Toward Sociolinguistic Corpora of Torlak

Towards sociolinguistic corpora of Torlak[1,2]

1. Introduction

The goal of this paper is to describe the process of creating two corpora aimed at documenting spoken Torlak with its core linguistic features, while also representing its geographic and sociolinguistic variability. At the time of writing, the two corpora–one capturing the Torlak vernaculars of Eastern Serbia, and one those of Western Bulgaria–differ markedly in size, as well as the approach to data transcription and processing. The paper initially describes both corpora, focusing in the second part on the Serbian *Spoken Torlak Dialect Corpus 1.0* and ongoing work on its 2.0 version. We start from the sampling, transcription and annotation of the data, and finish with the corpus publication and search options. Keeping in mind the complexity of spoken dialect corpora creation, we describe the conceptual and technical steps in corpus building and the specific challenges encountered at each level of data handling. Plans for a future integration of the two corpora into a single *Spoken Torlak Dialect Corpus* are also outlined.

While dialect corpora represent extremely valuable linguistic resources, compared to creating corpora comprising standard language written texts, collecting and processing oral data from dialect speakers is much more demanding. Firstly, taking a sociolinguistic approach, such corpora require a representative sample of speakers balanced across relevant demographic categories such as age, gender, educational level, or place of residence, which is often difficult to achieve due to population decline and/or dialect loss in some areas or age groups. Secondly, oral data must be audio and/or video recorded and then transcribed, which makes the process long and demanding in terms of organization and finances. Finally, once the data have been transcribed, processing the dialectal data has to be done at least partly manually, as the tools for automatic annotation tend to be available only for standard (written) varieties.

The corpora we describe contain data collected during fieldwork aimed at documenting the dialect and traditional culture of Eastern Serbia and Western Bulgaria, primarily within the projects *Guardians of the Intangible Heritage of the Timok Vernaculars* (financed by the Ministry of Culture and Information of the Republic of Serbia) and *(Dis-)entangling traditions in the Central Balkans: Performance and perception – TraCeBa* (see note 1 for details). Both projects had the goal of obtaining dialectologically relevant material, but with a focus on the actual, contemporary state of Torlak rather than an ideal described in

[1] The work reported in this paper was for the most part conducted within the project *(Dis-)entangling traditions in the Central Balkans: Performance and perception - TraCeBa*, funded by the Swiss National Science Foundation (grant agreement IZRPZ0_177557/1), the Russian Foundation for Basic Research (No. 18-512-76002) and the Serbian Ministry of Education, Science and Technological Development (grant agreement No. 401-00-00642/2018-09), within the FP7 ERA.Net RUS Plus programme. This support is gratefully acknowledged. We would also like to thank Andrey N. Sobolev and Barbara Sonnenhauser for their useful comments on the paper, and Tomaž Erjavec for his great help with publishing the corpus. More information can be found on the project website https://traceba.net.

[2] The authors' roles were as follows: M. Miličević Petrović: article writing and editing, coordination of normalization (Serbian); T. Vuković: corpus conception, data collection in Serbia, coordination of transcription (Timok), computational processing of Serbian data, article editing; M. Mirić: data collection in Serbia, coordination of transcription (Lužnica), article editing; D. Konior: data collection in Bulgaria, data transcription (Bulgaria), article editing; A. Escher: data collection in Bulgaria. We thank all informants who participated in the interviews and all colleagues who helped with different aspects of data collection, transcription or annotation.

the literature. Another shared objective was to investigate cultural practices and traditions in addition to language, with the *TraCeBa* project paying special attention to the role of boundaries (geographic, administrative, and others) in language production, and in the perception of the "other". It was thus important for the collected material to be suitable not only for the purposes of dialectological research, but also for the study of oral history, ethnography, and anthropological linguistics. At the same time, this multi-layered approach made it necessary to collect new data rather than relying on existing collections.

Both *TraCeBa* corpora consist of two subcorpora corresponding to the geographic areas from which the data were collected: Timok and Lužnica in Eastern Serbia, and Belogradčik and Tran in Western Bulgaria.[3] In addition, digitized materials from the *Sprachatlas Ostserbiens und Westbulgariens* (SAOSWB; Sobolev 1998; Sobolev, Gorlov 2021) are planned to be added as a separate corpus. On their own, these materials are not suitable for studying culture and traditions from the perspective adopted in the *TraCeBa* project, but they constitute a highly valuable dialectological source.

A number of extra-linguistic and methodological reasons contributed to the decision to create separate collections rather than a single corpus, at least for the time being: a) recent data from Eastern Serbia and Western Bulgaria and SAOSWB data were recorded in different time periods; b) even though open-ended interviews were always applied as a data collection method for the recent data, the methodology underwent certain modifications across areas, as the questionnaires were complemented with additional questions reflecting the researchers' needs and requirements (cf. Ćirković et al., this volume); c) the amount of recorded and/or transcribed material differs among the (sub)corpora (see below); d) the transcription was performed in different programs for Timok and Lužnica versus Belogradčik and Tran (see below); (e) at the time of writing, the data processing is at different stages for different corpora, and only some of them have become available in open access within the *TraCeBa* project.

The paper is organized as follows. In Section 2 the process of turning field recordings into text is presented, with a focus on data sampling and transcription. Section 3 deals with enriching text with annotation and metadata. Section 4 describes the process of publishing the corpus, with particular attention to the available search options. Finally, section 5 makes some concluding remarks.

2. From Field Recordings to Text

2.1. Data Sampling

[3] For practical reasons, the subcorpora are labelled Timok, Lužnica, Belogradčik and Tran. In geographic terms, Timok refers to the river Timok and its valley (with Knjaževac, Zaječar and Svrljig as the administrative centers), while Lužnica refers to the region around the town of Babušnica (and Bela Palanka to a certain extent). In these two regions, the data were collected in the rural areas, and to a lesser extent in some of the towns (e.g., Knjaževac). The two terms are also important because Timok and Lužnica vernaculars are part of the Timok-Lužnica dialect of the Prizren-Timok dialectal area of the Serbian language. On the other hand, Belogradčik and Tran are towns in Western Bulgaria, but the data were actually collected in the villages located in the areas surrounding the towns. Note that Torlak is also spoken in the northern part of North Macedonia and several other parts of Serbia; however, these areas are not included in the *TraCeBa* project.

