TeodoraVuković
Faculty of Philology
University of Belgrade
Maja Miličević
Faculty of Philology
University of Belgrade

UDC 811.163.42'322.2
DOI https://doi.org/10.18485/fid.2017.7.ch21

CREATION AND SOME IDEAS FOR CLASSROOM USE OF AN ELECTRONIC CORPUS OF THE DIALECT OF BUNJEVCI

Овај рад представља пројекат који је у току, а посвећен је изради електронског корпуса дијалекта којим говори заједница Буњеваца, у области Бачка на северу Србије. Материјал је сакупљен између 2009. и 2012. године у Суботици и околини; садржи око 60 сати снимака и процењено је да има 743.500 речи. Разрађујемо како су три снимка узорка трансформисана у (пилот) корпус формат, разматрајући избор лингвистичких и металингвистичких кодираних варијабли, као и описујући нормализацију усвојених стратегија како би се омогућило коришћење аутоматских алата за обраду корпуса, као и различитих врста упита. На крају, дати су примери како се овај корпус може користити у образовне сврхе.

Кључне речи: буњевачки дијалекат, дијалекатски корпуси, морфосинтаксичка анотација, нормализација.

1. Introduction

The dialect of Bunjevci is one of Serbia's minority languages. Due to its unique properties it is interesting for numerous linguistic explorations; however, it lacks resources that would enable or facilitate this. In this paper, we present initial steps and further plans for building a corpus of this dialect, which can be seen as a new resource for linguistic research, as well as a fertile ground for developing a more general methodology for dialect corpus building. We describe the steps involved in the creation of the pilot corpus we built, the problems we encountered, and the solutions we adopted; the methodology we used was partly based on existing dialect corpora, and partly on our own decisions. The corpus is derived from audio recordings of ethnolinguistic interviews, which were first transcribed,

and then normalized into standard Serbian, to enable further processing using existing computational tools. As the dialect of Bunjevci has recently been introduced in schools in Serbia, we also give some suggestions for possible uses of the corpus in the classroom.

Before moving on to the descriptions of the dialect of Bunjevci and the corpus we created, we wish to point out that while we do plan to make the final corpus publicly available, at the moment the pilot corpus can only be accessed by contacting the authors.

2. The dialect of Bunjevci

Bunjevci are an ethnic group that lives in the northern Serbian region of Bačka, primarily in the cities of Subotica and Sombor and the surrounding villages (Georgijević 1939, as cited in Bošnjaković 2013: 189). They constitute a group that has been widely studied from both historical-cultural and linguistic points of view. In the Serbo-Croatian literature, the linguistic variety spoken by Bunjevci is sometimes treated as a language, but it is more typically referred to as *bunjevački govor*‘ Bunjevac speech’, due to the lack of standardisation (see Bošnjaković 2013: 189-190). In this paper, we opt for the term ‘dialect’, following the English (narrower) rather than Serbo-Croatian (wider) use of the term.

The dialect of the Bunjevci belongs to the younger Ikavian group of the Štokavian dialects (Bošnjaković 2013: 189), and is the only Ikavian dialect in Serbia. Its distinguishable phonological features are Ikavian forms such as *mli**i**ko*‘milk’, *d**i**da*‘grandfather’, *ned**i**lja*‘Sunday’, *s**i**no*‘hay’, *tr**i**šnja*‘cherry’(standard Serbian equivalents being *ml**e**ko*, *d**e**da*, *ned**e**lja*, *s**e**no*, *tr**e**šnja*). Among the striking morphological characteristics we find truncated infinitives (e.g.*radit*‘work’, *kopat*‘dig’, *postignit*‘ achieve’, *doć*‘come’,whose standard equivalents are *raditi*, *kopati*, *postignuti, doći*), and 3rd person plural present tense forms of some verbs (e.g. *radidu*‘(they) work’,*gledaje*‘(they) watch’,*peču*‘(they) bake’, where the respective standard forms are *rade*,*gledaju*, and *peku*). Finally, the dialect of Bunjevci, like any other dialect, has many lexemes not used in standard Serbian, such as *dužijanca* ‘end-of-harvest holiday’ or *fanak*‘donut’.

3. Dialect corpora: rationale and some issues

3.1. Why and how to create a dialect corpus?

Over the past decades, digital textual collections in the form of language corpora have become widely used in linguistic research. Corpora are highly convenient sources of language data, as they tend to be large and they are not composed of raw text that users have to go through manually, but are segmented and annotated to enable advanced search types that cover anything from phonetic to syntactic phenomena, and help avoid unwanted data.

