Corpora and Processing Tools for Non-Standard Contemporary and Diachronic Balkan Slavic

Teodora Vuković **Nora Muheim** **Olivier Winistörfer**
Ivan Šimko **Anastasia Makarova** **Sanja Bradjan**

Slavic Departement, University of Zurich, Plattenstrasse 43, Zurich
teodora.vukovic2@uzh.ch, nora.muheim@uzh.ch,
olivier-andreas.winistoerfer@uzh.ch, ivan.simko@uzh.ch,
anastasia.makarova@uzh.ch, sanja.bradjan@uzh.ch

Abstract

The paper describes three corpora of different varieties of BS that are currently being developed with the goal of providing data for the analysis of the diatopic and diachronic variation in non-standard Balkan Slavic. The corpora includes spoken materials from Torlak, Macedonian dialects, as well as the manuscripts of pre-standardized Bulgarian. Apart from the texts, tools for PoS annotation and lemmatization for all varieties are being created, as well as syntactic parsing for Torlak and Bulgarian varieties. The corpora are built using a unified methodology, relying on the pest practices and state-of-the-art methods from the field. The uniform methodology allows the contrastive analysis of the data from different varieties. The corpora under construction can be considered a crucial contribution to the linguistic research on the languages in the Balkans as they provide the lacking data needed for the studies of linguistic variation in the Balkan Slavic, and enable the comparison of the said varieties with other neighbouring languages.

1 Introduction

Balkan Slavic (BS) languages are the eastern branch of South Slavic languages, which are known for their affiliation to the so-called Balkansprachbund. The languages that belong to this group are Bulgarian and Macedonian varieties as well as the Torlak dialects spoken in South East Serbia and West Bulgaria. These languages express typological differences from other Slavic languages. Some of the differentiating features are: complete or almost complete loss of the noun case inflection and fully or partially grammaticalized post-positive definite articles (Lindstedt, 2000). Many other differences lie in the nominal and verbal morphosyntax, adjectival morphology, and tense and mood system, as well as in the lexical and phraseological domain. Nonetheless, the area itself is not linguistically uniform. On the contrary - the diversification starts from the division into official standard languages, further separated by various phonological and structural isoglosses (Ivić, 1985; Institute for Bulgarian Language, 2018; Stojkov, 2002; Vidoeski, 1999). Variation occurs even in very small regions, such as the Timok area in South East Serbia, where substantial differences occur in the use of post-posed articles between villages in the mountains and valleys or urban areas (Vuković and Samardžić, 2018). There is a significant variation not only from an areal, but also from a diachronic point of view. Looking at older texts written in pre-standardized Bulgarian, phases of change in the language happening over time are noticeable.

This variation is best analyzed in non-standard varieties, unaffected or minimally affected by the prescriptive standard, and ideally in the form of spoken language since it represents arguably a more "natural"state. In this paper, we present three ongoing projects on South Slavic dialectology and diachronic linguistics, which are currently being elaborated at the Slavic linguistics department at the University of Zurich. In these projects, special focus is on places in time and space where the change happens and the underlying grammatical processes and structures. Apart from linguistic questions, the focus of our research is on the identification of potential geographical and social factors influencing changes in Bulgarian, Torlak and Macedonian.

For the purpose of this research, we are aiming at three corpora (and also processing

Proceedings of the Student Research Workshop associated with RANLP-2019, pages 62–68,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/issn.2603-2821.2019_010

tools) for modern and historical non-standard BS varieties that reflect diachronic and diatopic variation within the Balkan Slavic languages. The varieties covered are pre-standardized Bulgarian, Macedonian dialects, and Torlak varieties from Serbia and Bulgaria. Apart from the large collection of texts, the resources are equipped with part-of-speech (PoS) and morphosyntactic description (MSD) annotation, while some corpora also include a syntactic tree-bank and a layer of normalization. In order to make the corpora mutually comparable, we are developing and applying a uniform methodology. Methodology, corpora and tools are based on the existing standards used in the field as well as resources for the standard languages of the area, wherever available. This enables inter-comparability and reproducibility of the data. At the same time re-using already existing tools makes the process easier for the creators and it offers well-established methods too, which are discussed below.