The first step in creating the corpora required sampling the recordings according to the linguistic (dialectological), sociolinguistic and areal criteria defined by the *TraCeBa* project: a) the corpora should be relevant from a dialectological perspective and incorporate material representative of the Torlak vernaculars; b) the corpora should be relevant from a sociolinguistic point of view and include speech samples of older and younger, male and female native speakers of different educational backgrounds;[4] c) the corpora should contain the production of speakers living in rural and urban areas across the territory where Torlak is spoken.[5]

When it comes to the linguistic criteria, the presence of multiple dialectal features was taken into account when selecting the subset of data to be included in the corpora.[6] Since all of them were rarely attested in native speaker production, the sampling was based on the presence of certain core dialectal features at different levels of language structure. As for the Eastern Serbian vernaculars, the following set of features was applied:

1. At the phonetic-phonological level:
    a. stress
    b. the presence of the syllabic *l* (e.g., *slnce* 'sun')
    c. the presence of the semi-vowel schwa (e.g., *dən* 'day')
    d. the absence of the velar consonant *h* (e.g., *leb* 'bread')
2. At the morphosyntactic level:
    a. the analytic case system
    b. the postpositive article (i.e., postpositive demonstrative clitics)
    c. the analytic comparative with the prefix *po-* (e.g., *mlad* 'young' > *pomlad* 'younger')
    d. the absence of the infinitive form (and the exclusive use of *da* + present tense)[7]
    e. At the lexical level, the presence of distinctive vocabulary characteristic of the local Torlak vernaculars.

As for the Western Bulgarian vernaculars, the following features were considered central:

1. At the phonetic-phonological level:
    a. *u* as the reflex of the back nasal vowel
    b. schwa as the reflex of both *jer* semi-vowels (*ъ and *ь)
    c. *č* and *dž* as the reflexes of **tj* and **dj*
2. at the morphological level:
    a. inflection *-mo* in the first person plural present tense
    b. the first person personal pronoun *ja*

There was no specific number of features a speaker had to use in order to be included in the corpus; instead, the selection was based on the researchers' intuitions.[8] The task was overall very complex, as the informants often used a standard-like variety (close to either Standard

[4] Since the criteria under (a) and (b) are not always compatible with each other, in the particular case of Torlak precedence was given to criterion (a).

[5] The speakers gave oral consent to be recorded and for the (anonymized) data to be used for research purposes.

[6] For a detailed description of Torlak linguistic features in the dialects of Eastern Serbia and Western Bulgaria, see Sobolev et al., as well as Vuković et al. in this volume.

[7] Infinitives are often replaced by *da* + present tense in Standard Serbian as well, but they are not entirely absent.

[8] In the article by Sobolev et al., this volume, a broader set of dialectal features is discussed. The features used for selecting relevant recordings were chosen as the most prominent ones.

Serbian or Standard Bulgarian) in interviews led by researchers who were not native speakers of Torlak, or due to the gradual loss of the dialectal features in their idiolects.

Regarding sociolinguistic criteria, we aimed at selecting samples from dialectologically relevant speakers across different demographic categories, such as gender, age, educational level, and place of residence. However, it was impossible to obtain a large and balanced number of speakers in each of the desired groups. On the one hand, rural areas that are typically inhabited by the most representative dialect speakers are characterized by a population decline (see Šantić 2019 for data on the Timok area), which made it difficult to reach a sufficient number of informants. On the other hand, as younger speakers typically live in urban areas, attend school, and are widely exposed to media, their language is highly affected by the standard variety and often lacks even the core dialectal features (cf. Vuković et al., this volume). In addition, the tendency to use a standard-like variety characterizes even the older population from rural areas, especially in Western Bulgaria. When it comes to the Timok corpus, it includes 168 speakers, of whom 101 are women and 67 are men. For older speakers, information about age is available for 54 speakers and is inferred for an additional 15 speakers;[9] while most speakers are over 60, the corpus also includes 11 young speakers, high-school students aged 17 and 18, who were included despite speaking a more standard-like variety to allow for the study of language change across generations. Overall, the corpus is not fully representative in terms of age and gender, as there are more speech samples from female than male speakers, and more from older than younger speakers.

In order to adhere to the areal criteria and cover the vast territory where Torlak is spoken, we aimed at sampling the recordings from various locations. Where possible, we made sure to include data from locations evenly dispersed across the area, in order to represent the whole area and to allow the analysis of potentially relevant geographic factors, such as longitude, latitude and altitude, as well as spatial distances between villages and local town centers.[10] However, an ideal balance could not be achieved given that more samples were available from rural than from urban areas.

Since there were challenges in meeting each of the initial requirements, precedence was generally given to the dialectological criteria. The main information on the samples that were finally chosen (and transcribed) is summarized in Table 1; the problems associated with data collection and dealing with the recordings in the framework of the *TraCeBa* project are discussed in more detail in Ćirković et al. (this volume).

	Timok area (Eastern Serbia)	**Lužnica area (Eastern Serbia)**	**Belogradčik area (Western Bulgaria)**	**Tran area (Western Bulgaria)**
Fieldwork year(s)	2015, 2016, 2017, 2018	2018, 2019	2018	2019
Project(s)[11]	1, 2	2, 3	3	3

[9] The absence of information on age is due to the fact that the data were collected by multiple researchers, some of whom came from an anthropological tradition and considered it unethical to ask about age.

[10] This applies in particular to the Timok and Lužnica regions.