Unlike corpora of standard (written)language, which are fairly widespread, dialect corpora are still scarce, mostly due to the fact that dialect data is more difficult to collect and work with. At the same time, the scarceness makes existing corpora all the more valuable. Creating a corpus of the dialect of Bunjevci can greatly facilitate its description and its comparison with standard Serbian and other dialects of Serbo-Croatian (or other South Slavic languages);at the same time, this corpus can serve as a testing ground for developing tools and methods for handling dialect corpora in general. Finally, corpora can have numerous educational uses, and given that the dialect and the culture of Bunjevci are taught in schools, teachers could employ the corpus to find authentic examples of language use and include them in class materials, or students could use the corpus autonomously.

Several key points need to be decided upon before embarking on the creation of a corpus of this kind. The first one is the transcription method. The two main options, which can be seen as two extremes, are phonetic transcription using IPA symbols in order to literally transcribe non-standard phonology, and orthographic transcription using standard orthography, ignoring the dialectal characteristics, and delivering a text conforming to the standard. As the first type is difficult to search, and the second does not keep some of the dialectal characteristics, a third option has also been used – semi-phonetic transcription, i.e. the use of the regular alphabet to represent the morphophonological peculiarities not pertaining to the standard; such transcripts are easy to read and enable automatic processing.

Another major issue is the purpose of the corpus, which needs to be established in advance as it determines the levels of data segmentation and annotation, and the methods to be used. Depending on the researcher's' needs and interests, it is possible to annotate part-of-speech (POS) categories, syntactic relations, prosody, pragmatic markers, etc., where POS annotation and lemmatization are considered almost mandatory.

3.2. Dialect corpora of other languages

Dialect corpora demand more manual work than standard language corpora, and are thus quite rare. In most cases they are based on audio recordings that need to be transcribed and then transformed into a corpus. In this section, we briefly describe three well-known dialect corpora in order to illustrate how some of the problems relevant for spoken dialect data can be solved. The methods used in building these corpora have served as partial models for our corpus of the dialect of Bunjevci.

The Freiburg English Dialect Corpus – FRED (Anderwald & Wagner 2007) is assembled from traditional oral dialect data with the goal of enabling the study of non-standard morphosyntax. It contains approximately 2.5 million words derived from about 300 hours of recordings in the whole territory of the British Isles. The authors opted for an orthographic transcription with semi-phonetic elements; due to the focus on morphosyntax, pauses, laughter, hesitations, and similar elements were not marked. The Syntactic Atlas of the Dutch Dialects – SAND (Barbiers,Cornips,&Kunst 2007)covers 267 dialects in the Netherlands, parts of Belgium and France. It is based on syntax-related test sentences that the informants had to judge or translate, and on short spontaneous speech recordings for each dialect. The collected data was transcribed in Praat using orthographic transcription in order to make the sample more uniform and to enable automatic annotation. Dialectal forms were normalized to standard Dutch, with the exception of functional morphemes, which were transcribed as they were pronounced. In addition to the POS annotation, syntactic features, relations and word order are marked too. The Nordic Dialect Corpus – NDC (Johannessen, Vangsnes, Priestley, & Hagen 2014), a corpus of Northern Germanic dialects, contains around 2.8 million words and encompasses

dialects from Denmark, Faroe Islands, Sweden, Iceland and Norway. In the data collection process, 801 informants were asked to spontaneously converse (in pairs) for thirty minutes. For some dialects, interviews were also conducted. Orthographic transcription with semi-phonetic elements was used in this case too, solving the problem of multiple transcribers from multiple countries. This approach also enabled automated grammatical annotation using unified POS tags for different languages.

In sum, the three corpora have in common that (1)they were based on transcribed interviews,(2) the transcripts were orthographic with semi-phonetic elements, (3) there was some form of normalization in the transcripts, or in a separate layer added to the corpus, in order to facilitate corpus querying and enable the application of automatic processing tools.