2 Related Work

There is currently little data available in digital form that allows the analysis of the mentioned variation and change in BS. Bulgarian is the only language supplemented with a corpus of non-standard spoken varieties apart from the standard ones (Alexander, Ronelle and Zhobov, Vladimir, 2016; Alexander, 2015). Serbian only has corpora of written standard and computer-mediated communication (CMC) (CCSL; Utvić, 2011; Ljubešić and Klubička, 2014; Miličević and Ljubešić, 2016b). The only available searchable resource for Macedonian is an unannotated corpus of movie subtitles (Steinberger et al., 2014; Lison and Tiedemann, 2016). The mentioned corpora of standard language (or movie subtitles) are extremely valuable resources in itself, but they provide little-to-no insight into variation because they are a sample of only one variety by definition; furthermore, that standard-variety is by default controlled by people's prescribed conceptions of how language should be and how it should be written.

The only dialect corpus of BS, *Bulgarian Dialectology as Living Tradition* (Alexander, Ronelle and Zhobov, Vladimir, 2016; Alexander, 2015) is a database of oral speech comprising 184 texts from 69 Bulgarian villages, recorded between 1986 and 2013, and is 95,000 tokens in size. Texts were transcribed into Latin and Cyrillic script in order to make the data available to a wider audience. The texts are annotated with grammatical information, lemmas, English glosses for each token, and English translations for each line (see the annotated sentence in example 1, presented as it is in the corpus). The MSD are easily readable, but do not fin the standards used in the field. The database represents an impressive achievement, and a particularly valuable one given that it is the sole resource for dialectal research on BS at the moment.

(1)

hm	***kvó***	***se***
disc	what.Sg.N.Interr	Acc.Refl.Clt
.	какво	се
kázvaše		
say.3Sg.Impf.I		
казвам		

'Hmmmm. What is it called?'

The corpora described in this paper are created with the goal to be comparable with other existing related resources – be it corpora of dialects or of the standard language. The automatic processing tools developed for the language varieties included in the collection are created based on existing ones whenever possible. For example, the morphosyntactic tagger and lemmatizer for Torlak is an adaptation of the tagger for standard Serbian (Miličević and Ljubešić, 2016a).

An important tool for Serbian is the ReLDI tagger, which assigns morphosyntactic tags and lemmas to text (Ljubešić, 2016). The tagger was developed for standard Serbian, Croatian, and Slovene using a character-level machine translation method. It assigns tags specified by the MULTEXT-East standards (Krstev et al., 2004). The python package allows training for any other variety with an novel training set as an input (Ljubešić, 2016).

MULTEXT-East resources represent an extremely important milestone in the field of computational linguistic for South Slavic languages (Erjavec, 2010). The collection includes the manually annotated novel *1984* by George Orwell that can be used for training and testing of resources and specifications for morphosyntactic descriptions. The PoS labels are formulated as a string of characters, where each character stands for a grammatical category (e.g. the tag Ncfsn for the Serbian noun kuća, 'house', means 'Noun, common, feminine, singular, nominative').

The recently widely used convention of the Universal Dependencies (UD) database contains tools and specifications for the annotation of morphology and syntactic parsing for many languages, using universal grammatical categories founded in dependency grammar. The repository includes tree-banks and MSD taggers for the closely related South Slavic languages Serbian and Bulgarian.

3 Balkan Slavic Corpora

In order to bridge this gap and supply the materials necessary for the analysis of the multi-sided variation in BS, we are creating several spoken and historical corpora of varieties from the region, namely the territories in present-day Serbia, Bulgaria, and Macedonia. The individual corpora are presently at different stages of development. Contemporary materials have been collected in the past 10 years, while the historical data comprises some more recent 19th and 20th century resources that could be classified as micro-diachronical (representing a shorter time span), as well as older manuscripts dating from 16th-19th century, which provide a view on the language change on a larger scale. The goal is to establish a pipeline and tools that match the needs for non-standard varieties and produce comparable resources and would in turn allow the analysis of variation in a set of close languages.

3.1 Corpus Structure and Methodology

In order to make the corpora comparable among themselves but also with other corpora of neighboring languages, we are applying some standards from the field as well as creating them to fit a uniform structure. This will also result in ease of access and user-friendliness.

For texts which originate from audio recordings, transcripts have been made using Exmaralda (Schmidt, 2010) transcription software, developed specifically for linguistic transcription. The optical character recognition (OCR) for the printed texts and manuscripts was performed in Transkribus (Digitisation and Digital Preservation group, 2019), another piece of software created at the University of Innsbruck for automatic transcriptions of older texts and scripts.

Transcripts of spoken materials are segmented into utterances based on intonation or syntactic patterns, and each utterance is aligned with an interval on the recording. Texts that have been digitized from prints preserve the segmentation into sentences from the original version. Lastly, texts derived from written manuscripts, which do not always have clear sentence structure or punctuation, have been segmented into sentences in post-editing based syntactic structure and meaning.