[11] 1: *Guardians of the Intangible Heritage of the Timok Vernaculars*, financed by the Ministry of Culture and Information of the Republic of Serbia; 2: *Language, Folklore and Migrations in the Balkans*, financed by the

N of locations included in the corpus	64[12]	15[13]	5[14]	5[15]
N of transcribed hours	80	26	15.5	4.5
N of transcribed tokens	498,021	193,991	47,123	13,033

Table 1. An overview of the samples used.

In addition, transcripts from SAOSWB (Sobolev 1998; Sobolev and Gorlov 2021), covering 37 locations in Serbia and Bulgaria with 133,260 tokens, have been digitized and we plan on adding them as a separate corpus in the future.

2.2. Transcription

2.2.1. Timok and Lužnica (Eastern Serbia)

The selected speech samples from Timok and Lužnica were transcribed in the *Partitur-Editor* program from the software package *EXMARaLDA* (Schmidt 2009)[16], following the guidelines provided for transcribers (Vuković 2016). Each recording was uploaded to the software separately and thus associated with its transcript; information about the location, the speakers, the transcriber, etc., was then added to each transcript before the actual transcription began (see section 3.1 for details). Accurately reflecting all of the relevant oral and dialectal features was predicted to be a major challenge, and the transcribers were selected accordingly: they were associates external to the project, familiar with Torlak, who received training instructions by one of the *TraCeBa* team members, as well as a detailed sheet of instructions (Vuković 2016). The team members then verified the transcriptions; the problematic words or passages are marked and explained in a separate tier in *EXMARaLDA*.

Given that multiple speakers participated in each recording, the text was initially divided by entering the production of each individual speaker in a separate tier. The utterances of 15 researchers who participated in the data collection were also included in the transcripts (they constitute 10-15% of the text); the researchers' production was marked with RS_ plus their initials under a dedicated metadata attribute (see Table 2 below), so they may be omitted

Ministry of Education, Science and Technological Development of the Republic of Serbia (project No. 178010); 3: *TraCeBa* (see note 1 for details).

[12] Timok locations: Aldinac, Balanovac, Balinac, Balta Berilovac, Borovac, Buje, Bulinovac, Crni Vrh, Ćuštica, Debelica, Donja Kamenica, Drečinovac, Drenovac, Drvnik, Gabrovnica, Glogovac, Gornja Sokolovica, Gornje Zuniče, Gradište, Grlište, Guševac, Inovo, Jakovac, Jalovik Izvor, Janja, Jelašnica, Kandalica, Knjaževac, Koželj, Krenta, Lepena, Leskovac, Lokva, Mali Izvor, Marinovac, Minićevo, Novo Korito, Orešac, Ošljane, Petruša, Ponor, Pričevac, Radičevac, Ravno Bučje, Repušnica, Rgošte, Sastavci, Selačka, Stara Kalna, Staro Korito, Šarbanovac, Šesti Grabar, Štipina, Tijovac, Trgovište, Trnovac, Vasilj, Vlahovo, Vratarnica, Vrbica, Zorunovac, Žlne, Žukovac.

[13] Lužnica locations: Bratiševac, Dol, Draginac, Dučevac, Gorčinci, Gornji Striževac, Grnčar, Izvor, Kambelevac, Preseka, Radosin, Resnik, Studena, Šljivovik and Veliko Bonjince.

[14] Belogradčik locations: Čuprene, Stakevci, Ošane, Preužda and Gorni Lom.

[15] Tran locations: Velinovo, Jarlovci, Erul, Vidrar and Goročevci.

[16] https://exmaralda.org/en/

from corpus searches. Within each tier, the text was further segmented into utterances that form intonational, structural and meaningful units, whose duration usually varied between one and four seconds. Sentence structure was not explicitly marked by punctuation or other orthographic means. The only explicit structural division was created using the "event" units, based on time intervals (see Vuković 2021). Figure 1 illustrates a recording segment accompanied by a transcript.

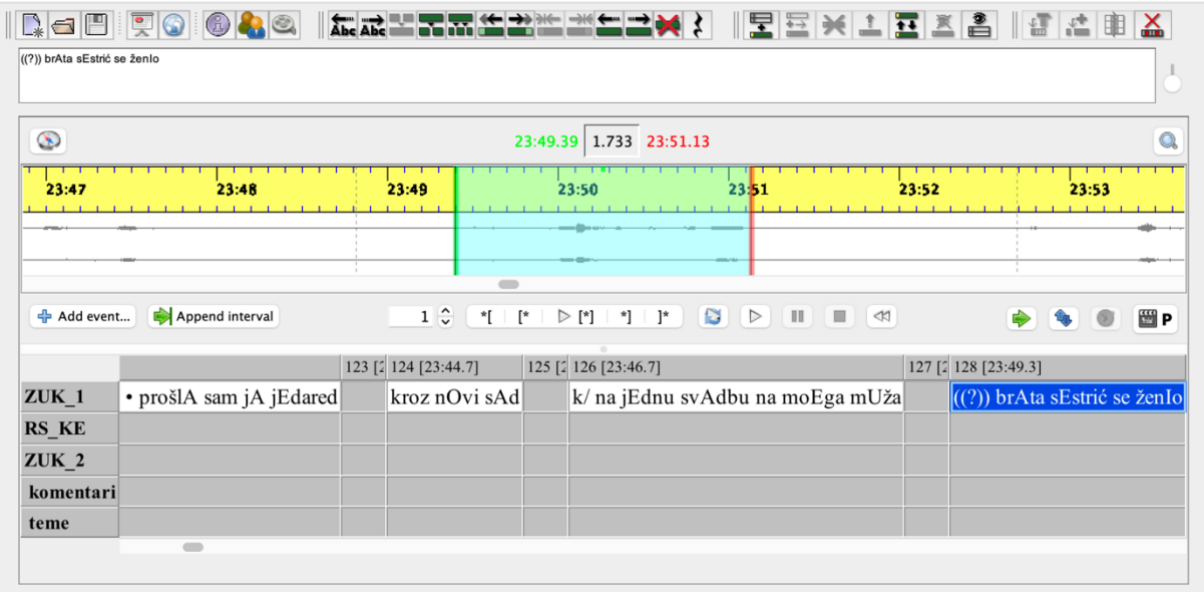

Figure 1. Transcript segment in *EXMARaLDA*.