4. The creation of a pilot corpus of the dialect of Bunjevci

4.1. Materials and the pilot sample

The sample used for creating the pilot corpus described in this paper is part of a larger collection of recordings made during fieldwork research of the Bunjevci dialect and culture in the period 2009-2012, conducted by researchers from the Institute for Balkan Studies of Serbian Academy of Sciences and Arts and Faculty of Philosophy of the University of Novi Sad. Their study was a qualitative one, focused on the rural dimension of the life and dialect of Bunjevci;the researchers employed semi-structured ethnolinguistic interviews and covered topics related to tradition, culture, autobiographical narratives, and everyday life. Over 30 hours of recordings were collected, parts of which (perceived as most relevant in terms of ethnolinguistic topics) were transcribed, for a total of about 18,000 words. The results were published in the monograph *Bunjevci:Etnodijalektološka istraživanja 2009* (Bošnjakovi ć& Sikimić 2013).

The sample we used for the pilot corpus was collected in 2009 in the area around the city of Subotica. Three recordings were selected and fully transcribed. The total duration of the recordings is 7 hours and 20 minutes, with the transcripts amounting to about 45,000 words. A total of twelve speakers participated in the recordings, including three interviewers who

predominantly used standard language, and whose turns were therefore excluded from the pilot corpus;they will be added to the final version, with the possibility of filtering them out in queries. Another participant whose production was excluded is a five-year-old child who spoke for about 20 seconds.That left us with eight speakers relevant as representatives of the Bunjevac dialect, six women and two men, with an average age of 63 years.

4.2. Transcription

The first version of the transcripts was created within the original ethnolinguistic study, which had objectives different from our own. The researchers transcribed what they found interesting and relevant based on two sets of criteria: linguistic (dialectological and sociolinguistic) on the one hand, and historic and cultural on the other. In other words, they selected for transcription those parts of the recordings, ranging from 2-3 up to over 20 minutes in length, that they found relevant for presenting the Bunjevac tradition and culture (see Sikimić 2013: 45).

The original transcripts did not include an elaborate internal segmentation, and each transcriber adopted a somewhat different system for segmenting text and for marking speakers and non-linguistic elements; for instance, some transcribers separated the interviewers' questions on a new line while others put them in brackets. Replicas by different informants were written in continuation, separated by dashes, in an attempt to create a form of "collective narrative" (Sikimić 2013: 49). The transcripts were segmented according to topic, but digressions were usually not separated. Non-linguistic elements (expression of emotions, non-verbal interactions among speakers, or interactions with the surroundings) were not marked. Two examples from the original transcripts are shown in (1).

(1)

a. Pa to se išlo na Bunarić kad je Proštenje, a sad sad se iđe. – Prva subata u mesecu.

BS: A zašto?

Misa. Razumete? Misa. Svaka prva subota u mesecu se ide na Bunarić, ima misa. – Ali kad počinje? – Pa u devet. – U maju, do septembra. – E sad, danas smo bili, bar koliko sam ja razumio.

b. Kako-s kad se ona bila mlada, ona je sad sedamdest osam godina, od-mene je starija dvi godne, godnu i po dana. Uglavnom, uskočla je. E onda mama joj nije dala ništa. A kod svekrove nećedu. – Jel već znate već šta to znači da uskoči? (SĐ: Da uskoči? To je da.) Da pobegne. (SĐ: Znam, znam.) to se išlo na Bunarić kad je Proštenje, a sad sad se iđe. – Prva subata u mesecu.

The transcribers' goal was to reproduce the spoken language from the recordings as faithfully as possible, without influencing the readability of the texts. For this reason they opted for an orthographic transcription system with semi-phonetic elements for dialect features such as *kruv* 'bread'*, vidili*'(they) saw', *odranit* 'raise'. A similar procedure was adopted for spoken language features such as phoneme omissions and phonetic quality changes (e.g. *očla* was not corrected to *otišla*'(she) went', *kašte* to *kažite* 'say', *is-kuće* to *izkuće* 'from the house', *vid'la*or *vidla* to *videla*'(she) saw'). Accent movements and phonetic words were marked with a hyphen (e.g. *od-kruva* 'of bread', *na-peć*'on the stove', ***Kako-ć*** *bit vrime?*'What will the weather be like?'). However, these interventions were not implemented fully consistently across the transcripts.

When these original transcripts were being adapted for corpus creation, we kept the orthographic transcription with semi-phonetic elements as the most readable alternative that still shows the characteristics of the dialect. We changed the way the texts were segmented, and each speaker's replicas were inserted on a new line; we also assigned each speaker a code following the order of appearance (GOV1, GOV2…, for Speaker 1, Speaker 2, etc.). Since only a single interviewer was present per recording, we adopted a unique interviewer code – ISTR (for *istraživač* 'researcher'). Replicas were not internally segmented into paragraphs; we did not mark overlaps, and we segmented the dialogs in such a way that they formed meaningful conversations.