Each corpus contains several layers. The minimum are the original text with automatically assigned PoS tags and lemmas, while some also contain some form of standardization or normalization. The structure of the corpus allows more information to be added over time (e.g. English glosses or an English translation).

When it comes to PoS and MSD annotation, the MULTEXT-East tag-set is used, because it is a widely accepted standard for morphologically complex languages of Eastern Europe. A further advantage is its easy adaptability to new grammatical categories, so the grammar of different varieties can be matched. We chose the MULTEXT-East tag-set over UD because they are mutually compatible. Namely, they both mark the same categories but in a different way (e.g. the MULTEXT-East tag 'Ncmsn' would be converted to the UD tag 'UPOS=NOUN, Case=Nom, Gender=Masc, Number=Sing'). They can be easily transformed from one to the other, and in fact, this has already been done for the Serbian UD Tree-bank (Samardžić et al., 2017).

All the corpora are provided with relevant metadata containing (where possible) age, gender, year of recording and main occupation of the informants as well as geo-spatial information about speaker locations. In the case of the pre-standardized Bulgarian corpus, the metadata base consists of approximate dating of the manuscripts and supposed location of its creation. The metadata for the dialectal Macedonian corpus is sometimes fragmentary because of the different working standards used and it is not possible to recover the lacking information. The metadata may be later used as a starting point for the analysis of the correlation of the linguistic data with non-linguistic factors.

The corpora are stored in files with XML markup in accordance with the TEI standards for spoken language and manuscripts (TEI, 2019). Aligned audio files are currently not supplemented with the recordings due to the

lacking infrastructure. However, recordings can be accessed on the project's YouTube channel (TraCeBa, 2015). The corpora will be made available online and freely accessible.

Each corpus has been tailored to match the methodology described above. This way different samples can be searched at the same time and the results compared. The following subsections present individual corpora on various BS varieties.

3.2 Torlak Corpus

The contemporary section of the Torlak corpus is based on fieldwork recordings from the Timok and Lužnica regions in South-East Serbia, and areas around Belogradčik in Western Bulgaria. The interviews have been transcribed using Exmaralda (Schmidt, 2010). The micro-diachronic part of the corpus includes dialect transcripts form East Serbia and West Bulgaria (Sobolev, 1998) collected by Andrey N. Sobolev in the 1990s, which have been digitized from the printed version using Transkribus (Digitisation and Digital Preservation group, 2019) as well as the data collected in the beginning of the 20th century by Belić (Belić, 1905) and Stanojević (Stanojević, 1911). Two parts of the collection have been completed so far. The collection from Timok contains around 350,000 tokens and the one from the 1990s has close to 100,000 tokens. The other data is currently being transcribed and will contain in total roughly 200,000 tokens.

Semi-phonetic transcription of spoken data have been made to reflect the spoken language as well as possible while maintaining a necessary level of readability. The transcripts of audio recordings and those of the printed interviews contain information about the accent position encoded in capital letters. They also include information about interruptions and overlaps, which is not available for the interviews recovered from print.

We have developed a PoS tagger and a lemmatizer for the contemporary spoken data from Timok and Lužnica using the ReLDI model. The training data and the lexicon combines Serbian and dialect material. For the Serbian part we used the SETTimes reference training corpus (Batanović, 2018) and the lexicon SrLex 1.2 (Ljubešić and Jazbec, 2016), both freely available. The dialect part consists of the 20,000 tokens, which have been pre-annotated with the ReLDI tagger, and then manually corrected and the lexicon derived from that sample. The accuracy of the tagger on the data Timok and Lužnica is 92.9% for the PoS labels and 93.9% for the lemmas. However, the accuracy is lower for the other sections of this corpus from the 1990s and earlier, and from Bulgaria. We are currently adding more manually annotated data from these sources to improve the results. An example of a sentence from the corpus annotated with MULTEXT-East PoS tags in the second line and lemmas in the third line is given in example 2.

(2)

On	***došAl***	***sInoč***	***iz***	***zAjčar***
Pp3msn	Vmp-sm	Rgp	Sg	Npmsa
on	doći	sinoć	iz	Zaječar

'He came last night from Zaječar.'

Apart from the morphological annotations, we are presently developing a UD-based tree-bank using the labels from the Serbian UD treebank (Samardžić et al., 2017). The data has been pre-annotated with the parser for Serbian and is being manually corrected in Arborator (Gerdes, 2013).