The transcription was semi-phonetic, i.e., the texts were written in the standard Serbian Latin script (with the addition of some special symbols; see below), but the phonological variability and features characteristic of the spoken language and the Timok and Lužnica vernaculars were preserved (e.g., *će a trAžiš tI tEa tvOe nOǵe* instead of *će **D**a trAžiš tI tE**J**a tvO**J**e nOǵe* 'you will look for those legs of yours'); a similar approach is adopted in other dialect corpora, such as the English FRED corpus (Anderwald, Wagner 2007). Although we aimed at transcribing the morpho-phonetic content accurately (covering even allophones in the speech of the same speaker), random and rare pronunciation errors and related phenomena were normalized in the transcripts, e.g., *gotina* was normalized to *godina* 'year', or *br((cough))at* to *brat ((cough))*.

The position of the accent was marked using upper case letters (e.g., *iz ovOj sElo* 'from this village'). Since Timok and Lužnica vernaculars differ from Standard Serbian at the phonological level, the transcription required introducing additional characters, such as *ə* for the semi-vowel schwa, *ǵ* for the palatalized *k*, *ƶ* for the voiced affricate *dz*. Additional symbols were also used for elements characteristic of spoken language (• for a pause, # for unintelligible parts of utterances, double letters for prolonged sounds, and / for interrupted parts of words or utterances). Non-transcribed parts of a recording were marked with NT, usually consisting of segments that contained personal information. Non-verbal elements with a clear communicative function (e.g., laughter) were also marked (e.g., ((laugh)), ((cough))).[17]

[17] For more details on transcription, see Vuković 2021.

A sample from a transcript from the Lužnica area is given below (Example 1, a female speaker aged 84).

(1) A transcription example from Eastern Serbia:

a mUž mi jOš nEe vOjsku služIl On u sedamnAesetu godIn se oženIl • mlAd
pa nEe
punolEtan ne mOž se vEnčamo
ne mOž da Idemo da se regIstruemo
i detE će bUde vAnbračno pa Idemo Oba pa ga prijAvimo • da je nAše
i kAd On d Ide u vOjsku onO Odeše
i pUno sOba zbrAše se tUva ispRćaj
mu praImo ispratImo ga dOle do • cEr
Imaše jOš vojnIci
kOlo • čEkaju vOjna kOla da dOjdu • nEma kOla
jA vIše da Idem u babUšnicu
jA detE ostavI pri bAbu
lEle ivAne jA mOra se vRtem detE swm ostaIla • Ajde Ajde Ajde Ajde
i otIdem
u babUšnicu
otIdem • dOle • dE sAmo jA i • pUno cElo rodA mUzika

The output of the transcription process is saved as an XML document, which stores the information on the transcript text, the utterances segmentation and the time intervals.

2.2.2. The Belogradčik and Tran Areas (Western Bulgaria)

Selected samples from locations in the western area of the town of Belogradčik were transcribed in *Microsoft Word*.[18] The members of the Russian *TraCeBa* team both transcribed the data and revised the transcripts themselves, taking turns in these tasks based on their knowledge of the languages and varieties involved (see Ćirković et al., this volume for more detail).

The recordings were transcribed in the form of a narrative, explicitly segmenting the text into utterances that follow the syntactic structure and prosodic elements (based on the researchers' intuition), which required the use of punctuation and other orthographic means typical of written texts (e.g., upper case for proper nouns and sentence beginnings). The researchers' production was also transcribed, and explicitly marked with the researchers' initials, similarly to what was done for the materials from Eastern Serbia (researcher production accounts for 5-10% of the total transcribed text).

The texts were transcribed using the Latin script, based on the transcription rules of the *Slavic Linguistic Atlas* (OLA 1994). In addition to the standard Latin characters, symbols from the Serbian Latin alphabet (*č*, *đ*, *dž, lj, nj*) were also introduced, as they reflect the phonetic features of the Bulgarian western border dialects quite accurately. The sonants *r̥*, *ŕ*, *l̥*, *ĺ* are marked by the corresponding IPA symbols. The Bulgarian medium vowel is symbolized by *ə*, and the palatalization of the consonants is demonstrated by an apostrophe (e.g., *Polaz'uvan'e* 'a ritual on St. Ignatius' day'). Some additional symbols are used where

[18] Due to having less experience with *EXMARaLDA* than the Serbian team, the Bulgarian team decided to first transcribe the data in a more familiar format, and adapt it later on.

necessary (superscript for coarticulation (e.g., *o*^{u}), and double letters for prolonged vowels (e.g., *bl'aagə* 'sweet, tasty'). The accent is marked by a straight vertical line before the corresponding vowel (e.g., *bašt'a* 'father'). The researchers' questions and comments, as well as the non-transcribed parts of the recordings are enclosed in square brackets.

Each transcript is preceded by information on the location, transcript number, the researchers' initials and the date. Additionally, when possible, the following information about the respective recording was provided in a separate .doc file: the informant's name and age, the topics discussed, the characteristics of the informant's language (dialect, standard-like variety, standard language), and the characteristics of the interview (mostly monologues, mostly dialogues, reading of regional folklore pieces, song singing, etc.).[19] A sample transcript segment from the village of Čuprene is provided below.

(2) A transcription example from Western Bulgaria.