Since it was decided that the corpus would be used primarily for morphosyntactic analyses,only the linguistic content of the interviews was considered important. As a consequence, we did not include any information about non-verbal or other non-linguistic content. To make the task of automatic processing and user querying easier,we tried to use as few symbols as possible in addition to the letters of the alphabet and punctuation

signs: apostrophes marking phonetic elisions were thus removed(e.g.*vid'la* was changed to *vidla*‘(she)saw’); hyphens were also removed because words needed to be separated, so we changed, for instance,*is-kuće* to *is kuće*. Irregularities such as self-corrections, errors, hesitations, and repetitions were not corrected and were transcribed literally. Two extracts from the final transcripts are given in (2).

(2)

a. GOV2:

Pato se išlo na Bunarić kad je Proštenje, a sad sad se iđe.

GOV1:

Prva subata u mesecu.

ISTR:

A zašto?

GOV2:

Misa. Razumete? Misa. Svaka prva subota u mesecu se ide na Bunarić, ima misa.

GOV3:

Ali kad počinje?

GOV2:

Pa u devet.

GOV1:

U maju, do septembra.

GOV2:

E sad, danas smo bili, bar koliko sam ja razumio.

b. GOV2:

Kako s kad se ona bila mlada, ona je sad sedamdest osam godina, od mene je starija dvi godne, godnu i podana. Uglavnom, uskočla je. E onda mama joj nije dala ništa. A kod svekrove nećedu.

GOV1:

Jel već znate već šta to znači da uskoči?

ISTR:

Da uskoči? To je da.

GOV2:

Da pobegne.
ISTR:
Znam, znam.

One last note concerns the fact that we left out from the transcripts parts of conversations that were deemed too private, and personal names were replaced with initials. We also omitted interchanges between the interviewers and the speakers about the organization of the research itself, as they involved more interventions on the part of the interviewers. Unintelligible words or sections were marked with *(...)*.

4.3. Normalization

To enable the use of tools for automatic processing designed for standard Serbian on dialect data, we decided to normalize the corpus by adding annotation layers in which non-standard forms are replaced by their standard equivalents. In this procedure we consulted a dictionary of standard Serbian (Stevanović,Marković, Matić,& Pešikan1990) and a dictionary of the Bunjevac dialect (Peić &Bačlija 1990). Due to requirements related to further processing, done at word level, each word was assigned a standard value; the transcripts were verticalized first, meaning that every word or sign was on a separate line, with normalized values added as new columns.

Language varieties can differ from the standard on many levels; we distinguished morphophonological, syntactic and lexical discrepancies as three separate layers of normalization. In the first layer we normalized the phonological features of the Bunjevac dialect (e.g. *vidila* to *videla* '(she) saw'), phonetic omissions and other changes (e.g. *volla*to *volela*'(she) loved'), as well as differences in morphemes (e.g. *mislit* was changed to *misliti*' think', *gledaje* to *gledaju* '(they) are looking').In the second layer we corrected syntactic differences (e.g. *Ne sćam se jana to*to *Ne sećam se ja toga*'I do not remember that'), and we translated lexemes that have an equivalent in standard Serbian (e.g. *ris*to *žetva* 'harvest'). In the third layer we only marked (with an asterisk) those lexemes that have no equivalent in standard Serbian. An example of the result can be seen in Table 1.

Original	NORM1	NORM2	NORM 3
samo	samo	samo	samo
imam	imam	imam	imam
šotošku	šotošku	šotošku	šotošku*
,	,	,	,
pa	pa	pa	pa
kad	kad	kad	kad
ne	ne	ne	ne
sćam	sećam	sećam	sećam
se	se	se	se
ja	ja	ja	ja
na	na	***	***
to	to	toga	toga
.	.	.	.

Table 1 – Normalization example

4.4. Annotation and lemmatization

Having normalized the data, we were able to use the annotation and lemmatization tool created for standard Serbian by Gesmundo and Samardžić (2012). As can be seen in Table 2, to look at the usefulness of normalization, the two processes were performed both on the original andon the normalized texts(which differed in 7-10% of the cases, depending on the recording).The tool performance on normalized data was by 2% better than the performance on the original texts (91 vs. 89% accuracy) for POS tagging, and by 3% better for lemmatization (97 vs. 94% accuracy), pointing to the usefulness of normalization for automatic corpus processing(details of the comparisons can be found in Vuković 2015).