3.3 Macedonian Dialect Corpus

The goal in this project is to create the first corpus of spoken Macedonian dialects, annotated with PoS tags and with lemmatized tokens. The data is mainly drawn from transcripts of field-work interviews with older people from different locations collected by Vidoeski from the 1950s until the 1970s. This text-collection also includes interviews from other researchers besides Vidoeski, of which some work is considerably older than Vidoeski's; several interviews are even dating back to 1892. All the texts have been published by Vidoeski 1999. The covered period of time gives the corpus a certain diachronic depth. The texts have been transcribed using Transkribus (Digitisation and Digital Preservation group, 2019). The modern state of Western Macedonian dialects is presented by the recent field data from multi-ethnic Ohrid, Prespa, Struga and Debar regions collected in 2013 - 2019 (Makarova, 2019). The contemporary data allows a contrastive analysis of the hypothesized change.

The data comes from diverse origins, so a unified metadata scheme cannot be applied to all the collections. As these interviews were not planned as one project, every researcher defined their own standard. To partially solve this challenge and guarantee some uniformity, a standardized frame is used, where potential gaps are clearly stated.

This allows the user to decide for themselves which amount of background information is enough (e.g. if they accept an unclear year of recording or no information about the speaker's sex) and whether they want to include such parts of the corpus with missing information in their research or not.

There are currently practically no automatic tools that could be used to do PoS annotation for dialectal Macedonian, so they need to be developed specifically for this corpus. The only previous attempt to produce an automatic tagger for Macedonian has been done by Aepli et al.2014, where they solely used part-of-speech categories with no morphology at all. In our approach, we will use the MULTEXT-East tag-set for Macedonian with minor modifications to accommodate the dialectal categories not present in the standard (such as nominal cases for instance) and the ReLDI tagger. To train the initial model we will use the manually annotated corpus and the lexicon provided in the MULTEXT-East collection for standard Macedonian. After the initial training with this material we will correct the results to take the dialectal forms and variations into account. The manually annotated sample will then be used to train a new model, suitable for the many dialects covered by the corpus. The following example shows one manually annotated sentence from the contemporary material (Makarova, 2019):

(3)

Pominav	***mnogu***	***ubo***	***detstvo***.
Vmia1s-anp	Rgp	Rgp	Ncnsnn
pomine	mnogu	ubavo	detstvo

'I remember a lot from childhood.'

3.4 Pre-Standardized Bulgarian

The corpus for pre-standardized Bulgarian contains texts from the period between the 16th and 19th century, mostly, but not exclusively from present-day Bulgaria. The texts are chosen according to the similarity of their language and the vernacular. Thus, Church Slavonic texts were generally avoided, but some of them were added for reference. The collection includes texts from the Damaskini tradition either as a whole (Kotel, Ljubljana, Loveč, Tixonravov and Svištov damaskini, Pop Punčo's miscellany, perspectively also manuscript NBKM 328 of Iosif Bradati), or as a parallel corpus of multiple versions of a recurring story, (e.g. *Life of St Petka*, *Second Coming of Christ* or multiple transcripts of the *Slavobulgarian Chronicle* by Paisius of Hilandar). Parallel corpora consisting of editions of the same text from various stages and dialectal or literary backgrounds, enable us to observe the development of linguistic features or orthographic influences independently of the differences caused by contents and genre. The manuscripts have been digitized from the printed or handwritten versions using Transkribus (Digitisation and Digital Preservation group, 2019).

The MSD labels are based on the MULTEXT-East specifications for Modern Bulgarian. The purpose of this corpus is to provide material for the study of the changes in the morphosyntactic features between Middle and Modern Bulgarian. This makes the the standardized Bulgarian tag-set unsuitable. To overcome instances of ambiguity within a text and within the corpus, we adapted it to reflect both archaic and innovative features. These include nominal case markers (e.g. dative and genitive-accusative being regularly marked on both masculine nouns and adjectives), verbal infinitives (e.g. *koi može iskaza* 'who could retell') and multiple options to mark the definiteness (short- and long-forms of the adjective, articles tagged as separate tokens). Phonetic ambiguities (e.g. /i/ and /y/) were resolved by conventions based on the development of the sound in the approximate area of origin of the text. In order to avoid any over-interpretation or bias, the tags used for cases refer to morphological and not syntactical (e.g. verbal voice) or semantic (e.g. difference between common and proper nouns) criteria.