Čuprene_20_AS, DK, AM_.28.07.18

15:29 N'ali ti kəz'uem, na Sv'eti Nik'olu, Svet'əc, na n'as, na 'ej t'am na dr'ugi čov'eci. 'I u n'ašija kr'aj te tov'a, r'iba, dok'aru, nal'i, tə se k'upi št'o t'u u St'akevci 'ili pa u rek'ata no edn'o vr'eme 'imaše r'iba mr'ena, mr'ena r'ipka, a tak'a, e te tək'ivə r'ipke bl'aagə bl'agə. Pa na m'ajka bašt'a i 'imaše vadan'ica, det se m'el'a ž'itoto i na j'azət zagr'adi ta vod'ata da ne m'ože d 'ide, a on'e d'ol'e se zb'eru r'ibete št'o 'ima od br'ašno [od tr'ice pad] i d'ojde 'on səs vr'eču s čuv'al zgreb'e i t'e t'i, d'aže sm'o niz'ali t'a is'ənu r'ibete. Ah'a. A 'inače za Sve'ti Nik'ola te tək'a, nal'i p'osno, 'ama r'iba se razreš'ava da se ed'e, nal'i na Sve'ti Nik'o:<lə>. I smo p'ekli. B'aba e prə'ila i r'ibenu čorb'icu na Sve'ti Nik'olə. 'Inače dr'ugo j'a če ne sm'o prəznuv'ali, Svet'əc n'ije bil'o na Sv'eti Nik'olə. Šar'ani təkv'oj ne sm'o p'l̥nili, nal'i ne 'e im'alo tejg'a, i p'osle v'eče se, nəl'i gov'orime za st'ari r'aboti, [...] a s'əga v'eč e s'ičko mod'erno, nal'i, n'e. [A. S. D'a. 'A z'a pol'aznikət, na k'oj d'en?..] Pol'aznikət e te tov'a na Ign'ažden, Ign'ažden! [A. S. A tov'a e kniž'ovno, Ign'ažden...] Da, Ign'atovden 'ili Ign'ažden, a m'i mu d'umamou Poləz'uvan'e, Pol'aznica. [...] Mɾ̥snite dn'i, p'ak tək'a, n'ištou se ne rab'oti, sv'adbe se ne pr'aju, n'ema god'ež, n'ali da se izgod''avu… [A. S. I v'elite god'eš?..] God'eš, zn'aeš kv'o zn'ači god'eš? [A. S. D'a d'a d'a.] Tək'a, god'eš, t'i če se ž'eniš, t'o ješ če 'ideš pri m'uš, sv'adba če pr'aiš. A pred'i da 'ideš da sv'adbə, d'ojde bašt'a i m'ajka na momč'e u d'omət na dev'ojčeto, na mom'ičeto. **17:57**

2.2.3. *Sprachatlas Ostserbiens und Westbulgariens* (SAOSWB)

SAOSWB (Sobolev 1998; Sobolev, Gorlov 2021) contains original texts collected between 1989 and 1995. The transcription of dialectal forms indicates the position of the accent with diacritics and uses special characters for phonemes or phonetic features that do not exist in the alphabets of standard Serbian and Bulgarian (e.g., *ə* in *dən* 'day', *l̥* in *bl̥'a* 'flea'). Allophones are also accurately marked, as they reflect internal variation in idiolects (e.g., *dan* / *də^{a}n* / *da^{ə}n* / *dən*). Researcher utterances are not included in SAOSWB.

In the process of digitizing SAOSWB, the original texts were firstly automatically converted into .txt format using the OCR program *Transcribus*,[20] which preserved all of the original phonetic symbols. For this conversion, a custom model was trained on 50 manually

[19] The Serbian team also saved separate *Microsoft Word* files with the informant's details, information on the recording location, date and settings, the researcher's name, the topics, the characteristics of the informant's language (dialect, standard-like variety, standard language), etc. These field protocols were used as a source of metadata entered in EXMARaLDA for each recording (see section 3.1).

[20] https://transkribus.eu/Transkribus/

transcribed pages of text. However, due to the levels of phonetic-phonological detail and variation being higher in SAOSWB than in the transcripts of recent field recordings, the transcription was simplified following the model adopted for the transcription of recent recordings, by using a custom Python script: for instance, the accented characters were initially replaced with upper case characters (the sentence structure was marked with punctuation, and proper names with a special character), and complex sounds were replaced by characters from the standard Serbian alphabet (e.g., *dž* instead of *ǯ*).[21]

An example from the *Transcribus* output and its simplified version are shown in (3) (taken from Makarova et al. 2020). The texts were exported in .txt format, and there are plans—outside the *TraCeBa* project—to convert them into TEI XML, after manual checks and corrections.

(3) Original and simplified versions of a text segment from SAOSWB

Transcribus output	**Simplified version**
I túj sam zapazíla túj gdé su pravíli kačámak, iskopáše bápku, pa nakládoše ógań, pa turíše kotə́l i napraíše kačámak, i túj mí večéramo túj. Í tiќe jútrom káže náž déda táj Pétar káže: “Déca, vrаćamo se, - káže. - Vójska ostupíla, mí se vrаćamo, - káže, - i ídemo, - káže, - kúći”. I pójdemo mí otúdaᵊkaj, i Kńáževac, u Kńaževaʒ, gorí, gorú ógńi, ógań gorí, ovdé gorí, ondéka. Onó Búgari kaᵊkó próšli, oní palíli tój. Palíli ovója dućán se zoválo, dućáni. Dućáni palíli.	I túj sam zapazíla túj gdé su pravíli kačámak, iskopáše bápku, pa nakládoše óganj, pa turíše kotə́l i napraíše kačámak, i túj mí večéramo túj. Í tiќe jútrom káže náž déda táj pétar káže: “Déca, vrаćamo se, - káže. - Vójska ostupíla, mí se vrаćamo, - káže, - i ídemo, - káže, - kúći”. I pójdemo mí otúdakaj, i Knjáževac, u Knjaževadz, gorí, gorú ógnji, óganj gorí, ovdé gorí, ondéka. onó Búgari kakó próšli, oní palíli tój. Palíli ovója dućán se zoválo, dućáni. Dućáni palíli.