ORIGINAL	POS1	LEM1	NORM1	POS2	LEM2	NORM2	NORM3
S	Sp	s	S	Sp	s	S	S
otim	Pd	otaj	tim	Pd	taj	tim	tim
siče	Nc	siča	seče	Vm	seći	seče	seče
snopove	Nc	snopova	snopove	Nc	snop	snopove	snopove
.	#	.	.	#	.	.	.
A	C	a	A	C	a	A	A
sitna	Af	sitan	sitna	Af	sitan	sitna	sitna
pliva	Nc	pliv	pleva	Nc	pleva	pleva	pleva
,	#	,	,	#	,	,	,
prvo	Rg	prvo	prvo	Rg	prvo	prvo	prvo
iđe	Vm	ići	ide	Vm	ići	ide	ide
slama	Vm	slamati	slama	Vm	slamati	slama	slama
iz	Sp	iz	iz	Sp	iz	iz	iz
doba	Nc	doba	doba	Nc	doba	dreša	dreša

Table 2 – Annotation and lemmatization example

4.5. Query options

As the corpus was not only annotated and lemmatized, but also processed in the Corpus Workbench platform, it allows for different types of queries using CQL – Corpus Query Language.Original texts as well as each additional layer can be searched separately, and any two or more levels of data can be searched simultaneously. For example, one can find all words in which an*i*in the Bunjevac dialect corresponds to an *e* in the standard, thus obtaining a list of all Ikavian forms; verbal or any other word forms can be looked up in a similar way. Normalized layers can also be of use to those who are not familiar with specific dialectal forms. The lemmatization layer can help users find all forms of a lexeme with a single query. On the other hand, part of speech categories can be used to find all the words belonging to a particular grammatical category; since the list of hits would be too long in this case, it is best to search in another layer simultaneously, for instance, all verbs that end in *-aje*. Some of the query options are illustrated in the next section.

5. Possibilities of classroom use

Most corpora can find a use in the language classroom, be it in the context of a native or non-native language. In the specific case of dialects, an additional dimension concerns the preservation of the dialect itself and of the (knowledge of) traditions associated with its speakers.

As far as the dialect of Bunjevci is concerned, the national minority status of Bunjevci ensures the dialect's presence in the Serbian education system, through the subject "Bunjevac Speech with Elements of National Culture". Several possible educational uses can thus be envisaged for the corpus once it is completed and made freely available, which could complement the existing teaching materials (in particular the recently published grammar and reading books; see Kujundžić Ostojić, Josić, & Tikvicki 2014, Savanov&BašićPalković2014). Some of these uses might be better suited for teachers,helping them in the preparation of class materials, but in many cases students could also query the corpus autonomously, during class or at home. Concrete examples are provided below:

(3) shows how the corpus could be employed for developing grammar exercises, while (4) relies on the fact that the texts in the corpus are about the traditional culture and history of Bunjevci, and as such can be used as material for teaching about topics such as customs and traditions, forgotten games, or the traditional way of life on the farm; lastly, (5)shows how the meaning of dialectal words can be extracted from the corpus.

(3) **(Non-)standard 3rd person plural present tense verb forms**
CQL query: [norm1=".+aju"]
Results sample:
28576: nolitna , onda su tili da <**dadu**> brata mi u dom . Ja sam
28843: kad je moj sin živ bio . <**Imadu**> dite , sa mojim sinom im
31583: tovo . I , a sad i pelene <**peglaje**> . Dok se ne krsti , dotl
44726: i kako se nrz , kako jaja <**šaraju**> . I onda imamo , ponedel
36700: l , e onda se kosti bolje <**otvaraje**> . E , barem tako su star
43190: , posle večere roditelji <**bacaju**> orahe , kod vrata , u sl
37271: ća . Et to je , ovako se <**smotaje**> , vidiš , ovako se smot

The CQL query shown in (3)searches through the normalized forms (at the morphophonological level) and detects words ending in *–aju*, i.e. 3rd person plural present tense verb forms; among the results, where by default only the original forms are displayed, it shows both verbs with the standard ending, and those with non-standard ones (*–du* and *–aje*). Used in an exercise, sentences such as those listed above could serve as a prompt for asking students to identify the different endings that are allowed in the dialect of Bunievci, look at the distribution of standard vs. non-standard endings, etc.

