mostly on the annotators (TimeBank and TimeEval, for example), whereas TimeBank-Dense prohibits the freedom of the annotator and provides an automated processing of time indicators. However, in our work we still rely on the judgment of the annotator, although that leads to more discussions about annotation solutions and therefore is much more time-consuming. We still try to follow guidelines in order to minimize errors and stay consistent with our annotating decisions. Keep in mind that we aim for a bigger goal - we want to point as exact time characteristics as possible, not just annotate the time in an Event chain. However, we follow the principle outlined by [8] that the placement of temporal

[6] https://framenet2.icsi.berkeley.edu/fnReports/data/frameIndex.xml?frame=Giving_birth (last visited: 05.11.2021)

Table 2
Some of the event labels in the annotation scheme.

Event	Roles
Donation	**donor** (person or organization) **recipient** (person or organization) **theme** (object) **mediator** (person or organization, it could be fund) **period–of–iterations** (time: the length of time from when the event denoted by the target began to be repeated to when it stopped) **goal** (situation: the goal for which the donor gives the theme to the recipient) **time** **place**
Giving–Birth	**brought–into–life** (the new self-motile creature produced from the mother and father) **parents** (the mother and father expressed together, for example “his parents” or “Penka and Toncho Ivanovi”) **mother** **father** **place** (the birth place — usually the name of a city, continent (rarely), country or hospital) **time** (the time of birth — usually it is a date, but can include hours, or it is just month and year)
Moving–in–Place	**agent** (a person) or theme (another type of object) **coagent** (another person or group of people the agent is moving with) **move-from** (the place from which the agent or the time moves) **move-to** (the place where the agent or the theme moves to) **time/beginning/end/duration** **purpose** (a situation or another event which causes the moving) **goal** (a situation/event to be achieved with the moving)
Kinship	**alter** (the person who fills the role named by the kinship term with respect to the ego; requires ego; if this is present, relatives are not used) **ego** (the person from whose perspective the kinship relationship is defined; requires alter; if this is present, relatives are not used) **relatives** (the combination of alter and ego together: for example “Ivanovi brothers”; if present, alter and ego are not used)
Characterisation	**characterised** (a person, organization, etc.) **characteristic** **evaluator** (the one who points out the characteristic: “Gorbachev insists for “second perestroika”, Russia considers him Judas.”) **source** (a document containing the characteristic)
. . .	. . .

features should cover the whole text and not just individual sentences (although formally INCEpTION works with segments for better visualization).

Regarding the role of the annotators, TimeBank relies almost exclusively on them, as temporal relations are only made when the relation is judged salient by the annotator. In a sense the complete trust of the annotator decision can mislead the data and their results - there is always the case with the individualization of decisions. On the other hand, the automation of the process can lead to loss of information, as it can only work on simpler cases and it cannot for sure do something about cases of exception. Like TempEval, we strive to cover the relations between all Events and Time not only at the level of the sentence, but at the whole text/piece of text. We rather work according to the method of [9] - we annotate multi-sentence segments of text rather than individual segments in order to understand the temporal flow of discourse. Also we follow the path of TimeBank Corpus, because we want to annotate "not simply times and dates, but events and temporal relations between events" [10].

The purposes of our annotation are related to preparing as many texts as possible, (semi)automating this process and linking the processed data, on the basis of which the Bulgarian Knowledge Graph should be built. For our purposes, it is necessary that the Events and the roles (participants) should be applicable to the processed information — for example, the biographies demand Events connected with the life of a person like *Birth*, *Death*, *Relation*, *Living*, *Education*, etc., historical documents are most likely connected with the beginning/end of an event, its participants, causal relationships, etc., important objects and places from history are often described by location, exterior and functions. The benefits of building an **AS** and its manual on a small but secure set of already established models and annotation formats have been discussed by Pustejovsky and A. Stubbs [11]. A leading advantage is that adapting, supplementing and expanding these already existing resources facilitates interoperability and the provision of usable data to other researchers working on similar tasks.

The annotators are also an important part of the annotation process — on one hand, their knowledge and schooling of language, and on the other hand, their familiarity with the processed texts domains. In most cases the information in the documents is not explicitly presented, and thus requires additional research from the annotator, processing and interpretation of the available data in a certain piece of text.

4. Challenges of the Annotation

In this section some problems are presented that we encountered during the annotation of about 350 documents (or 400 000 tokens) collected by CLaDA-BG partners — specialized articles, papers, biographies, archive documents, catalogues, etc., as well as freely available online resources such as Bulgarian Wikipedia and various dictionaries. The documents from Wikipedia were selected to show diversity of texts, related to different periods of Bulgarian history as well as different topics. The problems we face in the workflow are mainly in three directions: **context-related**: ambiguity, sign-referent-NE relation, missing context; **representational**, related to the way the information is presented with the scheme tools – two competing solutions available, predicate-Event relation; and **schema-based** — the scheme allows for more than one interpretation or does not cover some important facts; here we deal also with the identified gaps in the **AS**.

4.1. Challenges on Named Entities Level

The main problems with Named Entities in our annotation process are connected with the relations between name (sign) in the text, the actual referent of the name in reality and the NEs we choose by considering the context in which the sign is used. It is not a surprise that these names have similar behaviour as the common words — regular polysemy, homonymy. Here we present and discuss some examples from our annotation process, where we distinguish the sign level — the segment of the text that determines the name of a given entity, the Named Entity level which in the current **AS** is a combination of a sign and the Named Entity category as represented in Table 1, and the referent level — what the sign denotes in the world.

Ambiguity. The ambiguity on the Named Entities level occurs when there can be two or more different labels for the same element. For example:

(1) *the Serska detachment of the old voivode Stoyo Kostov formed near* **Rila Monastery**

In this sentence we have an Event of *Organization* in which the Rila Monastery, usually considered as an ORG, gets a **LOC** label, because of the context, as it represents the location of the establishment of organisation.

(2) *...the abbot of* **Rila Monastery** *Theodosius I Rilski*

In the second sentence we have a *Being-Employed* Event and by considering the context Rila Monastery is labeled as **ORG**, because it represents the body in which the occupation takes place. In both sentences the annotator should be aware of the context.

Relation Between NE and Its Referent. In Figures 2, 3, 4 we present three schemes which define the links between NEs, referents and signs. The relation between NE and its referent in reality depends on the different aspects of the context such as the attitude of the speakers, the temporal anchor in which it is used and other factors.

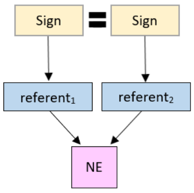


Figure 2: The same name is found in the text, i.e. two identical signs, but they have different referents and are marked with the same NE.

The first scheme is depicted in **Figure 2**. It demonstrates the case when the same sign (name) corresponds to different referents that are expressed by the same NE. Here are the example sentences:

(3) *With the treaty* **Bulgaria** *gives away South Dobrogea to Romania.*

(4) **Bulgaria** *and Frankish Empire start sharing borders.*

In the first sentence the name “Bulgaria” has for a referent the administrative unit of the Kingdom of Bulgaria (it existed from September 22, 1908 until September 15, 1946). In the second example, “Bulgaria” is used in the temporal context of the IX century (First Bulgarian Kingdom), that means it has different borders and administrative structure. However, in both cases, we have an independent administrative unit, so the two uses of the name “Bulgaria” are designated with the same Named Entity — **LOC-GPE**.

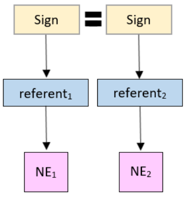


Figure 3: The same name is found in the text, i.e. two identical signs but they have different referents, and are marked with a different NE.

The scheme depicted in **Figure 3** shows cases in which same signs have different referents and different NEs. For instance, in the sentences (5) and (6) the name **Bulgaria** refers to the geographical region and to the state respectively. In sentence (3) it is annotated as **LOC** and in sentence (4) it is annotated as **LOC-GPE** because of the context in which they are used.

(5) *they aim to bring the liberation of* **Bulgaria** *with the help of the Russian troops*

(6) *after his return to* **Bulgaria**, *Sandanski is disappointed with his previous activity*

The last two examples — sentences (7) and (8) (repetition of sentence (6)) demonstrate the last scheme:

(7) *returns him to* **the Principality** *for treatment*

(8) *after his return to* **Bulgaria**, *Sandanski is disappointed with his previous activity*

In both examples, the referent is the same — the administrative unit of the Principality of Bulgaria, which existed from 1879 until 1908. Therefore, they are marked with the same NE — **LOC-GPE**, but the signs are different. This last case is when the same object in the world could have different names and they are annotated with the same NE category. The scheme is depicted in **Figure 4**. One example here is the usage of the exact name of a country and its colloquial name.

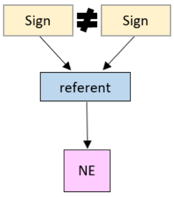
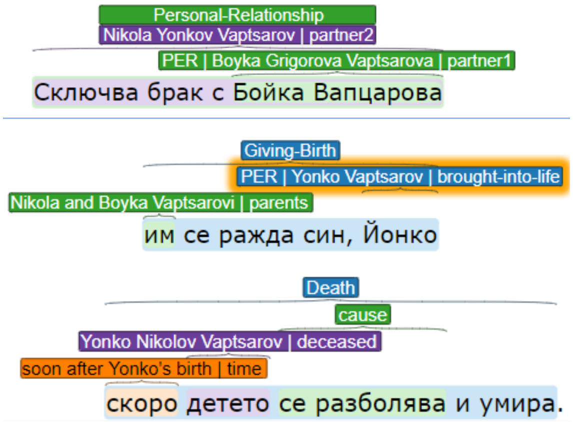


Figure 4: Different name is found in the text i.e. two different signs, they have the same referent and are marked with the same NE.

Although not all the cases are problematic, the annotators have to recognise them in the text and to annotate them correctly. This is important especially when a mapping to the knowledge graph identifiers is performed — since these identifiers are unique for each distinct referent.

4.2. Challenges on Event Level

An event can be expressed by a potentially endless set of paraphrases which semantically refer to the same annotation piece. In order to annotate them in a consistent way we need to have enough information within the context. The appropriate annotation also depends on the knowledge or interpretation of the annotators themselves. Thus we need strict rules stated in the guidelines how to interpret the text to be annotated and also how to treat the truly ambiguous cases. Here we present the main problems we encountered during the annotation process.


Personal-Relationship
Nikola Yonkov Vaptsarov | partner2
PER | Boyka Grigorova Vaptsarova | partner1
Giving-Birth
PER | Yonko Vaptsarov | brought-into-life
Nikola and Boyka Vaptsarovi | parents
Death
cause
Yonko Nikolov Vaptsarov | deceased
soon after Yonko's birth | time

Сключва брак с Бойка Вапцарова

им се ражда син, Йонко

скоро детето се разболява и умира.

Figure 5: *[Nikola Vaptcarov]* **marries** *Boyka Vaptcarova and they* **give birth to a son**, *Yonko, but soon the child gets sick and* **dies**.

Lack of enough context. This problem is related to the types of texts we focus on (about historical events and personalities), which require a temporal specificity. Usually the annotated documents belong to a large set of related materials, but the annotator deals with it in isolation. Event annotation requires the construction of event chains within the timeline. It is of great importance for structuring the Events included in the Bulgarian Knowledge Graph. However, it is the time boundaries in which an event occurs that are most often skipped or short-spoken in texts, especially when the texts are related to events far back in time. In these cases, a more vague specification of the temporal denote is often preferable. In most cases the events are reported in the text with relative expressions like *next year, afterwards, in five months, etc.* which need some anchors in order to receive appropriate interpretation during the annotation.

The annotator is the one to provide information about the time of the event (through Refers-to – a special layer for referring to other sources) when possible - such as to consult other sources — other documents in the set, sometimes additional sources like encyclopedic information or to use general knowledge about the events. Let us take for example **Figure** 5.

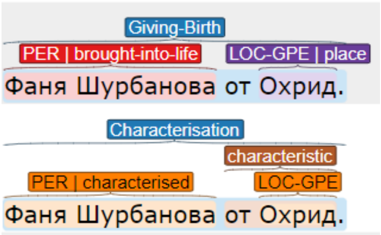
Here we see three separate Events that occur sequentially. Since the first one (*Relation*) has no specified time marker, and this marker cannot be obtained from the available context, the next Events (*Birth*; *Death*) depending on it also remain temporarily non-specific. The information gaps can be obtained only with the help of additional inquiry provided by the annotator.

In some cases it is not possible to add an exact anchor either and thus only time intervals are presented.

Very often the author of the text might not use a precise expression and that would allow more than one interpretation. In such cases, depending on the presuppositions in the text, two or more annotations are possible.

One example of this case is the interpretation of adverbials for locations expressed as prepositional phrases like in the following sentence (**Figure** 6):

"Fanya Shurbenova **from Ohrid** *[was a delegate at the meeting]".*



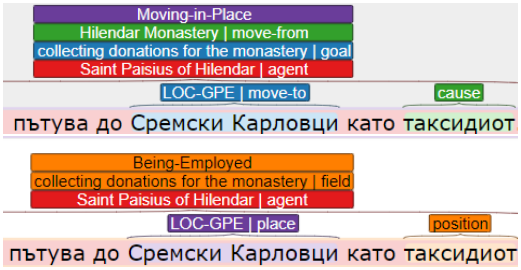
Фаня Шурбанова от Охрид.

Фаня Шурбанова от Охрид.

Figure 6: A case with more than one correct annotation: *Birth* or *Characterisation*.

In sentences of the type "someone comes from somewhere", some annotators prefer to use the Event *Birth*, others, however, consider that "comes from" can mean a living place, and not a birth place — there are many options, so in general, we use *Characterisation* Event (or another Event without a clear specification, a general one). Another example of the same kind is when the sentence expresses two predications, one of which is a secondary.

In such cases different annotators prefer to make explicit only one of the predications. For instance, in a sentence like *"In 1758 he travels to Sremski Karlovci as taxidiot* [7]*."* the noun "taxidiot" plays a twofold role — in **Figure** 7 the first part is for *reason* and the second part is for *occupation position*. These roles require different types of Events:


Moving-in-Place
Hilendar Monastery | move-from
collecting donations for the monastery | goal
Saint Paisius of Hilendar | agent
LOC-GPE | move-to
cause
Being-Employed
collecting donations for the monastery | field
Saint Paisius of Hilendar | agent
LOC-GPE | place
position

пътува до Сремски Карловци като таксидиот

пътува до Сремски Карловци като таксидиот

Figure 7: Event *Being-Employed* where "taxidiot" is the occupation position of the traveller — which could be an important fact about him.

Metonymic and metaphorical readings. Metonymic and metaphorical readings of some verbs and other predicates are frequent in language. They could be problematic when the figurative and the literal meanings could trigger two different types of Events in the annotation scheme.

In cases like these the annotator could mix the two annotations of *Birth* and *Establishment* (of group of people or organisation) because within the annotation scheme the two Events share many common roles. These cases can be also identified through consistency checks at later stages of the corpus development.

Figure 8 is an example of using the deverbal noun from the verb "to be born" to denote the establishment of a country.

Regular metonymy that is presented in examples like *"Sofia declared war on Turkey in 1912 and started the First Balkan War."* is relatively easy to recognize. The annotator has to pay attention that in this case the Named Entity "Sofia" represents the government of Bulgaria instead of a geopolitical location for the city of Sofia and we insert a comment that this is a metonymic use without changing the type of NE (we leave it as a LOC-GPE).