The literature of this period inherited the complex orthography of Church Slavonic. Already in the Middle Bulgarian period, it was the case that many of its rules were obsolete. Both Church Slavonic and vernacular literature attempted to follow these rules, but they weren't applied consequently. In the end, the same lemma may appear with different spellings, sometimes even within the same sentence. The different manifestations of the same lemma were partly unified by using a diplomatic transcription, eliminating ambiguous signs (e.g. accents, writing of /i/). Furthermore they were unified under the lemmas of the dictionary based on Tixonravov Damaskin, see (Demina, 2012). Turkish loanwords, Church Slavonicisms, and Russian words not included in this dictionary were lemmatized with the Etymological dictionary of BAN (Georgiev, 2006; Todorov, 2002) or with

Church Slavonic dictionaries, e.g. (Cejtlin, 1994).

The first instance of the PoS tagger and the lemmatizer was trained on a sample of 6000 manually annotated tokens using the ReLDI framework (Ljubešić, 2016). Given the unsatisfactory accuracy, we are in the process of adding more manually annotated training data. A sample of an annotated sentence from the corpus is given in the example 4. At the same time we are working on a UD-style tree-bank using syntactic labels taken from Bulgarian and Serbian specifications (Samardžić et al., 2017; Osenova and Simov, 2004).

(4)

Prědade+	***sŷ***	***dšá+***	***ta***	***bu***.
Vmia3s	Px—d	Nfsnn	Pa-fsn	Nmsdy
prědam	se	duša	ta	bog

'He surrendered his soul to the God.'

4 Conclusion

In joining our work on different languages with similar challenges, we are able to show how to deal with variation in corpora in a principled way and therefore contribute to the field of dialectology, on the one hand, and corpus linguistics on the other. Secondly, our approach demonstrates the fruitfulness of combining methodology for multiple similar projects, by taking advantage of the best practices and state-of-the-art methods and tools. The unified methodology in turn guarantees comparability of the data, which is required for the analysis of change and variation in several different varieties. The corpora under construction in the context of our projects can be considered a significant contribution to the linguistic research on the languages in the Balkans as they provide the lacking data needed for linguistic studies of BS, as well as comparison of the mentioned varieties with other neighbouring languages.

The output of these projects will be the corpora of spoken and written non-standard language equipped with with PoS annotation and lemmatization, as well as UD tree-banks. Additionally, the tools for automatic processing will be available for re-use, as well as training data and lexicons developed based on them. All of the resources will be made available online.

Acknowledgments

The development of the resources described in the paper has been funded by various sources: TraCeBa project (EraNet Rus Plus grant/Swiss National Science Foundation IZRPZ0_177557/1), Ill-bred sons project (Swiss National Science Foundation 100015_176378/1, Swiss Excellence Government Scholarship awarded to Teodora Vuković and Anastasia Makarova, Language and Space UZH, SyNoDe UZH, Slavisches Seminar UZH, Doctoral Program Linguistics UZH and Ministry of Culture and Information of Serbia. We would like to express our gratitude for their support.

References

Aepli, N., von Waldenfels, R., and Samardžić, T. (2014). Part-of-speech tag disambiguation by cross-linguistic majority vote. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 76–84.

Alexander, R. (2015). Bulgarian dialectology as living tradition: A digital resource of dialectal speech. pages 1–13.

Alexander, Ronelle and Zhobov, Vladimir (2016). Bulgarian dialectology as living tradition.

Batanović, Vuk; Ljubešić, N. S. T. (2018). Setimes.sr — a reference training corpus of serbian. pages 11––17, Ljubljana, Slovenia.

Belić, A. (1905). *Dijalekti istočne i južne Srbije*. Srpski dijalektološki zbornik. Sprska Kraljevska Akademija.

Cejtlin, Ralja Večerka; Radoslav Blagova, E. (1994). *Staroslavjanskij slovar: po rukopisjam X-XI vekov*. Russkij jazyk, Moskva.

Demina, Evgenia Mičeva; Vania Seizova, S. (2012). *Rečnik na knižovnija bălgarski ezik na narodna osnova ot XVII vek*. Valentin Trajanov, Sofia.

Digitisation and Digital Preservation group (2019). Transkribus. https://transkribus.eu/Transkribus/. [Online; accessed 09-July-2019].

Erjavec, T. (2010). Multext-east version 4 : multilingual morphosyntactic specifications, lexicons and corpora. In *V: Proceedings, LREC 2010, 7th International Conference on Language Resources and Evaluations*, pages 2544–2547.