3. Enriching text with annotation

Even though it is possible to create a spoken language corpus consisting only of transcripts, it is much more typical and more useful to enrich corpora with metadata and linguistic annotation so as to enable advanced search options. In the Serbian *TraCeBa* corpus, metadata concerning recordings and speakers are encoded, as are several types of linguistic annotation (part-of-speech (PoS) tags, and lemmas). Since the annotation is still ongoing at the time of writing, the present overview primarily captures the completed 1.0 version of the Timok subcorpus. The transcripts from Lužnica are expected to be annotated and added in the coming months, as a new subcorpus in the version 2.0 of the already published corpus. The Bulgarian corpus and the digitized SAOSWB data are currently independent entities that are planned be added to the Torlak corpus in future releases, contributing to a joint *Spoken Torlak Dialect Corpus*, but this will not be achieved within the *TraCeBa* project.

3.1. Metadata

[21] A similar procedure is likely to be used when uniting the current corpora in a single large corpus. We will also consider the possibility of aligning and including both the original and the simplified transcripts, enabling searches in the simplified version while also making the richer texts available.

In accordance with the spoken status of the data, and the sociolinguistic status of the corpus, the Serbian corpus encodes relevant contextual information about the recordings (i.e., transcripts) and about the speakers, so that both the context in which the speech event took place, and the characteristics of the speakers can be analyzed, providing a "situated speech" perspective typical of the ethnography of communication (see Bernardini et al. 2018, 27–28).

Each transcript, saved under a unique ID, contains information about the location the recording represents, the duration of the recording, the size of the transcript, and the speakers involved. The attributes used and a description of their values are shown in Table 2.

Attribute	Values
ID	a unique ID for each recording-transcript pair, in the format TOR_C_xxxx
LOCATION	the location the recording represents (place name)
LONGITUDE	location longitude, obtained from Google Maps (according to the WGS84 Mercator projection)
LATITUDE	location latitude, obtained from Google Maps (according to the WGS84 Mercator projection)
TRANSCRIBER	the full name of the transcriber
REC_DURATION	the duration of the source recording in the hh:mm:ss format
TOKENS	the number of tokens in the transcript
UTTERANCES	the number of utterances in the transcript
RESEARCHERS	the labels of the researchers conducting the interview
SPEAKERS	the ID of the speakers being interviewed (the same as in the TOR_C_speaker file)

Table 2. Attributes and values assigned to the individual recordings.

Each speaker is encoded in the corpus with an original ID in order to conceal their identity, and is associated with information about relevant extra-linguistic factors. The speaker metadata files contain the categories shown in Table 3. Due to the different methodologies of data collection depending on the researchers' backgrounds, metadata is not available for all speakers, and in some cases some specific pieces of information are lacking, sometimes pertaining to age and often related to education. In some cases, age could be guessed based on the speaker's appearance or from the narrative; these approximate numbers are marked with a tilde (e.g., ~60).

Attribute	Values
ID	a unique ID as displayed in the corpus, concealing information about the speaker, in the format TIM_SPK_xxxx
AGE	the age of the speaker at the time of recording
GENDER	the gender of the speaker
YEAR_OF_BIRTH	the year of birth of the speaker

EDUCATION	the education of the speaker
LOCATION	the location where the interview was recorded
ORIGIN	the birthplace of the speaker (if null, the same as the LOCATION)
OCCUPATION	the occupation of the speaker
DIALECT_Y_N	whether the speaker is a typical representative of the dialect (y/n)[22]
TRANSCRIPT	the IDs of the transcripts in which the speaker participates

Table 3. Attributes and values assigned to the speakers.

3.2. Linguistic Annotation

Two kinds of linguistic information are currently encoded in the Timok subcorpus (and will soon be added to the Lužnica subcorpus): PoS-tags (more specifically, MSDs, i.e., morphosyntactic descriptions containing information on parts of speech and relevant grammatical categories such as person or gender) and lemmas (base forms of words). A third layer whose addition is planned as a possible extension is normalization, i.e., information about the standard language forms that correspond to each of the corpus tokens.[23]

PoS-tagging and lemmatization are typically performed using automatic tools. However, the situation is less straightforward with dialect data. Among the dialects of Serbian, we can safely deem Prizren-Timok to be the most distant from the standard when it comes to the comparison of their grammatical structures. For this reason, existing tools for Standard Serbo-Croatian or Bulgarian could not be used without adaptation.

The starting point for the annotation tools were those available for Standard Serbian, specifically the ReLDI tagger (Ljubešić et al. 2016)[24]; the new tools built for Timok are an adaptation of this tool. The adaptation relied on a semi-automatically developed training set, built by using the ReLDI tagger and following up with manual verification and correction.[25] The training set consisted of dialect data combined with Standard Serbian data. The motivation behind using a combined dataset lay in the large vocabulary overlap between Prizren-Timok and Standard Serbian; in other words, good quality training material for Standard Serbian was seen as helpful for the annotation of the dialect data. The dialect portion of the training set amounted to 26,909 tokens, and it included samples from both the contemporary Timok materials (88%) and the SAOSWB (12%). The Standard Serbian portion was composed of 88,546 tokens from the manually verified *SETimes.SR* reference training corpus (Batanović et al. 2018).

The combined data were used to train the tagger and the lemmatizer for Torlak, resulting in 84.61% accuracy for the former and 92.62% for the latter (compared to only 68%

[22] Most speakers included in the corpus are representative of the dialect.

[23] Note that we did not adopt an approach in which normalization is performed first and used as input for PoS-tagging and lemmatization, as this approach would have required too much time allocated to normalization. It was instead decided to train the standard language tools on dialect data. Normalization of a smaller sample has been conducted with the goal of automating the process and facilitating corpus searches in the future.