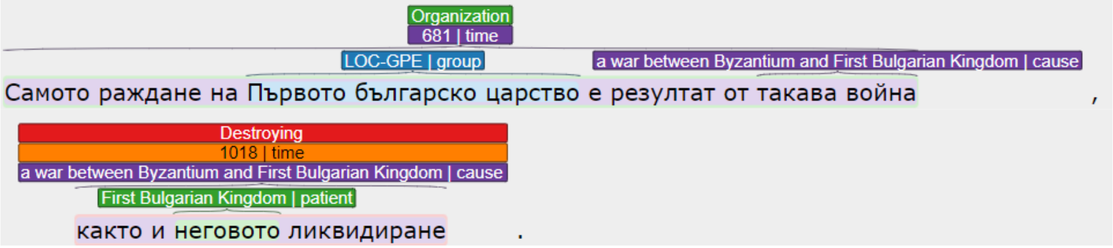
4.3. Gaps in the Annotation Scheme

The gaps in the annotation scheme are unavoidable at least in the following directions: missing types of Named Entities, missing types of Events, and missing text interpretations.

The current annotation scheme was designed with respect to a specific set of documents and thus many types of texts in the area of humanities remain uncovered or only partially covered by it.

We do not have specific examples of missing types of NEs. We think that the reason for this is that NEs are a much better established domain of annotation with a good label coverage.

[7] Taxidiot — A monk who travels to raise money for a monastery.


Organization
681 | time
LOC-GPE | group
a war between Byzantium and First Bulgarian Kingdom | cause
Destroying
1018 | time
a war between Byzantium and First Bulgarian Kingdom | cause
First Bulgarian Kingdom | patient

Самото раждане на Първото българско царство е резултат от такава война ,

както и неговото ликвидиране .

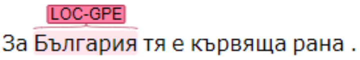
Figure 8: *"The birth of the First Bulgarian Kingdom is a result of such war, as well as its liquidation..."* (Metonymic and metaphorical readings.)

Still there are too general types like *Product names* (applicable in many cases where a more detailed hierarchy could be created) and *Miscellaneous* which are used relatively rarely at the moment.

Linking between the Text Type and the Annotation Scheme. This problem most often applies to philosophical and literary texts that mainly rely on labels interpretation, i.e. they represent the large percentage of exceptions in the annotation.

The scheme usually does not apply to them, since they contain metaphors, figurative uses, for which the only possible option is to leave a comment about their figurative nature, but annotate them only as if they are used in literal meaning. However, the information used for our annotation purposes (processing texts with specific information about important events in Bulgarian history and culture) is rarely found in such types of texts and vice versa.

When we encounter such a text in our documents, we skip it, because either the appropriate labels and mechanisms for its annotation are missing in our **AS**, or there is no fact-based information we can extract. For example see **Figure** 9:



За България тя е кървяща рана .

Figure 9: *For Bulgaria it is an open wound.* (Challenges with the linking between the Text Type and the Annotation Scheme.)

"Open wound" is an expression that means "reminding someone about an upsetting experience in the past which they would prefer to forget", it is an idiom. Thus it has a figurative meaning but with our **AS** we could not represent that characteristic, although we can annotate it with the Event *Characterisation*, for example. This case shows the need of modeling complex metaphors. However, we plan to explore this challenge in future.

The Usage of More General Annotation Solutions The majority of our problems are with missing types of Events within the current annotation scheme. The reason for this is the huge set of possible events in which the corresponding Named Entities could participate. For such cases we have used three more general Event types denoted by labels: *Characterisation*, *Have-Parts*, *Event*. The first two were included at the beginning only for specific cases, but over the time they were overloaded with additional examples. For instance, at the moment we have more than 2500 examples annotated as *Characterisation*. We are planning to detail the annotation scheme in order to cover more specific cases. **Figure** 10 presents a typical example.

5. Conclusions and Future Work

The problems and challenges stated above require a careful review and solutions. We aim to achieve this by making minimal changes to the **AS**, but also we want to avoid the risk of over-detailing the scheme, because it could lead to more errors in the future annotation process and to deficiency in the annotated data.

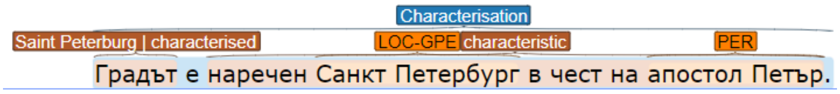
For example, at the moment we do not have an Event for 'naming something/someone'. Because of the importance of the naming Event we are currently extending the scheme in this direction. An example of this kind would be an Event for 'Naming something/someone' with the following specifications:

Naming Event

> *A person or an object is named after something/ someone and/or gets a nickname.*

Roles:

- **Name-Giver** — a person or an object, who gives another person or object a (different) name/nickname; may not be explicitly presented
- **Name-Recipient** — a person or an object, receiving (different) name/ nickname



Градът е наречен Санкт Петербург в чест на апостол Петър.

Figure 10: *Characterisation — "The city was named Saint Petersburg in honor of apostle Peter."* (Compromised Annotation Solutions).

- **Name** — including a nickname
- **Named-After** — the person or the object the recipient is named after
- **Time**
- **Place**
- **Containing-Event**

Please note that Time and Place accompany each event frame where they are adjuncts. Only in part of the events they are core arguments like in the event 'entering', for example.

The next step for us is the curation of all annotated files and the observations over concordances on the various Event types — especially the very general ones. These aim not only to eliminate the errors and suggest more appropriate solutions, but also to describe the event complexity per domain. And we also have an idea to expand the annotation of time indicators, similarly to TimeBank-Dense [7], in order to make relations between sentences we have already marked as carrying important information. This will allow us to arrange events chronologically, especially when definite time markers are missing.

A necessary step in our future work is to add identifiers to the Named Entities (i.e. to provide a Named Entity Linking), and thus to solve the ambiguity problem. We are reaching the stage when the manually annotated documents will be used to train automatic semantic analyzers.

Acknowledgments

This work is funded in part by CLaDA-BG, the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH, Grant number DO01-377/18.12.2020.

We wish to extend our special thanks to Nikola Ikonomov, Liudmila Dimitrova and Bilyana Borisova for their work on the annotated documents and also for their participation in the discussions about the problems we are mentioning above.

References

[1] K. Simov, P. Osenova, Integrated language and knowledge resources for clada-bg., Selected Papers from the CLARIN Annual Conference 2019 (2019) 137–144.

[2] L. Laskova, P. Osenova, K. Simov, Towards an interdisciplinary annotation framework: Combining nlp and expertise in humanities., in: Proceedings of CLARIN Annual Conference 2020, 2020, pp. 141–145. URL: https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf.

[3] N. Mostafazadeh, A. Grealish, N. Chambers, J. Allen, L. Vanderwende, CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures, in: Proceedings of the Fourth Workshop on Events, Association for Computational Linguistics, San Diego, California, 2016, pp. 51–61. URL: https://aclanthology.org/W16-1007.

[4] S. M. Yimam, I. Gurevych, R. Eckart de Castilho, C. Biemann, WebAnno: A flexible, web-based and visually supported system for distributed annotations, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 1–6. URL: https://aclanthology.org/P13-4001.

[5] S. Kübler, H. Zinsmeister, Corpus Linguistics and Linguistically Annotated Corpora, Bloomsbury, 2015.

[6] C. Beck, H. Booth, M. El-Assady, M. Butt, Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias, in: Proceedings of the 14th Linguistic Annotation Workshop, Association for Computational Linguistics, Barcelona, Spain, 2020, pp. 60–73. URL: https://aclanthology.org/2020.law-1.6.

[7] T. Cassidy, B. McDowell, N. Chambers, S. Bethard, An annotation framework for dense event ordering., in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 501–506. URL: https://aclanthology.org/P14-2082.pdf.

[8] N. Chambers, T. Cassidy, B. McDowell, S. Bethard, Dense event ordering with a multi-pass architec-

ture, Transactions of the Association for Computational Linguistics 2 (2014) 273–284. URL: https://aclanthology.org/Q14-1022.

[9] P. Bramsen, P. Deshpande, Y. K. Lee, R. Barzilay, Inducing temporal graphs., in: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing), Association for Computational Linguistics, 2006, pp. 189–198. URL: https://aclanthology.org/W06-1623.pdf.

[10] J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gauzauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro, The TIMEBANK corpus, Corpus Lingusitics (2003) 647–656. URL: http://ucrel.lancs.ac.uk/publications/cl2003/papers/pustejovsky.pdf.

[11] J. Pustejovsky, A. Stubbs, Natural Language Annotation for Machine Learning, O'Reilly, 2010.

WordNet in a Contrastive Bulgarian-English Perspective: Hierarchies and Linguistic Idiosyncrasies

Zara Kancheva, Ivaylo Radev, Barbara Miteva and Georgi Georgiev

Institute of Information and Communication Technologies - Bulgarian Academy of Sciences, Sofia, Bulgaria

Abstract
The paper reports on an effort to reconsider the representation of the BulTreeBank-WordNet (BTB-WN), a wordnet for Bulgarian, and its interlingual mapping to the Open English WordNet (OEW). While connecting the two wordnets, differences between the languages and between the lexicographic approaches emerge and it is our task to deal with these issues in order to preserve those differences and construct appropriate relations between the two languages. The article views the latest version of the BTB-WN – 4.0, and focuses on the consolidation of meanings in synsets, verification of the inherited from the OEW structure and relations and addition of new interlingual relations and relations between BTB-WN synsets. The paper shows our work approach and the evolution of decisions made when creating and expanding the BTB-WN, as well as our idea for introducing lemma markers which will provide additional linguistic information for the members of the synsets.

Keywords
Bulgarian, English, WordNet, Lexical Semantics, Semantic Relations, Linguistic Idiosyncrasies

1. Introduction

The lexical resource wordnet is viewed as one of the most popular and used providers of language data in the field of NLP for more than 50 languages. [1] Wordnet can be described as a kind of thesaurus containing groups of word senses and labels for the semantic relations among them. The role of wordnets in NLP tasks is to provide lexical and semantic data for tackling word-sense disambiguation (WSD), relation extraction, named entity (NE) and multiword expression (MWE) parsing, machine translation, etc.

As it was reported in a previous paper [1] the Bulgarian BulTreeBank WordNet (BTB-WN) has been created in three different ways: (1) by manual translation of English synsets from Core WordNet – a subset of Princeton WordNet (PWN) [2][2] – into Bulgarian. This step ensures comparable coverage between the two wordnets of the most frequent senses; (2) by identification of senses used in Bulgarian Treebank BulTreeBank (BTB); where the identified senses have been organized in synsets for the BulTreeBank WordNet and the newly created Bulgarian synsets were being mapped onto the conceptual structure of PWN. In this way, the BTB-WN was extended with real usages of the word meanings in texts. Also, the coverage of the core and base concepts for PWN has been evaluated over a Bulgarian syntactic corpus; (3) by sense extension, which includes two activities: a) detection of the missing senses of processed lemmas in BulTreeBank and adding them to the BTB-WN, and b) a semi-automatic extraction of information from the Bulgarian Wiktionary mapped to synsets from PWN and then manually checked. In both steps the interlingual relation between synsets in BTB-WN and PWN was maintained in order to provide transfer of information from PWN to BTB-WN such as lexical and semantic relations, and to support multilingual applications. When the development of PWN was suspended, we transitioned to the OEW [3].

Our work on BTB-WN started by employing the `expand model` [4] where a monolingual wordnet is created by transferring and translating synsets and structure from an existing wordnet – PWN in our case. We chose this approach because we did not have a structure of our own wordnet and the task to build it is very time and resource consuming. On this stage of the work we identified the need of some edits in the English structure. We are now reconsidering our approach and we are moving in the direction of the `merge model` where a monolingual wordnet is created by building the synsets and the structure from scratch and later mapped via interlingual relations to other wordnets (OEW for the purposes of our work).

Currently, BTB-WN 4.0 consists of about 25 250 synsets containing 50 000 lemmas and with this coverage comes the time for making the resource available – there will be a public version of the wordnet within the CLaDA-BG project accessible via the CLaDA-BG web portal. Some earlier versions of BTB-WN are available in Open Multilingual WordNet. [3]

Since 2020 we have changed the software system by

zara@bultreebank.org (Z. Kancheva); radev@bultreebank.org (I. Radev); barbi.miteva@gmail.com (B. Miteva); georgi_georgiev070@abv.bg (G. Georgiev)

[1] `http://globalwordnet.org/resources/wordnets-in-the-world/`

[2] The Core WordNet contains the 5000 most frequent synsets of PWN. `http://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt`

[3] `http://compling.hss.ntu.edu.sg/omw/`

which we edit the BTB-WN. Earlier we have exploited CLaRK System which is an XML-based system for language resources development [5] but recently we have started to use a system implemented within the CLaDA-BG project called bf CLaDA-BG-Dict. It is specially developed for the creation of BTB-WN and its mapping to OEW. bf CLaDA-BG-Dict supports a more global view on the consistency of the lexical data and the relation structure of BTB-WN. In this way, the editors are allowed to check synsets as part of the whole graph for the first time – there was no such option in the previously used software.

The goal of making BTB-WN accessible via web portal prompted us to work on new updated and refined version of the resource. With the help of bf CLaDA-BG-Dict we did extensive checking on the following aspects of BTB-WN: (1) the definitions of the synsets were extended with details, especially the ones belonging to synsets for adjectives and adverbs, which typically do not have thorough definitions in the explanatory dictionaries (for this purpose a comparison of several dictionaries was done, combined with a check of usages in the BulTreeBank corpora and in the internet); (2) the list of synonyms for each synset – several synonym dictionaries were compared for this task (for the contradictory cases we did a verification with usages in the BulTreeBank corpora and in the internet); (3) the mapping to the OEW; and (4) list of example sentences (they were excerpted from the BulTreeBank corpora and other freely available internet sources – news websites, Wikipedia, etc.). We are hoping to reach a state where each synset has multiple examples – not only with the lemmas, but with all word forms.

Here we report on some reconsideration over the structure of BTB-WN synsets and modification of the mapping between Bulgarian and English synsets. The mapping between Bulgarian and English reflects the idiosyncraries between the conceptual systems of the two languages. It could be summarized that the main problems are two: (1) in some cases the English hierarchy is not quite relevant for the Bulgarian concepts (for example, the synset 'numismatics' in OEW is defined as both the collection and study of money and has a hypernym 'collection', so it is not suitable for equivalent of any of the two Bulgarian synsets (for the collection and for the study)) ; (2) for some concepts there are no synsets in OEW, so the Bulgarian synsets for the corresponding English concepts do not have equivalent synsets – sometimes these are senses which are found in English vocabularies, they are just missing in EWN (for example 'paralympian'), but of course there are also cases when the missing sense is specific for Bulgaria (such as national cuisine, locations, etc.).

The paper is structured as follows: in section 2 some related works discussing similar difficulties in the mapping of wordnets are shown; section 3 outlines some typical issues in mapping two wordnets; section 4 presents a strategy to enrich synsets with more information and section 5 concludes the paper.

2. Related Works

Some of the challenges that were faced during the building and mapping of BTB-WN have been observed in the work on other wordnets, as it could be expected prevailingly in wordnets for Slavic languages, but also for Romanian (RoWordNet [6]) and Danish (DanNet [7]). This section presents plWordNet [8], CroWN [9], RussNet [10] and also several wordnets from the BalkaNet project[4] – Czech WordNet [11], Serbian WordNet [12] and BulNet [13]. One of the most challenging aspects in the work on Slavic wordnets proves to be their productive derivational morphology.

In a paper [14] dedicated to the mapping between the RoWordNet for Romanian and PWN the authors report on problems connected to lexical gaps and mismatches. They propose solutions such as mapping Romanian synsets to an empty synset or to a synset with a Romanian-specific hyponym. With the expansion of BTB-WN Bulgarian-specific hyponym synsets are becoming more frequent. Similarly, we do not perform interlingual mapping in cases when Bulgarian synsets do not have English equivalents, but in contrast with RoWordNet we do not use links to empty nodes.