Georgiev, V. (1972–2006). *Bălgarski etimologičen rečnik I-V*. Marin Drinov, Sofia.

Gerdes, K. (2013). Collaborative dependency annotation. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, Prague, Czech Republic. Charles University in Prague, Matfyzpress, Prague, Czech Republic.

Institute for Bulgarian Language (2018). Map of the dialectal division of the bulgarian language. http://ibl.bas.bg/bulgarian_dialects/. [Online; accessed 05-July-2018].

Ivić, P. (1985). *Dijalektologija srpskohrvatskog jezika: Uvod i štokavsko narečje*. Matica srpska, Novi Sad.

Krstev, C., Vitas, D., and Erjavec, T. (2004). Morpho-Syntactic Descriptions in MULTEXT-East - the Case of Serbian. *Informatica, No. 28*.

Lindstedt, J. (2000). Linguistic balkanization: Contact-induced change by mutual reinforcement. In Gilbers, D., editor, *Languages in contact*, number 28 in Studies in Slavic and general linguistics. Rodopi, Amsterdam.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Ljubešić, Nikola; Klubička, F. A. Z. and Jazbec, I.-P. (2016). New inflectional lexicons and training corpora for improved morphosyntactic annotation of croatian and serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Ljubešić, Nikola; Klubička, F. A. Z. J. I.-P. (2016). New inflectional lexicons and training corpora for improved morphosyntactic annotation of croatian and serbian. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož. European Language Resources Association (ELRA).

Ljubešić, Nikola; Klubička, F. A. Z. J. I.-P. (2016). New inflectional lexicons and training corpora for improved morphosyntactic annotation of croatian and serbian. In Chair), N. C. C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Ljubešić, N. and Klubička, F. (2014). Wac – web corpora of bosnian, croatian and serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29—35.

Makarova, A. (2019). Unpublished work.

Miličević, M. and Ljubešić, N. (2016a). Tviterasi, tviteraši or twitteraši? producing and analysing a normalised dataset of croatian and serbian tweets. *Slovenščina 2.0*, 4(2):156–188.

Miličević, M. and Ljubešić, N. (2016b). *Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets*. Slovenščina 2.0.

Osenova, P. and Simov, K. (2004). Btb-tr05: Bultreebank stylebook 05. Technical report.

Samardžić, T., Starović, M., Agić, v., and Ljubešić, N. (2017). Universal dependencies for serbian in comparison with croatian and other slavic languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 39–44, Valencia, Spain. Association for Computational Linguistics.

Schmidt, T. (2010). Exmaralda: un systéme pour la constitution et l'exploitation de corpus oraux. *Pour une épistémologie de la sociolinguistique. Actes du colloque international de Montpellier*, pages 319–327.

Sobolev, A. (1998). *Sprachatlas Ostserbiens und Westbulgariens*. Scripta Slavica. Biblion.

Stanojević, M. (1911). *Severno-timački dijalekat*. Dijalektološki zbornik. Sprska Akademija Nauka.

Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., and Gilbro, S. (2014). An overview of the european union's highly multilingual parallel corpora. *Language resources and evaluation*, pages 679–707.

Stojkov, S. (2002). *Bălgarska dialektologija*. Akad. izd. Prof. Marin Drinov, Sofia.

TEI, T. E. I. (2019). *P5: Guidelines for Electronic Text Encoding and Interchange*.

CCSL. Corpus of contemporary serbian language - official website (in serbian). http://www.korpus.matf.bg.ac.rs/prezentacija/istorija.html. [Online; accessed 08-March-2018].

Todorov, T. A., editor (2002). *Bălgarski etimologičen rečnik VI-VII*. Sofia.

TraCeBa (2015). Terenska istrazivanja. https://www.youtube.com/channel/UC4EpCSAnEb2RIsIRY7pfNdQ/feed. [Online; accessed 20-August-2019].

Utvić, M. (2011). Anotacija korpusa savremenog srpskog jezika. *INFOteka 12, br. 2*, pages 39–51.

Vidoeski, B. (1999). *Dijalektite na Makedonskiot Jazik - tom 2*. MANU, Skopje.

Vuković, T. and Samardžić, T. (2018). Prostorna raspodela frekvencije postpozitivnog člana u timočkom govoru. In Ćirković, Svetlana (Ed. in chief) and Andrej N. Sobolev, Barbara Sonnenhauser, Maja Miličević, Jelenka Pandurević, editor, *Timok. Terenska istraživanja 2015-–2017*. Narodna Biblioteka "Njegoš", Knjaževac.