[24] https://github.com/clarinsi/reldi-tagger

[25] The Torlak tagger, as well as the tagging and lemmatization files and instructions, can be downloaded from https://github.com/bravethea/Torlak-ReLDI-Tagger-2019.

for tagging and 77.5% for lemmatization obtained with the original version of the ReLDI tagger; see Vuković 2021).[26] The accuracy was measured on a test sample of 2,000 tokens (two transcripts) selected from the recent Timok data; this sample was annotated both manually and using the adapted tagger.[27] On average, between 7 and 15 out of 100 tokens were not tagged or lemmatized correctly when using the automatic tool. As expected, the errors mostly concerned dialect forms: for instance, pronouns were assigned a range of incorrect PoS tags (they were recognized as nouns, adjectives and verbs), and past participles ending in *-l* (instead of the standard language *-o*, e.g., *bil* instead of *bio*) were often labelled as adjectives. There were also issues with forms with syllabic *l*, *dž* (from **dj*, *dz*), forms without *h*, forms with new palatalizations (e.g., *devojća* 'girl'), and morphologically distinct endings. Additional problems were caused by spoken elements such as phoneme elisions, or truncated words (for more detail, see Vuković 2021). We expect the tagger to reach similar accuracy rates with the data from Lužnica, following a re-training phase on a more extensive training set that contains a manually annotated Lužnica sample. It should also be noted that more extensive accuracy evaluations are planned for once the Serbian corpus is complete; the current figures should primarily be taken as an indicator of initial training-induced improvements.

The PoS-tagging, i.e., the morphosyntactic annotation of the data, is based on the MULTEXT-East conventions; it mostly follows the specifications for Serbo-Croatian (Erjavec 2019), but due to the differences outlined below, it has been coded using dedicated Torlak specifications (Vuković 2020). An example of a tagged utterance is provided in Table 4; each character position in the tags denotes a grammatical category. For example, *Pp3msn* stands for the attribute-value pairs of Category = *P*ronoun, Type = *p*ersonal, Person = *3*, Gender = *m*asculine, Number = *s*ingular, Case = *n*ominative. The labels used are thus not just PoS tags, but full morphosyntactic descriptions (MSDs).

Text	Lemma	Tag	Translation
on	on	Pp3msn	he
toj	taj	Pd-nsn	that
videl	videti	Vmp-sm	saw
tam	tamo	Rgp	there
na	na	Sa	on
bugarsko	bugarski	Agpnsay	Bulgarian
prelazilo	prelaziti	Vmp-sn	crossed
se	sebe	Px---a	REFL

[26] Tagging and lemmatization were performed disregarding accent information in order to reduce variation and increase accuracy.

[27] The evaluation data are available at: https://tinyurl.com/eur2xmnz.

tam	tamo	Rgp	there
.	.	Z	.

Table 4. Example of an annotated utterance.

Several differences between Standard Serbian and Torlak made it necessary to adapt the original Serbo-Croatian tagset. Most importantly, the semantic category of definiteness is not grammaticalized in Standard Serbian, while in Torlak morphemes resembling the Bulgarian post-positive articles are added to nouns, adjectives, pronouns and numerals (e.g., *at* in *unukat* 'the grandson'). In order to annotate such morphemes, a new definiteness category was added to the standard tags, with three values: *v* for the speaker-proximal form, *t* for the hearer-proximal form and *n* for the distal form.[28] For example, *unukat* is annotated as *Ncmsn-t*: Noun, common, masculine, singular, nominative, animacy not marked, *t*-definite; no value for this category is assigned to tokens that do not contain a definiteness marker.

Other PoS-tagging decisions were related to some of the simplifications that have occurred in the Torlak grammatical system over time. Specifically, due to the process of morphological case loss, the sole oblique case, accusative in origin (e.g., *majku*.ACC.SG.F 'mother') performs all oblique case functions, including those expressed with other case forms in Standard Serbian (e.g., *sas*/*sa majku*.OBL.SG.F vs. Standard Serbian: *sa majkom*.INS.SG.F 'with mother'). Since the specific function of the oblique case endings is often difficult to detect automatically (many prepositions can take different cases), feminine nouns ending in *-u* are always annotated, on purely morphological grounds, as accusative. Masculine animate nouns, ending in *-a* in the oblique case, correspond to the homonymous genitive and accusative forms in Standard Serbian. In the Timok subcorpus, the oblique case for such nouns is annotated as genitive if the noun is preceded by a preposition requiring the genitive in Standard Serbian, e.g., *od čoveka.*GEN.SG.M 'of a man', and as accusative in other cases. The same applies to the endings of adjectives and pronouns, e.g., genitive in *od mojega čoveka* 'of my man'. Note that Standard Serbian case forms are also often used by speakers of Torlak (see Vuković et al, this volume), which was the main reason for keeping the standard case tags rather than introducing a new generic tag for "oblique" in the Torlak tagset.[29]

In addition to the above, since the corpus contains notations referring to the spoken language elements that are not found in the original tagset, the category *X* is also used, as illustrated in Table 5. This label is visible in the text output of the tagger, but it is not displayed in the TEI XML version of the corpus (see section 4.1), where it is replaced with the standardized markings such as <*pause*/>.

word	**lemma**	**tag**	**Translation**
kad	kad	Cs	when

[28] The post-positive demonstrative clitics are derived from demonstratives and signify the three-way distance reference: the speaker-proximal *ovəj/ovija* 'this', the hearer-proximal *təj/tija* 'that', and the distal form *onəj/onija* 'that over there'.

[29] In other words, while this decision does not fully capture Torlak as described in traditional dialectology, it appears more suitable for present-day vernaculars, which are heavily influenced by the standard language.

((?))	((?))	X	((?))
sečémo	seći	Vmr1p	cut
koláč	kolač	Ncmsan	cake

Table 5. An example including an additional tag for spoken language elements.

As for lemmatization, the main problem arose from the fact that Torlak does not have an infinitive form for verbs. Even though the third person singular present is used as the word form representing the lemma in other infinitive-less South Slavic languages (e.g., Macedonian), for the Timok sample it was deemed more appropriate to keep the infinitive as the lemma, for the same reasons outlined above (related to the frequent "mixed" use of Torlak and Standard Serbian), and in order to make the most of the available resources for Standard Serbian. The same decision was adopted even for dialectal verbs such as *oratim* 'I speak', for which an infinitive form *oratiti* was constructed, following the rules of Serbian verbal morphology. This decision will be highlighted in the instructions for corpus use, and it will be accompanied by examples.