The DanNet [15] for Danish follows the `expand model`, like BTB-WN, and starts with the translation of 5000 base concepts from English into Danish, and subsequently links the vocabulary to the PWN. However, there is a difference between the approaches to the creation of base concept synsets in DanNet and BTB-WN – for DanNet a semi-automatic translation and creation of Danish concepts are performed, but in BTB-WN the base concepts are manually translated. DanNet and BTB-WN observe similar issues in the creation of interlingual relations. Both wordnets face the difficulty of deciding which is the corresponding English synset when two very similar synsets exist, which is the proper hyponym of a non-English synset, how to handle gaps in the PWN vocabulary and the lack of exact equivalents.

The plWordNet [8] for Polish followed a variant of the `merge model`, so a semantic structure was built for it and mapped to the English structure. One of the most outstanding features of plWordNet concerning relations is the rich group of adjective relations and most of all the fact that "they follow the same hierarchical, hyponymy-based model as nouns" [16]. Additionally adjectives have relations like `gradability`, `near-synonymy and modifier`, `antonymy/antonym`, `cross-categorical synonymy/pertainym`

[4] http://www.dblab.upatras.gr/balkanet/

and derivativity/derivationally related form. In BTB-WN there is a similar approach – the sem-derived-from relation, just like the Polish derivationally related from) serves to link semantically and derivationally related adjectives and nouns. In contrast with BTB-WN perfective and imperfective forms of verbs in plWordNet are divided in separate synsets, but verbs have various relations which outline such features that will be presented with lemma markers (see section 4) in BTB-WN. Likewise, diminutives in BTB-WN will labeled on the lemma level and in plWordnet – by the diminutivity relation.

A Slavic wordnet following the expand model is the Croatian WordNet (CroWN) [9], which was build by the translation of PWN, EuroWordnet and BalkaNet base concepts with special attention on preserving the specificity of Croatian. The authors suggest that typical Croatian (and Slavic) phenomena should be integrated in the structure of the wordnet and outline verb aspect and word derivation as the most prominent and problematic cases for the creation of the CroWN. It could be summarized that these two features are indeed challenging/difficult for all Slavic wordnets. In CroWN and in BTB-WN the perfective and imperfective verb pairs are members of one synset where the sense is common and modifies the same action, but in the explicit senses the synsets must contain only the correct pair member.

Several attempts for creation of Russian wordnets are known, but only the RussNet, which is 'developed from scratch', [10] is relevant for comparison with BTB-WN in respect to their way of creation. Two methods for synset formatting are used in RussNet: substitution method and semantic similarity [10]. The reason for that is that in Russian (similarly to Bulgarian) there are 'many words which are not interchangeable in a context because of the syntactic, stylistic, expressive differences, but they are considered by native speakers as having similar meanings, denoting the same objects, entities, etc.' Thus, the need to include new links in the wordnet emerges. For example, in RussNet are distinguished three types of synonymy between literals: 1) synonymy — 'relations between words which have the same root and different sets of affixes. They are not expressive and their senses differ so slightly that not every native speaker (researcher) is able to explain the distinction between them'; 2) near-synonymy – relations between: a) verb and abstract nouns, denoting processes of the same nature, b) adjectives and abstract nouns, denoting characteristics and qualities, c) adjectives and nouns, d) verbs and adjectives; 3) derivational-synonymy — relation between neutral words and their expressive derivatives. In BTB-WN on the other hand only one type of synonymy is used. The function of the near-synonymy relation (to show the derivational connections between the words) in BTB-WN is performed by the semi-derived-from and semi-derived-to relations and the function of the derivational-synonymy relation which links neutral and expressive words will be performed by lemma markers in BTB-WN. Additionally, the research discusses the use of derivational-hyponymy relation between verbs, nouns and adjectives with slightly different types of meaning. For example, for verbs are used specific attributes such as: 1) x-has-inchoative or x-has-specified-duration 'for actions restricted in time duration (inchoatives)'; 2) x-has-specified-recurrency 'for actions repeated only once or several times'; 3) x-has-specified-number 'for actions, having many objects involved'. Again, in BTB-WN verbs with such differences in the sense are united in one synset and are planned to be specified with lemma markers.

The following three wordnets are part of the BalkaNet project which brings together wordnets for different Balkan languages and Czech. In the creation of Czech WordNet there have been observed several issues, that are encountered also in the work on BTB-WN and they are related with the features of the Slavic languages, mainly their inflectional nature and very productive derivational morphology. The observation made in [11] that 'in some standard cases the straightforward English translation equivalents either cannot be easily found or have to be substituted by the various syntactic constructions often depending on context' is completely relevant to the case of BTB-WN. There are many similarities between the Czech WordNet and BTB-WN – they both use the PWN relations and include aspect verb pairs (perfective and imperfective verbs), reflexive verbs, prefixed and polyprefixal verbs, diminutives and noun gender pairs, but there are differences in the approach towards them. For instance, the aspect verb pairs in Czech WordNet are linked with special internal language relations and the prefixed verbs are in separate synsets in any case, while in BTB-WN the perfective and imperfective verbs are not at this point marked with any additional information and the prefixed verbs are treated differently depending on their sense (they are in one synset with the verb with general meaning if they differ only in the aktionsart (lexical aspect)). One more example can be made with the diminutives – in Czech WordNet their synsets have particular attributes and in BTB-WN they are labeled on the level of lemma. Both wordnets introduce derivational relations, but in Czech WordNet they are much more thorough, which is strongly motivated by the presence of grammatical case in Czech – there have been included several EuroWordNet derivational relations, additional relations following the main types of derivation processes in Czech and also valency frames are integrated in the Czech verb synsets [11], [17].

The Serbian WordNet (SWN) follows the expand model and is mapped to the PWN. It has inherited relations from the PWN whereas the BTB-WN inherits the ones from the OEW: hypernym, near_antonym, holo_part, verb_group, holo_member, be_in_state, subevent, causes, derived and particle. Additionally, SWN has a special interlingual relation – eq-synonyms, which is used between the PWN, SWN and the French wordnet when there is a one-to-one correspondence between the synsets. In SWN diminutives are also incorporated, as in BTB-WN, but there is a lack of information on how they are presented. Similarly to other Slavic wordnets, in [12] augmentatives and diminutives of nouns, possessive adjectives, verb form aspect and transitiveness, as well as the cross-PoS problem are considered problematic, but unfortunately there is no available information on the approach towards them. The 'cross-POS problem' is also faced in BTB-WN. This is the case when a noun in English corresponds to a different part of speech in Serbian (for example the English noun 'sort' corresponds to a pronoun in Serbian – 'nekakav', and also the English noun 'peer' corresponds to an adjective in Serbian – 'ravan'). Another task faced in SWN, that is still not tackled in BTB-WN is finding and marking the most prominent sense for a lemma. For the purpose were calculated occurrences of keywords in subcorpora and the most prominent sense of a lemma was determined.

One more wordnet that can be compared with the BTB-WordNet is the Bulgarian WordNet (BulNet). In BulNet the relations has-subevent and is-sub-event-of are introduced [18], [19]. In it verbs are connected in the matter of temporal inclusion (co-extensiveness or proper inclusion) and temporal exclusion (backward presupposition or causation). In comparison, for verbs BTB-WN uses only the relations originating from PWN such as entails and is-entailed-of. For example, in the BTB-WN хъркам *harkam* 'snore' is related with entails to спя *spya* 'sleep' while in the BulNet хъркам *harkam* 'snore' is connected with is-subevent-of to спя. Moreover, in the BulNet there is a derivational relation for participles – has-participle. In BTB-WN there is no similar relation, because a decision was made to include in synsets only participles which are used as adjectives and the rest are included in the morphological paradigm of the verb. In contrast to BTB-WordNet, in BulNet the perfective and imperfective verbs are divided in separate synsets in order to distinguish them. The nouns for females and males denoting professions, roles, nationalities and animals in BulNet also are divided but they are connected with relation. In addition, the diminutives in BulNet are in a separate synset from the source nouns to which they are hyponyms, whereas in the BTB-WN they are united in one synset.

3. Mapping Between the Relation Structures

As it is presented in [20] (and also in many works on wordnets for other languages) both the merge and expand model have some disadvantages, so better results are achieved with the combination of the models – "if the expand method is chosen, the language resource suffers from lack of nativeness of the hierarchy and relations. If the merge method is followed, the language resource differs too much from other similar resources and it is time-consuming to map it back to them".

The initial approach adopted in BTB-WN included the following interlingual relations: full correspondence (one-to-one); partial correspondence (one-to-many or many-to-one); forced connectivity (re-design of Bulgarian definition); common general meaning; resolving metonymies; incorrect and extended correspondence [20]. We have modified this approach and applied new relations. A new interlingual relation is the near-equivalent-to – it is used when there is no exactly equivalent English synset (for example there is a near-equivalent-to relation between the Bulgarian synset for книжарница *knizharnitsa* 'bookshop' and the English synset for 'bookshop', because the meaning of книжарница is a shop in which books, stationary and writing materials are sold and the OEW definition includes only books). It is also used together with the equivalent-to relation to link one Bulgarian synset to several English synsets. For Bulgarian synsets without English equivalents we are trying to find at least one relation which led us to the introduction of the semantically-derived-from and semantically-derives-to relations, which link adjectives and nouns. For example, the adjective библиотечен *bibliotechen* meaning 'related to library' has a semantically-derived-from relation towards the noun библиотека *biblioteka* 'library' and the noun has a semantically-derives-to relation towards the adjective. As future work we plan to add similar derivational relations between other parts of speech as well. Nevertheless, we have currently prioritized specifically those between adjectives and nouns due to the large number of such cases.

3.1. Challenges in Transferring OEW Hierarchy to BTB-WN

The first challenge is POS-related – it refers to the structural relations transferred from the OEW to BTB-WN. A good example are the adjectives. The OEW differentiates two kinds of adjectives: adjectives (a) and satellites (s) that are organised in clusters around a head adjective. The satellite synset is defined as 'synset in an adjective cluster representing a concept that is similar in meaning to the

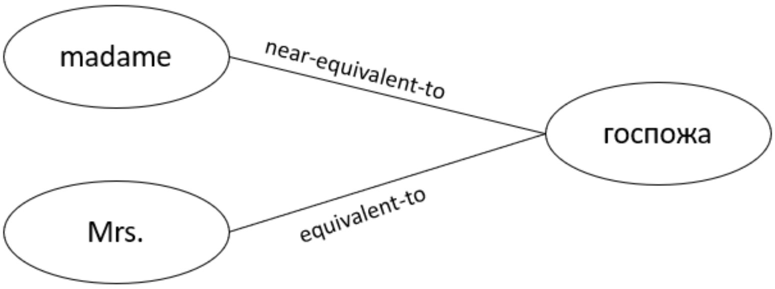


Figure 1: Figure 1: Example of a many-to-one interlingual relation

concept represented by its head synset'. [5] For the time being we do not find it necessary to distinguish adjectives and satellites. That is until we find a precise criterion by which to determine them. So BTB-WN contains only four parts of speech at this moment: nouns, verbs, adjectives and adverbs.

The second challenge is concept-related – it regards the granularity of the OEW, where the lexicographers encode very subtle differences in meaning which results in the presence of many synsets with similar meanings. We prefer the opposite approach – consolidation of the hierarchy, and therefore – unification of very similar meanings, supported by detailed definitions and diversified examples. If we completely follow the structure of OEW this would make BTB-WN an English language-dependent wordnet and it would be very difficult to map it with wordnets in other languages. So our strategy is to unite two or sometimes more highly similar English synsets and map them with the one corresponding Bulgarian synset using the `near-equivalent-to` relation. For example the synset for the adverb 'repeatedly' has no structural relation with the synset 'over and over' which is defined as 'repeatedly'; the Bulgarian verb впрягам *vpryagam* in the sense of 'getting a pack animal ready to pull a cart by putting a harness on it' can be mapped to three synsets in OEW – 1) 'inspan' 'attach a yoke or harness to', 2) 'harness' 'put a harness' and 3) 'yoke' 'put a yoke on or join with a yoke'; and the noun госпожа *gospozha* 'a form of address for a (presumably married) woman' has relations with two synsets in OEW – 'madame' 'title used for a married Frenchwoman' and 'Mrs.' 'a form of address for a married woman' (see Figure 1) because in several Bulgarian explanatory dictionaries these two words are presented as synonyms and we prefer to follow this approach in BTB-WN.

The issue with the granularity can be observed also in the opposite direction. In some cases the Bulgarian language differentiates more concepts than English. An example can be made with the lemmas леля, тетка *lelya, tetka* 'my mother's sister', леля *lelya* 'my father's sister', стрина, стринка, чинка *strina, strinka, chinka* 'the wife of my father's brother' and вуйна, уйна, учинайка *vuina, uina, uchinaika* 'the wife of my mother's brother' which can all be linked to one English synset with `near-equivalent` relations – 'aunt' 'the sister of your father or mother; the wife of your uncle'.

In some hierarchies a change is needed because the structure creates logical inconsistencies, for example the noun маниак *maniak* 'maniac' defined as 'a person who has an obsession with or excessive enthusiasm for something'. As shown in Figure 2 it has hypernyms 'enthusiast' ('a person having a strong liking for something'), then 'admirer' ('someone who admires a young woman') and later 'lover' ('a person who loves someone or is loved by someone'). The hypernym 'lover' narrows down the meaning of its hyponyms by defining them as someone who loves or is loved by someone else, but 'maniac' and 'enthusiast' are not used only for love for a person – there are usages for people who love arts, sports, etc.

Another issue met during the work on BTB-WN concerns the participles. We include in synsets only the participles that are used as adjectives in Bulgarian, e.g. решен *reshen* 'resolved', 'solved' meaning 'that has got a solution or explanation' which is a lexicalised partici-

[5] `https://wordnet.princeton.edu/documentation/wngloss7wn`

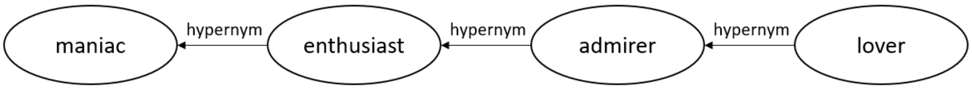


Figure 2: Figure 2: Example of a problematic hierarchy chain

ple commonly used as an adjective (решен проблем *reshen problem* 'solved problem'). OEW is rather rich with participles, but a decision was made not to create synsets for those of them that are not used as adjectives in Bulgarian, because such influence from English is not beneficial for BTB-WN. For example the participles харесван *haresvan* 'liked' meaning 'that is found pleasant, attractive', боядисан *boyadisan* 'painted' meaning 'that is covered with paint' and строшен *stroshen* счупен, *schupen* 'broken' meaning 'that is made into pieces by physical force' have synsets in OEW, but can not be found in the Bulgarian dictionaries.

An interesting example (shown in Figure 3) for differences in the hierarchy can be made with the synset for 'sibling' ('a person's brother or sister'). Such term does not exist in Bulgarian and we were considering either dropping it of the hierarchy or creating an artificial Bulgarian synset that combines both 'brother' and 'sister'. However, it turned out that neither 'brother', nor 'sister' are included as hyponyms of 'sibling', but 'twin', 'triplet', etc. are. The hypernym of 'brother' is 'male sibling' that has a hypernym 'kinsman' which is a hyponym of 'relative'. Analogous to this is the case with 'sister' – the hypernyms in its hierarchy are 'female sibling', 'kinswoman', 'relative'. Thus, 'sibling' and 'male sibling'/'female sibling' are not structurally related to one another although they are semantically connected. Furthermore, in Bulgarian there are no accurate equivalents for 'kinsman','kinswoman' and 'male sibling', 'female sibling' (the last two of which seem rather artificial, maybe they only have a structural role for the purpose of organizing other concepts). After all, the question is not just whether it is beneficial to create an artificial synset for 'sibling', but how in general to connect брат *brat* 'brother' and сестра *sestra* 'sister' to the OEW. Furthermore, we do not find it appropriate that 'brother' and 'sister' are not structurally related to 'twin', 'triplet', etc. since the concepts of those words are inevitably connected – a 'twin' is either a brother or a sister.