(4) **Tradition-specific concepts or wider topics**
CQL query: set Context 2 s
[lemma="dužijanca|beba"]
Results sample:
4567: Di je bilo svetenje , u Tavankutu , kad ja nisam tela ić . Jako šteta lani niste bili kod nas , ali neka , ali možemo vas pozvati , u Tavankut na <**dužijancu**> , tamo se svi običaji bunjevački vide , to je vraćeno

unazad skroz , i nošnja , i mi smo prošle godine imali i baš smo i nošnju , i sve baš , tu je , mislim bila sam tu među organizatorima . To je pravo bunjevački stvarno .

37298: .) Evo vidiš , to je suknena pregača . Et to je , ovako se smotaje , vidiš , ovako se smota , to je nova , još ni niko nije , a ona tak metla doli , to se ovako priko <**bebe**> metne , vidiš , kad oćeš digod ić . Al to mamina pregača koju koristi , razumiš , suknena , to je , to je vuna , vako , priko bebe se metne , priko dunjice .

In example (4), the context is first set to showing a sentence to the left, and a sentence to the right from the one containing the required key word; two lemmas are subsequently searched for, giving as the result pieces of texts dealing with the chosen topics. A similar query can help quickly identify portions in the material where specific topics are discussed, making a theme-based selection of didactic material relevant for the study of the Bunjevac culture much more efficient.

(5) **Discovering word meanings**
CQL query: show +norm2
"parasnička|kuružna|kiseline"

65: To/To s/se kazala/zvala <**parasnička/seljačka**> peć/peć se/se zvala/zvala ./.

72: Ložila/Ložila se/se <**kuružna/kukuruzovina**> u/u nju/nju i/i onda/onda pečemo/pečemo kruv/kruh i/i eto/eto ./.

83: Kiselili/Kiselili <**kiseline/jogurt**> ./.

8997: Opere/Opere se/se i/i on/on pusti/pusti svoju/svoju ,/, on/on je/je pun/pun <**kiseline/kiseline**> ,/, to/to je/je u/u stvari/stvari želudac/želudac ,/, ne/ne ,/, gde/gde se/se zadržava/zadržava hrana/hrana ./.

Example (5) illustrates yet another advantage of a normalized dialect corpus. In addition to allowing for a search of dialectal words in authentic language use, the corpus lets users find meanings of words by displaying the normalized text alongside the original. In this particular example, the second – lexical – level of normalization is displayed, enabling the users to see the standard language "translations" of the dialectal lexemes.

6. Conclusion

Working on the corpus of the dialect of Bunjevci, we have come up with a methodology that has not only proven to be efficient in the case of this corpus, but can also be applied to other dialect corpora. In particular, orthographic transcription with semi-phonetic elements is easy to use, while it maintains the dialectal and spoken language features, essential for a non-standard language resource; normalization allows for the use of POS annotation and lemmatization tools created for the standard language, and enables a wide range of potentially useful queries. The corpus can be of use to linguists and beyond; as it contains different kinds of information about Bunjevci, their past and present, tradition and culture, it can also be useful for ethnographers, anthropologists, historians, sociologists, etc. As we have shown, it can in addition be employed in teaching about the linguistic properties of the dialect of Bunjevci, or their culture. In the future, we plan to add more texts and extend the pilot version, and we intend to make the audio recordings available alongside the corpus.

References:

Anderwald, L., & Wagner, S. (2007). FRED – The Freiburg English Dialect Corpus: Applying corpus-linguistic research tools to the analysis of dialect data. In J. C. Bael, K. P. Corrigan,& H. L. Moisl (Eds), *Creating and Digitizing Language Corpora. Volume 1: Synchronic Databases* (pp. 35-53). New York: Palgrave Macmillan.

Barbiers, S., Cornips, L., & Kunst, J. P. (2007). The Syntactic Atlas of the Dutch Dialects (SAND): A corpus of elicited speech and text as an online dynamic atlas. In J. C. Bael, K. P. Corrigan,& H. L. Moisl (Eds), *Creating and Digitizing Language Corpora. Volume 1: Synchronic Databases* (pp. 54-90). New York: Palgrave Macmillan.

Bošnjaković, Ž. (2013).Govori bačkih Bunjevaca i tradicionalna dijalektologija. In Ž. Bošnjaković & B. Sikimić (pp. 187-230).