3.2. Missing Concepts in OEW

The gaps that we observe in OEW are either differences between the English and Bulgarian or are a matter of approach. An example for the differences between the languages could be made with the noun военна служба *voenna sluzhba* 'military service' which cannot be linked directly to military service in OEW. In Bulgarian the concept denotes the exercise of the citizen's duty to serve in the armed forces for set amount of time and the English concept revolves around the group of people serving in the armed forces. Another possible translation in English is the term 'conscription', but its synset in OEW is categorised as an act, its hypernym 'mobilisation, militarisation' and hyponym 'levy en masse, levy' are also acts, so this synset could not be considered equivalent with the Bulgarian synset.

We observed that the OEW lacks figurative meanings which is an example of a gap that depends on the matter of approach. However, figurative meanings are included in BTB-WN and we have to find an appropriate position and relations in the hierarchy for such terms since they do not have appropriate equivalents. For example, самоубиец 'suicide, a person who kills himself intentionally' is present in OEW, but its figurative meaning 'a person whose lifestyle and behaviour have destructive effect on their carrier, health, etc.' is missing. The synset for the literal meaning of прозорец *prozorets* 'window' has equivalent in OEW but its figurative meaning – 'something that provides wide opportunities to acquire knowledge, career progress, etc.' – does not.

In order to connect such cases to the semantic structure we are using the hypernym-hyponym relation and looking for an appropriate position for them in the hierarchy – for example they could be placed under their suitable hypernym. This task elaborates additionally with the artificial synsets in OEW.

OEW does not cover all parts of speech – it excludes pronouns, prepositions, conjunctions, particles and interjections. This is another obstacle for the mapping because if we include these parts of speech in BTB-WN they will not have OEW equivalents, so they will be provided with additional information by lemma markers. It would be

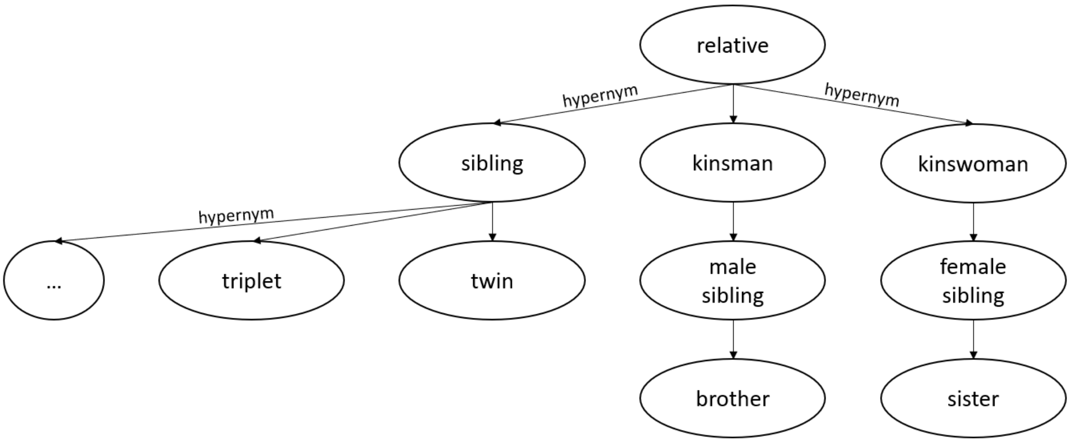


Figure 3: Figure 3: Example of concept disjointedness

beneficial to incorporate these currently missing parts of speech in BTB-WN, because the good POS coverage helps for better performance in different NLP tasks. For example, the interjection ох oh ‘used to express pain, sadness, relief, etc.’ has developed a meaning of a noun ‘synonym of exclamation, groaning, moaning’ used in a phrases such as няма ох *nyama oh* – ‘there is no oh’. Furthermore, there are some differences between the categorisation of some of the parts of speech in English and in Bulgarian. For example: не *ne* ‘used to express negation in different parts of the sentence’ in Bulgarian is considered to be a particle. However, the appropriate equivalent in OEW is not, defined as ‘negation of a word or group of words’, and it is marked as an adverb. Such examples show the need of new relations and as part of our future work we will consider the application of cross-POS interlingual relations, so we can incorporate different parts of speech.

Another challenge arises when the OEW does not include a certain concept although it is found in English dictionaries. For example the Bulgarian synset for виртуален *virtualen* ‘virtual’ in the sense of ‘something, that is created electronically as a simulation, an analogue of certain physical object and real phenomena and with which a person can come into contact, interaction’ does not have an exact English equivalent because *virtual* appears in OEW in two synsets with the following definitions: 1) ‘existing in essence or effect though not in actual fact’; 2) ‘being actually such in almost every respect’. The proof that this is not a case of conceptual difference between the two languages is that this meaning of ‘virtual’ is found in the Oxford Learner’s Dictionary [6] – ‘(computing) made to appear to exist by the use of computer software, for example on the internet’.

Also, a solution has to be found for the mapping of Bulgarian reflexive verbs. Their English analogous meaning is usually expressed by the general form of the verb. This way one English synset covers several Bulgarian meanings. For example – сресвам се *sresvam se* ‘to comb myself’ does not have an exact equivalent – the English synset for ‘comb’, ‘comb out’ and ‘disentangle’ ‘smoothen and neaten with or as with a comb’ has equivalent relation with the Bulgarian synset for the transitive verb сресвам *sresvam*, среша *sresha*, разресвам *razresvam*, разреша *razresha*, etc. ‘arrange, smooth hair, mane, etc. with a brush or comb’.

One inconsistency in OEW presents a serious issue for our work. It concerns the nouns for men and women denoting professions, roles, nationalities and animals – these noun pairs sometimes are combined in one synset, in other cases they are divided. The reason for the inconsistency is usually, but not always, the lack of nouns for women in English such as *teacheress as counterpart of teacher. Our decision is to always unite these nouns in BTB-WN. Occupations like ‘actor’ and ‘actress’, ‘waiter’ and ‘waitress’ are divided in separate synsets in OEW, but many similar are combined in one synset. When the nouns for men and women are in separate synsets in OEW this leads to problems with the structure of the relations. In some cases the lemma for nouns for women is a hyponym of the lemma for men which is not the most precise relation, and when we put both in one synset it is very difficult to find the appropriate position in the hierarchy. In other cases the nouns for men and women for professions and roles

[6] https://www.oxfordlearnersdictionaries.com/definition/english/virtual?q=virtual

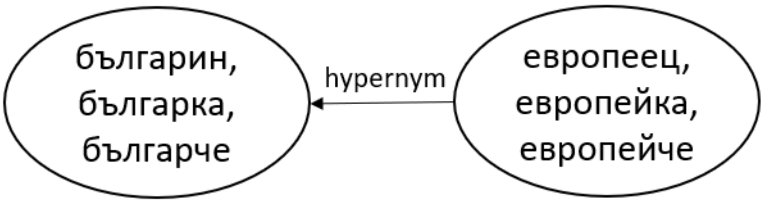


Figure 4: Figure 4: Example of Bulgarian synsets for ethnicities

are not structurally related. For example, 'widower' has a hypernym 'man' and 'widow' has hypernym 'woman' but between the synsets of 'widower' and 'widow' there is no formal relation although the role of the two concepts is basically the same.

One more example of this issue is related to the nationality terms – in English there are no nouns for denoting women and children, but in Bulgarian there are and our idea is to unite them in one synset – българин, българка, българче[7], *balgarin, balgarka, balgarche* 'Bulgarian man, Bulgarian woman, Bulgarian child' as shown in Figure 4.

Another similar case is related to the synsets for female and male animals in OEW – often the domestic animals are divided (like 'hen' and 'cock', 'male horse' and 'mare', 'billy' and 'nanny', 'cow' and 'bull'), but for wild animals the synsets are inconsistent. For example, there is only one synset for wolfs, the available in dictionaries female form 'she-wolf' is not present, but for 'fox' there are two synset – one general and one for the female animal – 'vixen'. This means that the corresponding Bulgarian synsets again would not be the same, they will have different type of relations.

4. Additional Linguistic Information

We plan to incorporate lemma markers with additional linguistic information in the BTB-WN synsets. We believe, that this will help in two directions – 1) to avoid having too many and very similar synsets; and 2) to distinguish the semantic nuances of lemmas united in one synset (for example the lemmas for the general and the diminutive form or for formal and slang words). One part of this linguistic information will be in the form of POS tags derived from a morphological dictionary [21] that is attached to BTB-WN.

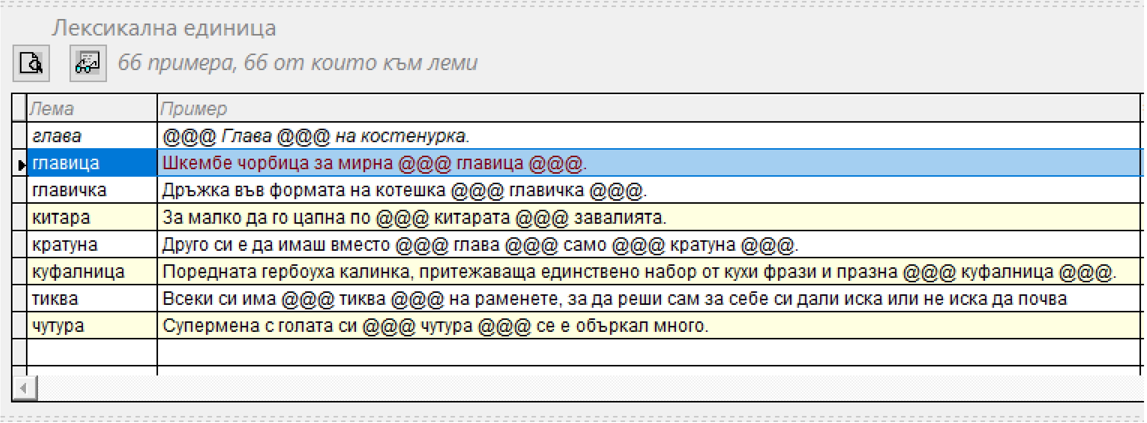
The approach for combining lemmas with slightly different semantic nuances is borrowed from explanatory dictionaries – [22] and [23], and bringing together the forms for men and women for professions, roles, functions, etc. in one synset follows [22].

These additional lemma markers will make the BTB-WN more thorough and consequently it is going to be much more suitable for semantic parsing of all kinds of texts – from dialectal and old texts, to children books and colloquial language.

We plan to introduce markers with additional linguistic information for:

1. Bulgarian verbs with prefixes that bear semantics of start, end, duration, repeatability, etc. of the action. Synsets with these verbs can not have an equal English synset so we decided to map them with the synset for the general meaning of the verb and label the different forms with their specific semantic features on the level of lemma. For example the verbs чета *cheta* 'read' and зачитам *zachitam* зачета *zacheta* 'start reading' will be united in one synset; the second verb will have a lemma marker that further describes its meaning.
2. The diminutive forms of the nouns will be in one synset with the general form, but will have a special lemma marker. For example: стол *stol* 'chair' and столче *stolche* 'chair-diminutive'. Another useful aspect of this approach is that this way there will not be several synsets with only one different feature – diminutiveness, because in Bulgarian diminutives can have more than one meaning. The general one is that something is very small or very young, but they can also express gentle or diminishing, humiliating attitude. An example for that is the synset европеец *evropeets* 'European', европейка *evropeika* 'European woman' and европейче *evropeiche* 'European-diminutive' where the diminutive can refer to a European child

[7] българче is also the form for the diminutive of *Bulgarian*, but the example is not relevant to it



Лексикална единица

66 примера, 66 от които към леми

Лема	*Пример*
глава	*@@@ Глава @@@ на костенурка.*
главица	Шкембе чорбица за мирна @@@ главица @@@.
главичка	Дръжка във формата на котешка @@@ главичка @@@.
китара	За малко да го цапна по @@@ китарата @@@ завалията.
кратуна	Друго си е да имаш вместо @@@ глава @@@ само @@@ кратуна @@@.
куфалница	Поредната гербоуха калинка, притежаваща единствено набор от кухи фрази и празна @@@ куфалница @@@.
тиква	Всеки си има @@@ тиква @@@ на раменете, за да реши сам за себе си дали иска или не иска да почва
чутура	Супермена с голата си @@@ чутура @@@ се е объркал много.

Figure 5: Figure 5: The synset for 'head' with colloquial and offensive (кратуна *kratuna*, куфалница *kufalnitsa*, тиква *tikva*, чутура *chutura*), diminutive (главица *glavitsa*, главичка *glavichka*) and slang lemmas (китара *kitara*) from the **CLaDA-BG-Dict**.

or to a European with some ironical attitude. One more example is the synset адвокат *advokat* 'lawyer', адвокатка *advokatka* 'female lawyer' and адвокатче *advokatche* 'lawyer-diminutive' where the diminutive can refer to four different senses according to the context: 1) a young lawyer; 2) affectionate attitude; 3) belittling attitude; 4) negative attitude.

3. Another type of markers will be used to label archaic, dialectal, slang, informal, vulgar, offensive, etc. terms. For example:
    - archaic terms – храна *hrana* 'food', ядене *yadene* 'food' and пища *pishta* 'food-archaic'.
    - dialectal term – магданоз *magdanoz* 'parsley', мерудия *merudiya* 'parsley-dialectal' and меродия *merodiya* 'parsley-dialectal'.
    - slang term – работя *rabotya* 'work', трудя се *trudya se* 'work' and бачкам *bachkam* 'work-slang'.
    - vulgar term – глупак *glupak* 'stupid person', идиот *idiot* 'idiot', кретен *kreten* 'cretin' and тиквеник *tikvenik* 'simpleton'
4. Marker for multi-word expression lemmas. MWEs are defined as 'combinations of two or more words that are typically used to express a specific concept. (...) these combinations are stored in the mental lexicon of native speakers and as a whole refer to a (linguistic) concept' [24]. For example: ръчно *rachno* 'by hand' and на ръка *na raka* 'by hand' have the exact same meaning, but are different in structure, so the approach towards such cases in BTB-WN is to unite the MWE with its synonyms in one synset.
5. Derivational relations between different parts of speech. These markers are planned to be used for lemmas rather than synsets because if they are applied on synsets they would be appropriate for members of one synset that are not derivationally related. This lemma marker will be used a lot especially for the very well-known example of the conversion of English nouns to adjectives. Because of this it is common that the synsets for Bulgarian adjectives do not have appropriate equivalents. Therefore, by the usage of derivational lemma markers we will be able to cope with this issue. An example could be made with the nouns диамант *diamant* and елмаз *elmaz* (both meaning 'diamond'). The related adjective диамантен *diamanten* is going to be linked to диамант ('diamond'), but not to елмаз and analogously – елмазен *elmazen* is going to be related to елмаз, but not to диамант.

5. Conclusion

We could summarize that the recent work on the BTB-WN for Bulgarian includes consolidation of meanings in synsets, verification of the inherited from the Open English WordNet structure and relations, modification and editing where necessary as well as addition of new interlingual relations and relations between BTB-WN synsets.

Overall, the challenges we encountered in the mapping the BTB-WN and OEW are based on missing concepts in one of the wordnets, distinctions between the two languages and differences in the way that the resources are built. The discrepancies are solved with interlingual relations and relations between lemmas and synsets in the

BulTreeBank WordNet.

The plans for future work concern the addition of more parts of speech in BTB-WN, introduction of cross-POS relations and reconsideration of the POS categories – the OEW distinguishes 26 categories for nouns, three for adjectives, one for adverbs and 15 categories for verbs, which are currently used in BTB-WN.

Acknowledgments

This work was supported by *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant number DO01-377/18.12.2020. and partially supported by the *Bulgarian Ministry of Education and Science* under *the National Research Programme 'Young scientists and postdoctoral students'* approved by DCM # 577 / 17.08.2018.

In addition the work of Zara Kancheva has been accomplished with the partial support by the Grant No. BG05M2OP001-1.001-0003, financed by *the Science and Education for Smart Growth Operational Program (2014-2020)* and co-financed by the European Union through *the European structural and Investment funds*.

References

[1] P. Osenova, K. Simov, The Data-driven Bulgarian WordNet: BTBWN, Cognitive Studies | Études cognitives 18 (2018).

[2] C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press, 1998.

[3] J. P. McCrae, E. Rudnicka, F. Bond, English WordNet: A new open-source WordNet for English, K Lexical News (2020) 37–44. URL: https://doi.org/10.5281/zenodo.4382320. doi:10.5281/zenodo.4382320.

[4] J. Ellman, Eurowordnet: A multilingual database with lexical semantic networks: Edited by piek vossen. kluwer academic publishers. 1998. isbn 0792352955, Natural Language Engineering 9 (2003) 427 – 430. doi:10.1017/S1351324903223299.

[5] K. Simov, A. Simov, H. Ganev, K. Ivanova, I. Grigorov, The CLaRK System: XML-based Corpora Development System for Rapid Prototyping, Proceedings of LREC 2004 (2004) 235–238. URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/258.pdf.

[6] D. Tufiş, E. Barbu, V. Mititelu, R. Ion, L. Bozianu, The Romanian Wordnet, Romanian Journal on Information Science and Technology. Special Issue on BalkaNet 7(2-3) (2004) 107–124.

[7] B. S. Pedersen, S. Nimb, J. Asmussen, N. H. Sørensen, L. Trap-Jensen, H. Lorentzen, DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary, Language Resources and Evaluation 43 (2009) 269–299.

[8] M. Derwojedowa, M. Piasecki, S. Szpakowicz, M. Zawisławska, Polish Wordnet on a shoestring, Proceedings of Biannual Conference of the Society for Computational Linguistics and Language Technology, 2007 (2007) 169–178.

[9] I. Raffaelli, M. Tadic, B. Bekavac, Ž. Agic, Building croatian wordnet, in: Proceedings of GWC, 2008, pp. 349–360.

[10] I. Azarova, O. Mitrofanova, A. Sinopalnikova, M. Yavorskaya, I. Oparin, Russnet: Building a Lexical Database for the Russian Language, Proceedings of Workshop on WordNet Structures and Standardisation and How this affect Wordnet Applications and Evaluation (2002) 60–64.

[11] K. Pala, P. Smřz, Building Czech WordNet, Romanian Journal of Information Science and Technology 7 (2004) 79–88.

[12] C. Krstev, G. Pavlovic-Lazetic, D. Vitas, I. Obradović, Using textual and lexical resources in developing Serbian Wordnet, Romanian Journal on Information Science and Technology 7 (2004) 147–161.

[13] G. T. Koeva, S., A. Genov., Towards Bulgarian Wordnet, Romanian Journal on Information Science and Technology 7(1-2) (2004) 45–60.

[14] D. Cristea, C. Mihăilă, C. Forascu, D. Trandabat, M. Husarciuc, G. Haja, O. Postolache, Mapping Princeton WordNet synsets onto Romanian Wordnet synsets, Romanian Journal of Information Science and Technology (ROMJIST) 7 (2004) 125–145.

[15] B. S. Pedersen, S. Nimb, I. R. Olsen, S. Olsen, Merging DanNet with Princeton Wordnet, in: Proceedings of the 10th Global Wordnet Conference, Global Wordnet Association, Wroclaw, Poland, 2019, pp. 125–134. URL: https://aclanthology.org/2019.gwc-1.16.

[16] E. Rudnicka, W. Witkowski, M. Piasecki, A (non)-perfect match: Mapping plWordNet onto PrincetonWordNet, in: Proceedings of the 11th Global Wordnet Conference, Global Wordnet Association, University of South Africa (UNISA), 2021, pp. 137–146. URL: https://aclanthology.org/2021.gwc-1.16.

[17] K. Pala, D. Hlaváčková, Derivational relations in Czech WordNet, in: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 75–81. URL: https://aclanthology.org/W07-1710.

[18] S. Koeva, Derivational and morphosemantic rela-

tions in Bulgarian Wordnet, Intelligent Information Systems 16 (2008) 359–369.

[19] S. M. Svetla Koeva, T. Tinchev, Bulgarian Wordnet - structure and validation, Romanian journal of information science and technology 7 (2004) 61–78.

[20] P. Osenova, K. Simov, Challenges Behind the Data-driven Bulgarian WordNet (BulTreebank Bulgarian Wordnet), in: Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets, co-located with 1st Conf. LDK 2017, Galway, Ireland, 2017, pp. 152–163.

[21] D. Popov, K. Simov, S. Vidinska, A Dictionary of Writing, Pronunciation and Punctuation of Bulgarian Language, Atlantis LK, Sofia, Bulgaria, 1998.

[22] D. Popov, Bulgarian Explanatory Dictionary. Forth revised and extended edition, Nauka i izkustvo, Sofia, Bulgaria, 1994.

[23] E. Pernishka, E., Dictionary of the Bulgarian language, Academic publishing company "Prof. Marin Drinov", 2001.

[24] S. Sprenger, Fixed expressions and the production of idioms, Ponsen and Looijen BV, Wageningen, 2003.

Implementation of Humanities and Social Sciences Data Storage, Retrieval and Curation Environment for the National Library "Ivan Vazov" – Plovdiv needs

Desislava Paneva-Marinova[1], Maxim Goynov[2], Lubomir Zlatkov[3], Detelin Luchev[4], Radoslav Pavlov[5], Lilia Pavlova[6]

[1, 2, 3, 4, 5] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, bl.8, Acad. G. Bonchev Str., 1113 Sofia, Bulgaria
[6] Laboratory of Telematics, Bulgarian Academy of Sciences, bl.8, Acad. G. Bonchev Str., 1113 Sofia, Bulgaria

Abstract

This paper presents the design process of implementation of the humanities and social sciences data storage, retrieval and curation environment (DSRCE) for the needs of the National Library "Ivan Vazov" – Plovdiv, Bulgaria, one of the biggest content providers in CLaDA-BG (the Bulgarian National Interdisciplinary Research E-Infrastructure for Bulgarian Language and Cultural Heritage Resources and Technologies). The implementation of an environment for storage, extraction and curation of humanities and social sciences in the Plovdiv library will support the users' access to the extensive funds of the library. This will give researchers more opportunities to investigate, manage and publish their research data.

Keywords

Research e-Infrastructure, Digital Humanities, Digital Content Management Systems.

1. Introduction

The potential of information technology to support the work of researchers in the humanities and social sciences by offering solutions for the intelligent digital management and presentation of national cultural heritage objects and knowledge increases every year. CLaDA-BG (the Bulgarian National Interdisciplinary Research E-Infrastructure for Bulgarian Language and Cultural Heritage Resources and Technologies) [1], in the spirit of European CLARIN and DARIAH), is a leading initiative in Bulgaria, aiming to promote the innovative use of resources in order to encourage the sustainable development of European cultural landscapes in a digital environment. CLaDA-BG's technological partners are creating technologies and tools to unify the processes of access, preservation and use of Bulgarian language and cultural historical content in accordance with the established best practices and regulations in the field.

This paper presents the design process of implementation of the humanities and social sciences data storage, retrieval and curation environment (CHCS-DSRCE) for the needs of the National Library "Ivan Vazov" – Plovdiv, one of the biggest content providers in CLaDA-BG. The prototype is developed by the IMI-BAS team and aims to provide flexible and efficient access to multimedia representations of cultural and historical artifacts, supporting a variety of forms and formats of digital information content and rich functionality for interaction. Emphasis is placed on the storage, retrieval and curation of data and metadata for target objects.

Plovdiv's library is the second largest library in Bulgaria and serves as the second national repository for Bulgarian literature. The library

CLaDA-BG 2021 Conference: Language Technologies and Digital Humanities in Bulgaria 2021, September 06--07, 2021, Varna, Bulgaria
EMAIL: dessi@cc.bas.bg (D. Paneva-Marinova); goynov@gmail.com (M. Goynov); lyubcho@gmail.com (L. Zlatkov); dml@math.bas.bg (D. Luchev); radko@cc.bas.bg (R. Pavlov); pavlova.lilia@gmail.com (L. Pavlova)
ORCID: 0000-0001-5998-687X (D. Paneva-Marinova); 0000-0003-0926-5796 (D. Luchev)

owns a rich, multipurpose fund of over 1,460,000 library units – scientific literature and fiction; manuscripts, incunabular, rare and valuable publications; a rich reference fund; Bulgarian and foreign periodicals; audiovisual and digital documents; original works of art. The library offers digital catalogues of the library units, published after 1996, including a digital collection of Slavic manuscripts and incunabular, rare and valuable publications. The implementation of the humanities and social sciences data storage, retrieval and curation environment in Plovdiv library will support the users' access to the extensive funds of the library. It will give the researchers more possibilities in the investigation, management, and publication of their research data. The implementation will also support the researchers' collaborative work, knowledge sharing and participation in social and cultural life.

2. Core environment and used technologies

The Humanities and Social Sciences Data Storage, Retrieval and Curation Environment is a web-based software environment, supporting a variety of digital cultural units and rich functionality for interaction, with an accent on components providing storage, retrieval and management of data and metadata. The implementation is based on the experience and knowledge gained from previous developments of the IMI team on digital content management systems (including digital libraries, digital repositories, galleries, *etc.*) preserving the valuable Bulgarian cultural heritage: Bulgarian iconographic art, Bulgarian ethnographic and folklore artifacts, medieval and early modern Bulgarian hagiographical texts, in combination with ethnological data and visual sources, *etc.* [2, 3, 4, 5, 7].

The platform is a web-based software environment that provides the following basic functional components: *a metadata management* and *presentation functional module (incl. specific services), a metadata model management module, administrative services* that are linked to a media repository and a user *data repository* [6]. The basic prototype of CHCS-DSRCE stores and manages the digital analogues of cultural heritage objects presented in text (*via* .pdf files, fully corresponding to book media), graphic, video, audio formats, or other media objects as well as the relevant metadata. The resources are digitized and made available by the CLADA-BG partners.

The *module for management of the metadata model* includes a service for building the descriptive schemas (descriptor structures) for cultural objects, the so-called *model builder*. The service manages the "object metadata descriptive structure", supporting flexible opportunities to build, edit and extend its meta descriptors (also called items). Every item has an inherent predefined data type specification, determined by the object's data. While standard data types (*e.g.*, Text, Textarea, Text (multilingual), Textarea (multilingual), Number, Date, Time, Object, Array, Dropdown, Single choice, Multiple choice, File or Hidden) are available, the environment also allows for customized data types, such as domain specific predefined nonscalar object types, thus allowing definition of nested structures. The current implementation utilizes the later feature and, as a result, eliminates the need for redefinition of common descriptive items for each specific object type.

Similar descriptive items available for several types of objects could be aggregated in a separate general scheme, which unites the common elements of the objects, with the specifics of the different types of objects added as additional characteristics of their respective descriptive schemes. For example, the creation of a descriptive model of an object of type "Book", could potentially trigger data sharing with the common characteristics of other objects (*i.e.*, "Periodicals", "Photography", *etc.*). These common characteristics for different objects can be distinguished in an "Identification" model of *an abstract type of object*. Subsequently, the object "Book" will have all the descriptive items of the model of the abstract object type "Identification" (such as "Signature", "Inventory №", "Name", *etc.*), which are common to other types of objects, plus all specific descriptive items that are specific for the object "Book".

This approach is also fundamental for the efficient use of the platform's search engine, enabling, simplifying and boosting the performance of a unified search mechanism, allowing the users to perform complex search queries within the whole content, using simple search operations.

The model builder could also create and manage relationships between objects showing their complex or heterogeneous descriptive structure. The model builder could also support preliminary defined descriptive schemas and standards in the cultural heritage field, including Dublin core, CIDOC-CRM, *etc.*

The *functional module for managing and presenting metadata* implements the basic activities related to the creation and management of metadata for cultural objects: adding, storing, editing and deleting metadata; searching, selecting (filtering), accessing, viewing and displaying metadata. For the creation of the metadata, functionalities are provided to optimize the input, including tree structure of the annotation template, reuse of metadata, suggesting already entered metadata values, auto completion, automated metadata/object import, *etc.* Specific services, as a part of this module, include functionalities closely related to the content provider needs, such as advanced collection creation, management and curation (thematic collections, time dependent collections in a calendar structure, *etc.*), search in the text media objects, advanced objects preview and ordering, different device support, *etc.* The content provider specifies the metadata categories that will be used for grouping the objects into collections in the library. Moreover, the creation, management, and use of a dictionary of specific terms, included in the description (metadata) of the stored objects are implemented in the module. The searching and tagging of a term (word form) is implemented automatically throughout the database. There are also options for placing links to pop-ups with text and media files (images), basic data analysis/synthesis, *etc.* Aggregation of objects is done based on their common (one or more) characteristics, depending on the specific model and application domain. The system summarizes groups of objects based on the aggregated data in order to improve the organization of object representation in subsequent analysis.

For each cultural object, all metadata is stored in the *media repository*. This metadata is represented in catalog records that point to the original media file(s) associated with each object. The *user profile repository* manages all user data and its changes.

The *Administrative Services panel* offers mainly user data management, metadata export, tracking services, analysis services, *etc.*

User data management covers activities related to registration (including via OAuth technology - Sign up with Google, Facebook, *etc*.), data changes, setting the access level, *etc.*

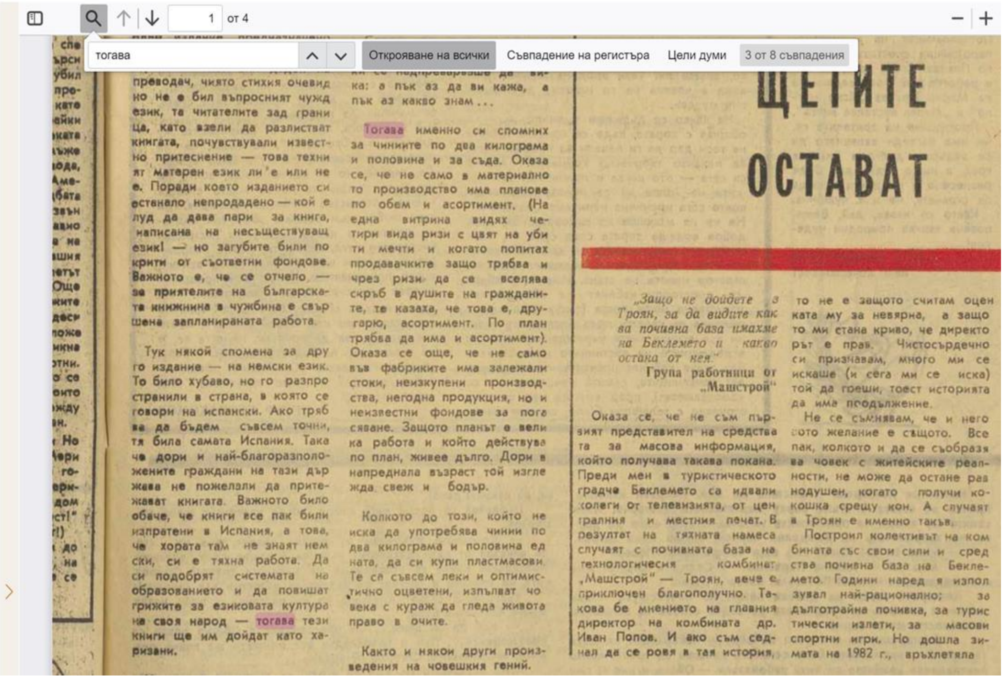
The platform is fully based on open-source software. The back-end is powered by load balanced Node JS Express 4 server behind Apache 2.4 with advanced firewall and security protection. For database management system, we use the non-relational MongoDB Community Server 4.4. This type of database is suitable for data with complex and diverse structure, and therefore is the preferred solution in this use case. The front-end is based on the MVVM design pattern and is built with the Vue JS framework which allows good management for dynamic content, complicated forms, and web pages in combination with reduced network load and good page performance. Bootstrap 4 is used in order to achieve responsiveness over different devices and to make the application more accessible (incl. for screen readers, *etc.*). Packaging and deployment are configured using WebPack 5. The application source code is managed by a Git version control system. Additional components for media management (converting, indexing, *etc.*) like FFmpeg, ImageMagick, PostScript utilities, and others are integrated.