Bošnjaković, Ž., & Sikimić, B. (2013) *Bunjevci: Etnodijalektološka istraživanja 2009.* Subotica/Novi Sad: Nacionalni savet bunjevačke nacionalne manjine/Matica srpska.

Chambers, J. K.,& Trudgill, P.(2004). *Dialectology*. 2nd edition. Cambridge: Cambridge University Press.

Evert, S. (2010). *The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial*. Retrieved on 10 January 2016, from http: //cwb. sourceforge. net/files/CQP_Tutorial/

Georgijević S. (1938). Bački bunjevački govor.*Godišnjak Zadužbine Sare i Vase Stojanovića,VI*, 23-32.

Gesmundo, A., & Samardžić, T. (2012). Lemmatisation as a tagging task. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Volume 2: Short Papers* (pp. 368-372). Jeju Island, Korea: Association for Computational Linguistics.

Johannessen, J. B.,Vangsnes, Ø. A., Priestley, J., & Hagen, K. (2014).A multilingual speech corpus of North-Germanic languages. In T. Raso & H. Mello (Eds).*Spoken Corpora and Linguistic Studies*(pp. 69-83). Amsterdam: John Benjamins.

Kujundžić Ostojić, S., Josić, R., & Tikvicki, J. (2014). *Klupče. Moja prva bunjevačka gramatika od 1. do 4. razreda osnovne škule*. Beograd: Zavod za udžbenike.

Peić, M., & Bačlija, G. (1990). *Rečnik bačkih Bunjevaca*. Novi Sad: Matica srpska.

Savanov, M., & Bašić Palković, N. (2014). *Bunjevačka čitanka za 1.i 2. razredo snovne škule*. Beograd: Zavod za udžbenike.

Sikimić, B. (2013). Između dijalektologije i antropologije: bunjevačka terenska građa. In Ž. Bošnjaković & B. Sikimić (pp. 13-68).

Stevanović, M., Marković, S., Matić, S., & Pešikan, M. (1990). *Rečnik srpskohrvatskoga književnog jezika*. 2nd edition (reprint). Novi Sad: Matica srpska.

Tanasijević, I., Sikimić, B., & Pavlović-Lažetić, G. (2012). Multimedia database of the cultural heritage of the Balkans. In *Proceedings of LREC 2012*(pp. 2874-2881).

Vuković, T. (2015). *Izrada modela dijalekatskog korpusa bunjevačkog govora* (Unpublished master's thesis). Belgrade: Faculty of Philology, University of Belgrade.

Abstract

This paper presents an ongoing project devoted to building an electronic corpus of the dialect spoken by the Bunjevci community, in the northern Serbian region of Bačka. The material discussed was collected between 2009 and 2012 in the city of Subotica and its surroundings; it amounts to approximately 60 hours of recordings and an estimated 743,500 words. We elaborate on how three sample recordings were transformed in (pilot) corpus format, discussing the choice of linguistic and metalinguistic variables coded, and describing the normalization strategies adopted in order to enable the use of automatic corpus processing tools, as well as different types of queries. Lastly, examples are provided of how the corpus can be employed for educational purposes.

Keywords: Bunjevac dialect, dialect corpora, morphosyntactic annotation, normalization.

Biographical statement

TEODORA VUKOVIĆ, MA, first did a BA in General Linguistics and then proceeded to graduate from the Faculty of Philology at the University of Belgrade with an MA thesis titled *Creation of a Pilot Corpus of the Dialect of Bunjevci*. She is an external associate researcher at the Institute for Balkan Studies of the Serbian Academy of Sciences and Arts.

E-mail: bravethea@gmail.com

MAJA MILIČEVIĆ, PhD, is an Assistant Professor in Applied Linguistics in the Faculty of Philology at the University of Belgrade. Her recent publications include a monograph about experimental methods in second language acquisition research (*Eksperimentalne metode u istraživanjima usvajanja drugoga jezika*, co-authored with Tihana Kraš), and papers "Translation between L2 acquisition and L1 attrition: Anaphora resolution in Italian by English-Italian trainee translators" (*Applied Linguistics*, co-authored with Tihana Kraš) and "From EPIC to EPTIC – Exploring simplification in interpreting and translation from an intermodal perspective" (*Target*, co-authored with Silvia Bernardini and Adriano Ferraresi).

E-mail: m.milicevic@fil.bg.ac.rs