3. Actual or anticipated outcomes in the intelligent content management context

The CHCS-DSRCE prototype produced to manage digital collections of the National Library “Ivan Vazov” – Plovdiv includes all the basic modules presented in the previous section extended with several specific functionalities. The focus is on the intelligent library units and metadata management. Plovdiv library owns rich digital catalogues of language and cultural heritage units, that need to be managed and presented in their full brilliance, providing the online readers with maximum access to the knowledge accumulated in its depositories. The big volume of library units requires well defined functionality for search, ordering, flexible access and preview, satisfying librarians’ and viewers’ needs.

The CHCS-DSRCE Plovdiv library prototype provides specific services for indexing Portable Document Format (PDF) objects, providing the opportunities for full-text search in the objects’ content (Figure 1). Functionalities for presentation of PDF objects in a web environment (without browser add-ons requirements), with options for search and visualization of the results

СТЪРШЕЛ , Брой № 2208, 03.06.1988



преводач, чиято стихия очевидно не е бил въпросният чужд език, та читателите зад граница, като взели да разлистват книгата, почувствували известно притеснение — това техният матерен език ли'е или не е. Поради което изданието си останало непродадено — кой е луд да дава пари за книга, написана на несъществуващ език! — но загубите били покрити от съответни фондове. Важното е, че се отчело — за приятелите на българската книжнина в чужбина е свършена запланираната работа.

Тук някой спомена за друго издание — на немски език. То било хубаво, но го разпространили в страна, в която се говори на испански. Ако трябва да бъдем съвсем точни, тя била самата Испания. Така че дори и най-благоразположените граждани на тази държава не пожелали да притежават книгата. Важното било обаче, че книги все пак били изпратени в Испания, а това, че хората там не знаят немски, си е тяхна работа. Да си подобрят системата на образованието и да повишат грижите за езиковата култура на своя народ — тогава тези книги ще им дойдат като харизани.

ка: а пък аз да ви кажа, а пък аз какво знам ...

Тогава именно си спомних за чиниите по два килограма и половина и за съда. Оказа се, че не само в материалното производство има планове по обем и асортимент. (На една витрина видях четири вида ризи с цвят на убити мечти и когато попитах продавачките защо трябва и чрез ризи да се вселява скръб в душите на гражданите, те казаха, че това е, другарю, асортимент. По план трябва да има и асортимент). Оказа се още, че не само във фабриките има залежали стоки, неизкупени производства, негодна продукция, но и неизвестни фондове за погасяване. Защото планът е велика работа и който действува по план, живее дълго. Дори в напреднала възраст той изглежда свеж и бодър.

Колкото до този, който не иска да употребява чинии по два килограма и половина едната, да си купи пластмасови. Те са съвсем леки и оптимистично оцветени, изпълват човека с кураж да гледа живота право в очите.

Както и някои други произведения на човешкия гений.

ЩЕТИТЕ ОСТАВАТ

„Защо не дойдете в Троян, за да видите каква почивна база имахме на Беклемето и какво остана от нея."

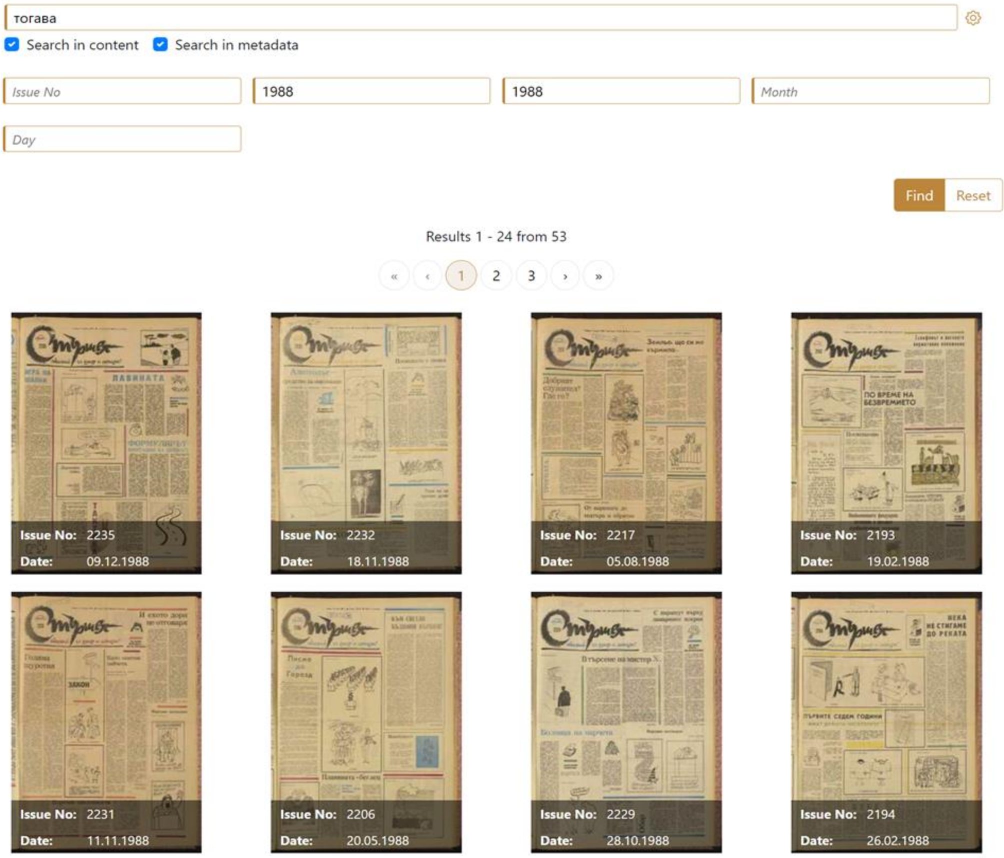
Група работници от „Машстрой"

Оказа се, че не съм първият представител на средствата за масова информация, който получава такава покана. Преди мен в туристическото градче Беклемето са идвали колеги от телевизията, от централния и местния печат. В резултат на тяхната намеса случаят с почивната база на технологичесия комбинат „Машстрой" — Троян, вече е приключен благополучно. Такова бе мнението на главния директор на комбината др. Иван Попов. И ако съм седнал да се ровя в тая история, то не е защото считам оценката му за невярна, а защото ми стана криво, че директорът е прав. Чистосърдечно си признавам, много ми се искаше (и сега ми се иска) той да греши, тоест историята да има продължение.

Не се съмнявам, че и неговото желание е същото. Все пак, колкото и да се съобразява човек с житейските реалности, не може да остане равнодушен, когато получи кокошка срещу кон. А случаят в Троян е именно такъв.

Построил колективът на комбината със свои сили и средства почивна база на Беклемето. Години наред я използувал най-рационално; за дълготрайна почивка, за туристически излети, за масови спортни игри. Но дошла зимата на 1982 г., връхлетяла

Figure 1: Full-text search in the digital objects' content



тогава

☑ Search in content ☑ Search in metadata

Issue No | 1988 | 1988 | Month

Day

Find Reset

Results 1 - 24 from 53

« ‹ 1 2 3 › »

Issue No: 2235
Date: 09.12.1988

Issue No: 2232
Date: 18.11.1988

Issue No: 2217
Date: 05.08.1988

Issue No: 2193
Date: 19.02.1988

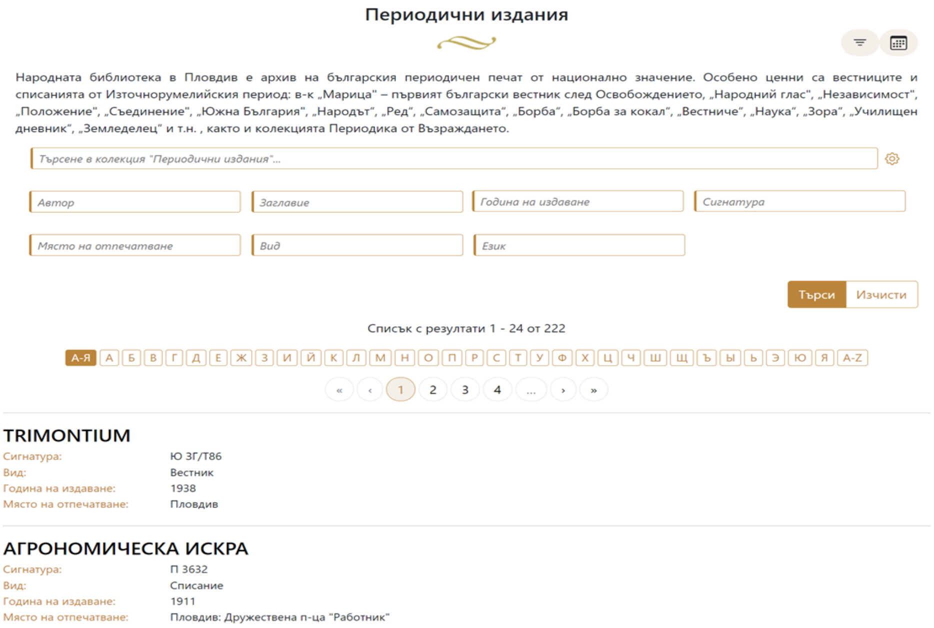
Issue No: 2231
Date: 11.11.1988

Issue No: 2206
Date: 20.05.1988

Issue No: 2229
Date: 28.10.1988

Issue No: 2194
Date: 26.02.1988

Figure 2: Content search functionality applied to all editions from 1988



Периодични издания

Народната библиотека в Пловдив е архив на българския периодичен печат от национално значение. Особено ценни са вестниците и списанията от Източнорумелийския период: в-к „Марица" – първият български вестник след Освобождението, „Народний глас", „Независимост", „Положение", „Съединение", „Южна България", „Народът", „Ред", „Самозащита", „Борба", „Борба за кокал", „Вестниче", „Наука", „Зора", „Училищен дневник", „Земеделец" и т.н. , както и колекцията Периодика от Възраждането.

Търсене в колекция "Периодични издания"...

Автор | Заглавие | Година на издаване | Сигнатура

Място на отпечатване | Вид | Език

Търси | Изчисти

Списък с резултати 1 - 24 от 222

А-Я А Б В Г Д Е Ж З И Й К Л М Н О П Р С Т У Ф Х Ц Ч Ш Щ Ъ Ы Ь Э Ю Я A-Z

« ‹ 1 2 3 4 ... › »

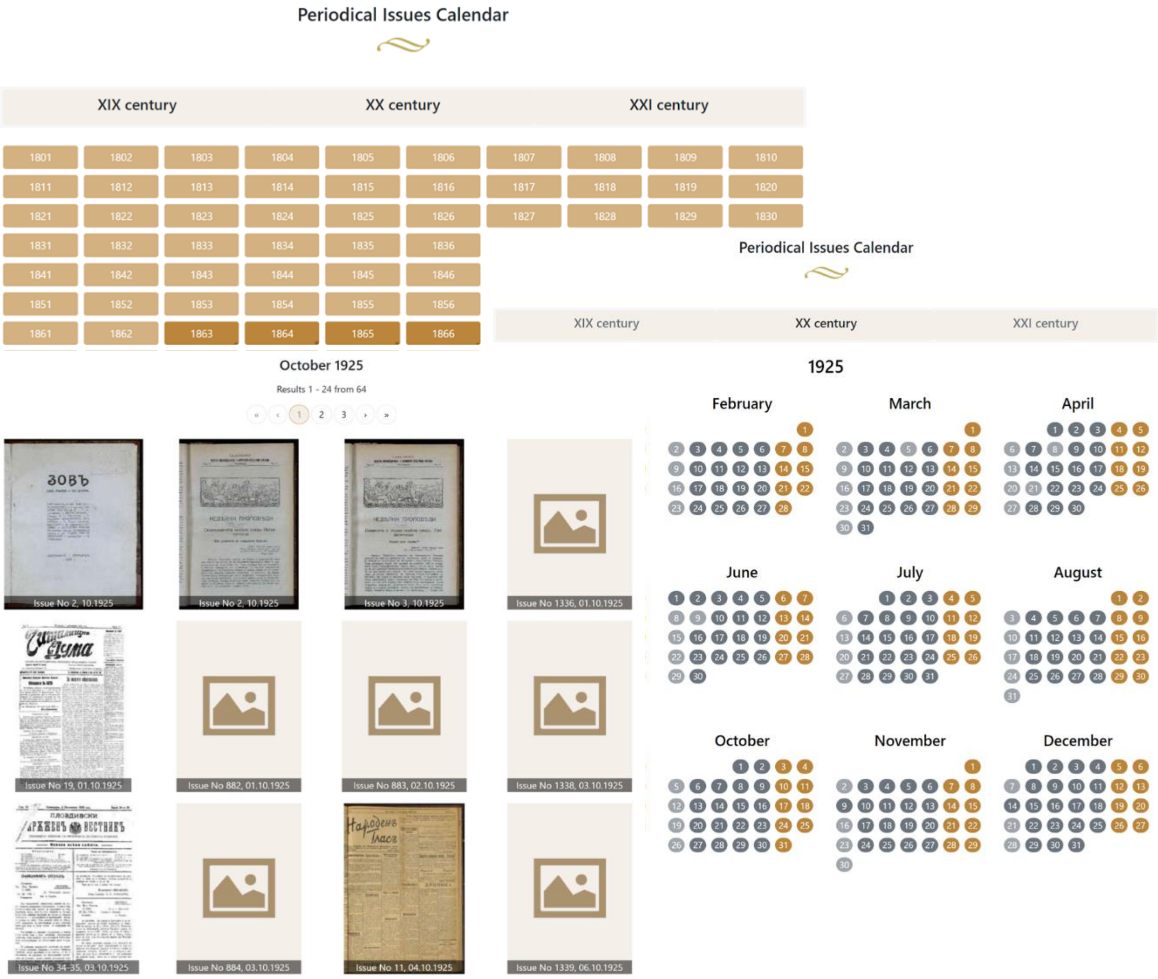
TRIMONTIUM

Сигнатура:	Ю 3Г/Т86
Вид:	Вестник
Година на издаване:	1938
Място на отпечатване:	Пловдив

АГРОНОМИЧЕСКА ИСКРА

Сигнатура:	П 3632
Вид:	Списание
Година на издаване:	1911
Място на отпечатване:	Пловдив: Дружествена п-ца "Работник"

Figure 3: Library units ordering



Periodical Issues Calendar

XIX century | XX century | XXI century

1801	1802	1803	1804	1805	1806	1807	1808	1809	1810
1811	1812	1813	1814	1815	1816	1817	1818	1819	1820
1821	1822	1823	1824	1825	1826	1827	1828	1829	1830
1831	1832	1833	1834	1835	1836				
1841	1842	1843	1844	1845	1846				
1851	1852	1853	1854	1855	1856				
1861	1862	1863	1864	1865	1866				

October 1925

Results 1 - 24 from 64

« ‹ 1 2 3 › »

Issue No 2, 10.1925 | Issue No 2, 10.1925 | Issue No 3, 10.1925 | Issue No 1336, 01.10.1925

Issue No 19, 01.10.1925 | Issue No 882, 01.10.1925 | Issue No 883, 02.10.1925 | Issue No 1338, 03.10.1925

Issue No 34-35, 03.10.1925 | Issue No 884, 03.10.1925 | Issue No 11, 04.10.1925 | Issue No 1339, 06.10.1925

Periodical Issues Calendar

XIX century | XX century | XXI century

1925

February
1
2 3 4 5 6 7 8
9 10 11 12 13 14 15
16 17 18 19 20 21 22
23 24 25 26 27 28

March
1
2 3 4 5 6 7 8
9 10 11 12 13 14 15
16 17 18 19 20 21 22
23 24 25 26 27 28 29
30 31

April
1 2 3 4 5
6 7 8 9 10 11 12
13 14 15 16 17 18 19
20 21 22 23 24 25 26
27 28 29 30

June
1 2 3 4 5 6 7
8 9 10 11 12 13 14
15 16 17 18 19 20 21
22 23 24 25 26 27 28
29 30

July
1 2 3 4 5
6 7 8 9 10 11 12
13 14 15 16 17 18 19
20 21 22 23 24 25 26
27 28 29 30 31

August
1 2
3 4 5 6 7 8 9
10 11 12 13 14 15 16
17 18 19 20 21 22 23
24 25 26 27 28 29 30
31

October
1 2 3 4
5 6 7 8 9 10 11
12 13 14 15 16 17 18
19 20 21 22 23 24 25
26 27 28 29 30 31

November
1
2 3 4 5 6 7 8
9 10 11 12 13 14 15
16 17 18 19 20 21 22
23 24 25 26 27 28 29
30

December
1 2 3 4 5 6
7 8 9 10 11 12 13
14 15 16 17 18 19 20
21 22 23 24 25 26 27
28 29 30 31

Figure 4: Calendar of the periodic editions

in a certain document, are integrated in the system.

The prototype provides advanced search applied simultaneously in the metadata and the digital objects' content (Figure 2).

The library units could be selected through ordered or alphabetic lists (Figure 3) or calendars (Figure 4).

For improvement of the context-based digital library content usage for research and e-learning in the development of the digital library, the National Library "Ivan Vazov" – Plovdiv, as a content provider, specifies the metadata categories that will be used for grouping the objects into collections in the digital library. Moreover, the creation, management, and use of a dictionary of specific terms, included in the description (metadata) of the stored objects, are also implemented. As a result, the digital library also supports an automatically implemented searching and tagging of a term (word form) throughout the database, providing advanced search functionalities, applied simultaneously in the metadata and the digital objects' content, along with a full text and regular search in objects with different structure, but sharing common characteristics, and options for placing links to pop-ups with text and media files (images), basic data analysis/synthesis, *etc.*

The objects are aggregated on the basis of their common (one or more) characteristics, depending on the specific model and application domain. The system summarizes groups of objects based on the aggregated data to improve the organization of object representation in subsequent analyses.

Following on the concept of a system, with customizable functionalities according to specific research and e-learning applications, the provided technological support for different devices (*i.e.*, PC, smart TV, tablet, smart phone) and the customized preview options enable accessibility and use of the digital library environment.

Most of the digitised materials, those of high cultural and historical value and expired copyright in the collections of the National Library "Ivan Vazov" – Plovdiv, are predominantly from the period before the Bulgarian language's last Orthographic Reform of 1945, which poses a number of problems for the accuracy of the optical character recognition (OCR) tasks, which the platform's search algorithm must take into account in order to achieve the most adequate results. Accounting for these problems within the system provides additional opportunities for better content search and context-based digital library content usage for research and e-learning, as not all researchers and (especially) students are familiar with the Bulgarian orthographic norms of the past.

4. Conclusions

The research activities of IMI-BAS within CLaDA-BG aim to consistently build scientific and information infrastructure, integrating research, education, preservation, promotion and sustainable use of national cultural heritage. It seeks to explore the potential of information technology to support the work of researchers in the humanities and social sciences by offering solutions for the intelligent digital management and presentation of national cultural heritage objects and knowledge.

A full text search improvement will be implemented using a dictionary mapping between the Bulgarian language spelling rules before and after 1945.

This will allow users to search effectively in content (before 1945) without necessary knowledge of the old Bulgarian spelling rules. An algorithm for managing hyphenations in the search indexes will be implemented as well.

The presented prototype is focused on objects having a lot of text content (periodical issues and books). The next steps for the implementation concern development of the front-end part for the other types of objects of the Plovdiv Library – manuscripts, photographs, graphic publications, maps, audio visual content, *etc.* Very important steps before launching the live version of the environment are to perform security, load, stress and performance tests, to scale properly the hardware, to define proper backup policies, which will guarantee high availability of the environment and a good protection level for the users, content and its respective metadata.

5. Acknowledgements

This work is funded in part by CLaDA-BG, the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH, Grant number DO01-377/18.12.2020.

6. References

[1] Simov, Kiril. “Integrated Language and Knowledge Resources for a Bulgarian-Centric Knowledge Graph.” Digital Presentation and Preservation of Cultural and Scientific Heritage. Vol. 9, (2019). 65-74.

[2] D. Paneva-Marinova, R. Pavlov, K. Rangochev, Digital Library for Bulgarian Traditional Culture and Folklore, in Proceedings of the 3rd International Conference dedicated on Digital Heritage (EuroMed 2010), 8-13 November 2010, Lymassol, 2010, pp. 167-172

[3] D. Paneva-Marinova, M. Goynov, D. Luchev, Multimedia Digital Library: Constructive Block in Ecosystems for Digital Cultural Assets. Basic Functionality and Services. LAP LAMBERT Academic Publishing, Berlin, Germany, 2017.

[4] Paneva-Marinova, Desislava, Jordan Stoikov, Lilia Pavlova, and Detelin Luchev. “System Architecture and Intelligent Data Curation of Virtual Museum for Ancient History.” SPIIRAS Proceedings, 18, 2 (2019), 444-470.

[5] Stewart, Radovesta, Maria Zheleva-Monova, Yanislav Zhelev, Lilia Pavlova, Detelin Luchev, Desislava Paneva-Marinova, Radoslav Pavlov. “The Orthodox Icons Collection of the Regional Historical Museum—Burgas: A Digital Library for Iconographic Objects.” Digital Presentation and Preservation of Cultural and Scientific Heritage, Vol. 5, (2015). 157-172.

[6] Luchev, Detelin, Maxim Goynov, Desislava Paneva-Marinova, Jordan Stoykov, and Lilia Pavlova. “Synergy of National Cultural Heritage and Technology.” Digital Presentation and Preservation of Cultural and Scientific Heritage, Vol. 11, (2021). 281-286.

[7] Stewart, Radovesta, Yanislav Zhelev, and Maria Monova-Zheleva. “Development of Digital Collections of Intangible Cultural Heritage Objects – Base Ontology.” Digital Presentation and Preservation of Cultural and Scientific Heritage. Vol. 11, (2021). 51-56.

Approaches to the protection of audio files in BULGARIAN LABLING CORPUS

Dimitar Popov, Velka Popova, Krasimir Kordov, Stanimir Zhelezov and Radostina Iglikova

Konstantin Preslavsky University of Shumen, Universitetska str. 115, 9700 Shumen, Bulgaria

Abstract

This article discusses the problems of protection of multimodal corpora with human speech, which are being developed in the Laboratory of Applied Linguistics (LabLing) at the University of Shumen. Ensuring the protection of audio files in the BULGARIAN LABLING CORPUS, which are provided for free access to users, is the main goal of this study. To achieve this goal, two approaches have been chosen - cryptographic and steganographic. Cryptographic and steganographic methods for protection of BULGARIAN LABLING CORPUS audio files have been proposed and verified. It has been proven that the algorithms for the implementation of the proposed methods have a high level of reliability and security, which makes them extremely suitable for the purpose of this study.

Keywords

spontaneous speech corpus, labling corpus, sound files protection, cryptography, steganography

1. Introduction

This article presents the problems of protection of multimodal corpora with human speech, which are being developed at the Laboratory of Applied Linguistics (LabLing)[1] at the University of Shumen. LabLing is part of the consortium of the Bulgarian national research infrastructure for resources and technologies for language, cultural and historical heritage, integrated within CLARIN and DARIAH (CLaDA-BG)[2].

One priority of the researchers from LabLing is to monitor the individual speech development of several Bulgarian children by conducting longitudinal observations. The methodology of Brian MacWhinney [1] was used for optimal multifaceted visualization of speech through the interactive multimodal system of the CHILDES platform, where the integrated presentation of empirical data through transcripts of audio and video recordings is possible, which are simultaneously linked to several modes of communication [2]. As a significant result of the work of the LabLing team the publication of the pilot version of the first Bulgarian CHILDES - corpus can be highlighted – BULGARIAN LABLING CORPUS[3].These data will definitely be of great importance for the formation and creation of a national interdisciplinary electronic infrastructure in the process of integration and development of electronic resources in the Bulgarian language. Therefore, the construction of LabLing CORPUS is a priority task of the CLaDA-BG consortium. In relation to that, there is an immediate need to ensure the protection of audio files in BULGARIAN LABLING CORPUS, which are provided for free access to users. This paper seeks a solution to this problem in the paradigm of cryptography and steganography.

2. Cryptographic protection of audio files of the Laboratory of Applied Linguistics

2.1. Why cryptography?

The use of cryptography for this purpose is one of the most common methods of information protection in the transmission of data on computer networks and in the exchange of information in communication channels between remote objects.

Cryptographic means of protection are special methods and means for transforming information, as a result of which its content is masked. Cryptographic transformations change the components of the messages (letters, words, numbers) in an implicit form through special algorithms, code keys or hardware solutions.

2.2. Which cryptographic methods are used?

Cryptography uses two types of cryptographic methods depending on the secret keys that are used for encryption and decryption - symmetric and asymmetric.

The concept of asymmetric cryptographic methods was introduced by W. Diffie and M. Hellman of Stanford University nearly 40 years ago. They propose the idea of creating cryptographic systems using a public key, thus

✉ labling@shu.bg (D. Popov); v.popova@shu.bg (V. Popova); krasimir.kordov@shu.bg (K. Kordov); s.zhelezov@shu.bg (S. Zhelezov); r.iglikova@shu.bg (R. Iglikova)

[1] http://labling.fhn-shu.com/home.htm

[2] https://clada-bg.eu/en

[3] https://childes.talkbank.org/access/Slavic/Bulgarian/LabLing.html

giving a new direction in the development and research of cryptographic methods [3]. In their article, they justify and define the use of public key systems called Public Key Systems (PKS). All asymmetric cryptographic algorithms are characterized by the use of key pairs, and the recognition of one of the keys cannot be used to calculate the other. The encryption process is performed with the public key, which is non-secret, and the decryption process is performed with a secret (private) key known only to the recipient of the message. In symmetric encryption, the same secret key is used to encrypt and decrypt messages.

The communication process proceeds in the following steps:

1. The sender of the message uses an encryption method, transforming the incoming message with the secret key;
2. As a result, an encrypted message is received, which is sent to the recipient;
3. The encrypted message reaches the recipient, who uses the secret key to recover the incoming message.

The symmetric approach in cryptography is also referred to as conventional cryptography. Symmetric cryptographic algorithms are divided into block and stream. In streaming encryption, each character is transformed independently of the others using the secret key, and in the block approach, the message is divided into blocks, and the transformation of the characters in the block is highly dependent.

To choose a cryptographic method, it is important to compare the advantages and disadvantages of the two types of cryptographic methods [3]. In terms of speed, symmetrical methods are preferred to asymmetrical ones.

2.3. Use of pseudo-random sequences for encryption

Pseudo-random number generators are a class of cryptographic primitives that are a major building component of any symmetric cryptographic system that performs streaming encryption. A true random sequence is that sequence of bits (0 and 1) for which the knowledge of the arbitrary subset of its elements does not give any information about the other bits. Examples of such sources are: the decay of the nucleus of a radioactive element, the thermal noise of a diode or resistor, the sound from a microphone or video input from a camera, the instability of the oscillator frequency, and others. The use of such sources is associated with a number of technical difficulties [4, 5], so the so-called pseudo-random series (PRS), and the generators of such PRS are called pseudo-random generators (PRG). Traditionally, linear-feedback shift registers (LFSR) and feedback with carry shift registers (FCSR) are used as approaches in the construction of PRG.

In recent years, chaotic maps have been used in the construction of PRG. This is due to their chaotic behavior and better cryptographic protection performance [6, 7]. This paper describes the implementation of pseudo-random generator based on two maps of this type - duffing map and circle map. The method itself is described in detail in [8]

The main steps of the algorithm that implements the method are the following:

1. The generator is initialized;
2. The samples are converted into binary form;
3. The samples are encrypted with pseudo-random sequence;
4. The samples are combined into an encrypted file.

2.4. Verification of the method

The main purpose of the cryptographic analysis is restoring the plain message from the encrypted message. In this section, in order to prove the audio encryption efficiency, we performed various empirical tests to compare plain files and their corresponding encrypted files.

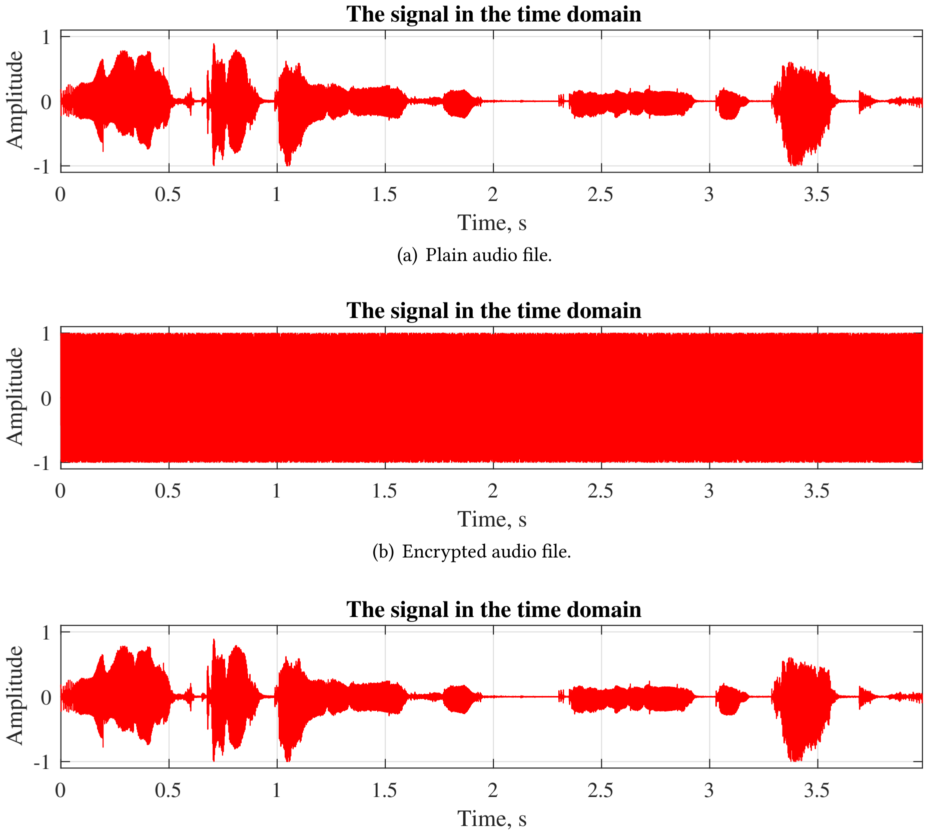
2.4.1. Waveform Plotting

One of the most common approaches, concerning audio signal analysis is waveform plotting to display the audio signal amplitude distributed in time. To compare the plain audio files with the encrypted ones we present the visualization of one of the tested files. Figure 1(a) represents the waveform of a normal file before encryption, Figure 1(b) represents the changes in the file after encryption and Figure 1(c) demonstrates the restored file after decryption.

The difference between the plain file plot and the encrypted file plot is an indication of successful encryption. Furthermore, the strong difference also means the original file cannot be restored even partially.

2.4.2. Spectrogram Plotting

The spectrogram plotting is another important approach for analyzing audio signals. In this case the main focus is the frequency of the sound against time domain. Comparing plain files with encrypted files allows us to see the difference between the files and to evaluate the proposed audio encryption algorithm. Figure 2(a) shows the spectrogram of a plain file, Figure 2(b) represents the changes in the file after encryption and Figure 2(c) demonstrates the restored file after decryption. The spectrogram plot of the encrypted file means the frequency of the original



(a) Plain audio file.

(b) Encrypted audio file.

(c) Decrypted audio file.

Figure 1: Waveform Plotting.

signal in the plain file is completely destroyed. This test is another indicator of the high encryption properties of the proposed audio encryption algorithm.

The proposed cryptographic algorithm relies on permutation-substitution architecture realized by using chaotic circle map and modified rotation equations. Extended cryptographic analysis is performed for testing the proposed method for security. The waveform plots and the spectrograms of the tested audio files demonstrate the changes in encrypted files compared to plain files.

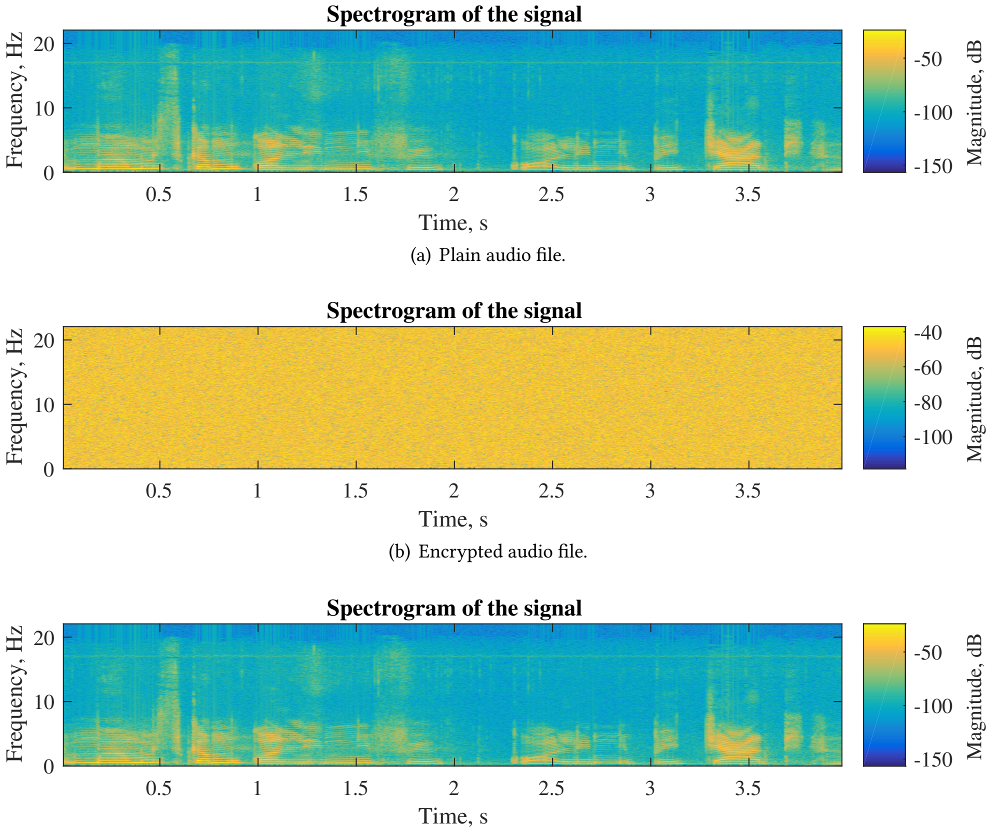
This method is suitable for information security by data transformation, but in case the data must be handled in a plain form, it must be protected from illegal use by proving authorship. For this reason, it is necessary to extend the information protection with a steganography approach.

3. Steganographic protection of audio files of the laboratory of applied linguistics.

Part of the activity of the laboratory of applied linguistics is related to the creation and processing of audio files with human speech. Recordings are subject to copyright of the laboratory and steganography methods may be used for their protection.

3.1. Why steganography?

Steganography is a scientific field of application, a set of technical skills and art for ways to hide the fact of transmission (availability) of information [9]. High-tech steganography is a term used by some authors to summarize the directions for hiding messages using communication and computer technology, nanotechnology and modern advances in biology. Steganography encompasses methods using redundancy in the binary representation



(a) Plain audio file.

(b) Encrypted audio file.

(c) Decrypted audio file.

Figure 2: Spectrogram Plotting.

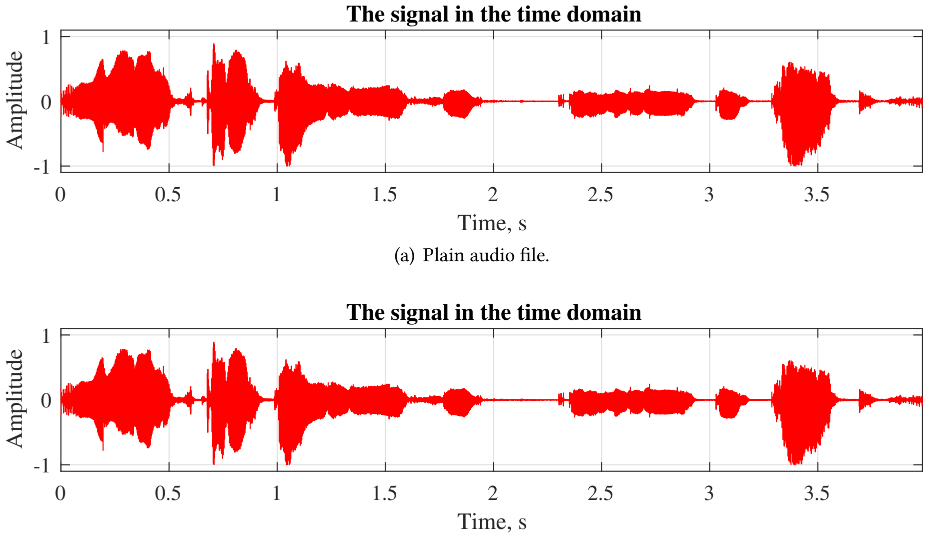
of multimedia information.

Steganographic methods (stegomethods) allow hiding data in different containers: text documents (electronic articles, books, letters), in graphic files (drawings, banners, photographs), in video files (videos, movies, animation), in audio files (music works, speech, natural sounds), in the code of HTML pages, in movie subtitles, in messages transmitted via SMS, MMS, chat, blogs, etc. Text messages can be hidden in unused areas of Flash memory, hard drives, and optical drives. Given that each type of container has different formats, and various methods can be used to hide the information, it is clear how multidimensional the steganographic tasks are.

Each of the multimedia containers has its own characteristics and each of them requires the use of specific methods for embedding and retrieving hidden information. Multimedia steganography is one of the most studied areas of computer steganography. It covers methods using the excess in the binary representation of visual and sound information. Digital images, digital music and digital video are represented by matrices of numbers that encode the intensity of colors or sound signals in space and/ or time. The lower digits of digital readings contain a very small payload for the current parameters of sounds and images. Filling them with other data does not significantly affect the quality of perception of the image or sound by people.

3.2. Which steganographic embedding methods and algorithms are used?

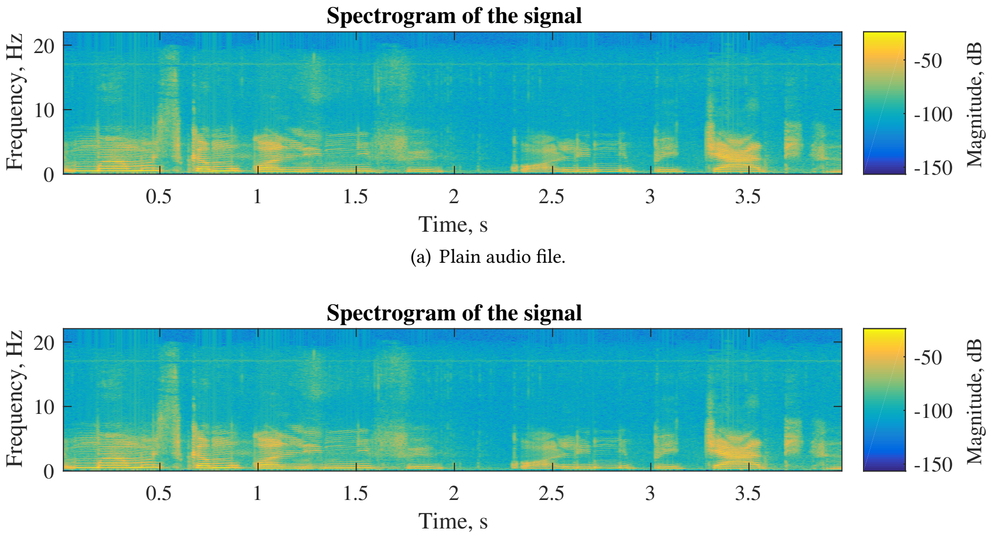
Digital images, digital music and digital video are represented by matrices of numbers that encode the intensity of colors or sound signals in space and/ or time. The lower grades of digital readings contain a very small payload for the current parameters of sounds and images. Filling them with other data does not significantly affect the quality of perception of the image or sound by people. When used with image containers, for example, about 100 KB of information can be embedded in an 800 KB



(a) Plain audio file.

(b) Stego audio file.

Figure 3: Waveform Plotting.



(a) Plain audio file.

(b) Stego audio file.

Figure 4: Spectrogram Plotting.

image file without significantly altering the container image. In an audio container - quantized sound with a duration of 1 sec., with a frequency of 44 KHz and an accuracy of 8 bits, stereo mode, the popular LSB method allows you to hide a message of about 10 KB. This leads to about 1% change in the value of the amplitude of the sound signal, which is practically impossible to detect by most people when listening to the audio file.

In recent years, many scientific developments have examined the possibilities for steganographic exchange of sensitive information provided by different types of multimedia containers - graphic files, audio files and video files. Most of them rely on a combination of different methods of steganography, and sometimes cryptographic methods are implemented in order to increase security. A typical example of this is the use of pseudo-random number generators for scattered steganographic embedding [10].

3.3. Use of pseudo-random sequences in LSB embedding

Numerous scientific publications in recent years have paid serious attention to spread spectrum steganography methods. Methods using the selection of pseudo-random positions to embed hidden information have become especially popular. These methods use different approaches to generate these pseudo-random sequences. Some of them are based on feedback shift registers [11], Others are based on different types of chaotic maps. In [8] is proposed a method of steganographic embedding in images using a pseudo-random generator based on duffing map and circle map. From these studies it is clear that this method has a high level of security and reliability, and is also fast. This makes it very suitable for use in steganographic protection of audio files. The realization of the method in steganographic protection of audio files with human speech is presented in the current development. For this purpose, empirical material from the database of the Laboratory of Applied Linguistics is used. The algorithm for its implementation contains the following steps:

1. The text information is transformed to vector V of binary sequence using ASCII table values of the characters;
2. Initializing the generator;
3. Choosing random samples (in chaotic order) from an audio, using the constructed pseudo-random generator for embedding information;
4. Using traditional LSB samples modification for hiding the values of the vector V, leaving no traces of steganography.

3.4. Verification of the method

3.4.1. Waveform Plotting

To compare the plain audio files with the stego ones we present the visualization of one of the tested files. Figure 3(a) represents the waveform of a normal file before and Figure 3(b) represents a stego file after embedding.

The lack of differences in the two files indicates that the selected method leaves no traces. Therefore, the main task of steganography, the hidden steganographic information to be invisible, is fulfilled.

3.4.2. Spectrogram Plotting

Comparing plain files with stego files shows that there is no difference between the files and allows to evaluate the proposed audio stego algorithm. Figure 4(a) shows the spectrogram of a plain file and Figure 4(b) represents the changes in the file after embedding. This test is another indicator of the high stego level of the proposed audio steganographic algorithm.

4. Conclusion.

Ensuring the protection of audio files in the BULGARIAN LABLING CORPUS, which are provided for free access to users, was the main goal of this study. To achieve this goal, two approaches were chosen - cryptographic and steganographic.

From the research described in this article, it is clear that the proposed cryptographic algorithm meets all modern requirements for cryptographic protection of information. This algorithm has a proven high level of reliability and security, which makes it extremely suitable for achieving the goal in the present study.

The same conclusion could be drawn for the proposed steganographic method. During the verification of the proposed method, it was proved that the information that marks the audio files in the BULGARIAN LABLING CORPUS as part of the database of LabLing laboratory is completely invisible. This proves the applicability of the chosen method and the validity of the implemented stego algorithm in the protection of LabLing audio files.

Acknowledgments

This research was partially funded by CLaDA-BG, the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH, Grant number DO01-377/18.12.2020.

References

[1] MacWhinney, B., Wagner, J.: Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion, Ausgabe 11. pp.154–173 (http://www.gespraechsforschung-ozs.de) (2010)

[2] Popova, V., Popov, D.: Multimodal Presentation of Bulgarian Child Language. In: Speech and Computer (SPECOM 2015). 17th International Conference, SPECOM 2015, Athens, Greece, September

20-24, 2015, Proceedings (A. Ronzhin, R. Potapova, N. Fakotakis, eds.). Springer International Publishing Switzerland, pp. 293–300 (2015).

[3] Diffie, W., Hellman, M.: New directions in cryptography. IEEE transactions on Information Theory, **22**(6), 644-654 (1976).

[4] Tamimi, A. A., Abdalla, A. M.: An audio shuffle-encryption algorithm. In Proceedings of the World Congress on Engineering and Computer Science, San Francisco, CA, USA, 22—24 October 2014.

[5] Sathiyamurthi, P., Ramakrishnan, S.: Speech encryption using chaotic shift keying for secured speech communication. EURASIP Journal on Audio, Speech, and Music Processing, **2017**(1), pp.1-11 (2017).

[6] Kordov, K. M.: Modified Chebyshev map based pseudo-random bit generator. In AIP Conference Proceedings, **1629**(1), pp. 432–436. American Institute of Physics (2014).

[7] Kordov, K.: Signature attractor based pseudorandom generation algorithm. Advanced Studies in Theoretical Physics, 9(6), pp. 287–293 (2015).

[8] Kordov, K., Zhelezov, S.: Steganography in color images with random order of pixel selection and encrypted text message embedding. PeerJ Computer Science, **7**, e380 (2021).

[9] Stanev, S., Szczypiorski, K.: Steganography Training: a Case Study from University of Shumen in Bulgaria. International Journal of Electronics and Telecommunications, **62** (2016).

[10] Kordov, K.: A novel audio encryption algorithm with permutation-substitution architecture. Electronics, **8**(5), 530 (2019).

[11] Kordov, K.: Modified pseudo-random bit generation scheme based on two circle maps and XOR function. Applied Mathematical Sciences, **9**(3), pp. 129-135 (2015).