Another issue concerned the use of semi-phonetic transcription, which led to inter-speaker and areal phonetic variability being explicitly shown in the data, making a single word seem like several different words. For instance, several variants of the lexeme *san* 'sleep/dream' can be encountered in the recordings: [*sъn*], [*son*], [*san*], [*saᵇn*], and [*sъᵃn*], for which three variants of transcription were used: *sən*, *son*, and *san*. All of them are assigned the lemma *san*, thus encompassing the variation and making it easier for future corpus users to find the different phonetic realizations (see section 4.2).[30]

Finally, by virtue of the transcription methodology, the corpus texts contained information about the position of the word accent. It was marked manually by using capital letters (e.g., *ovdE* 'here'). At the corpus processing stage, the capitals were replaced with accented letters (*ovdé*); section 4.2 provides information on how accented letters can be used in corpus searches.

As concerns a possible layer of normalization, the presence of such a layer would enable corpus users to access the dialect content through a more familiar variety (Standard Serbian in the case of the Timok subcorpus), as normalization can be understood as translation from the dialect into the standard language, and it can simplify searches in cases where multiple variants of a word occur. In the specific case of the *TraCeBa* corpus, only morphological normalization is being implemented, i.e., the normalization of individual word forms (e.g., from the dialectal *pituval*, to the standard equivalent *pitao* 'asked', or from the dialectal *tuj* to the standard *tu* 'here/there'). An attempt at automatic normalization has already been made, using the machine translation method enabled by the ReLDI tagger. A set of ca. 21,000 tokens were normalized manually and used as a training set. However, given the novelty of the task, this training set was not large enough for the automatic tool to reach an acceptable accuracy level. Manual normalization of a further 20,000-token data set is currently being finalized (the disagreements between two annotators are being resolved), and

[30] This was the approach adopted in the manual part of the annotation. It is possible that the decision was not always implemented correctly by the automatic tagger.

the machine translation method will be re-tested once this set is complete. Insights from historical computational linguistics will be used to improve the procedure.

4. Publishing the Corpus

4.1. Corpus Building

In order to be published as a full-fledged digital corpus, the transcripts need to be indexed and made searchable, in addition to being processed through metadata coding, morphosyntactic tagging, and lemmatization. This process of corpus building requires the data to be in appropriate formats to be uploaded to a corpus management platform. In the case of the *TraCeBa* corpus, after all the necessary tools were developed and the transcripts were annotated, the corpus was made searchable and thus prepared for publication in the CLARIN.SI repository,[31] the Slovene branch of the European CLARIN ERIC infrastructure. Corpora stored in CLARIN.SI are also integrated into the *NoSketch Engine* platform (*NoSkE*, Rychlý 2007) and the *KonText* concordancer (Machálek, Křen 2013; Machálek 2020).[32] The corpus in XML/TEI and derived verticalized format, as well as the concordancer access link are available at http://hdl.handle.net/11356/1281.[33]

The *TraCeBa* corpus file(s) were initially encoded in XML (using the TEI standard). In this format, the individual corpus files start with a header containing information about the transcript (project information, details about the media behind the transcript, the speakers in the transcript etc.),[34] a timeline element with the time intervals, and the corpus text. Figure 2 shows an example: inside the <u> element, the <w> element is used for individual words, and this is where the annotation information is kept (lemma and tag (marked as 'ana', and with an 'mte' prefix for MULTEXT-East specifications)); <pause> is used for silent pauses, <gap> for text that was unintelligible or omitted for other reasons (such as personal data protection); additional attributes include <vocal> for laughter, cough, filled pauses and other human non-verbal sounds, and <incident> for non-human sounds. Information about the speaker is also encoded, using the 'who' attribute (initials are used for the researchers conducting the interview, and dedicated IDs for dialect speakers).

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
  [...]
  </teiHeader>
<text>
   <body>
      <timeline unit="s" origin="#TOR_C_0001.T0" corresp="#TOR_C_0001.wav">
         <when xml:id="TOR_C_0001.T0" interval="645.825" since="#TOR_C_0001.T0"/>
         <when xml:id="TOR_C_0001.T1" interval="646.165" since="#TOR_C_0001.T0"/>
         […]
      </timeline>
  <u xml:id="TOR_C_0001-u28" start="TOR_C_0001.T657" end="TOR_C_0001.T658" who="#TIM_SPK_0001">
      [...]
```

[31] http://www.clarin.si/

[32] https://nlp.fi.muni.cz/trac/noske, https://github.com/czcorpus/kontext

[33] Remember that the current text refers to the Timok subcorpus.

[34] The specifications used in the headers can be seen at https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html.

```
    <unclear reason="unintelligible"/>
    <pause xml:id="TOR_C_0001-u28-w11"/>
    <w xml:id="TOR_C_0001-u28-w12" lemma="e" ana="mte:I">é</w>
    <w xml:id="TOR_C_0001-u28-w13" lemma="pa" ana="mte:Cc">pa</w>
    <w xml:id="TOR_C_0001-u28-w14" lemma="tu" ana="mte:Rgp">túj</w>
    <w xml:id="TOR_C_0001-u28-w15" lemma="sedeti" ana="mte:Vmp-sm">sédite</w>
    <vocal xml:id="TOR_C_0001-u522-w1" type="laugh"/>
    <incident xml:id="TOR_C_0015-u51-w1" type="noise"/>
   </u>
</body>
 </text>
 </TEI>
```

Figure 2. An excerpt from the *TraCeBa* XML file.

The final corpus file, derived from the XML document(s), is in the vertical format typically required by corpus processing platforms. It contains all the transcripts with annotation and a minimal XML mark-up. This format is optimized for indexing and consultation with *NoSkE*.

The corpus, together with the tools and training sets, is published under the Creative Commons Attribution-NonCommercial 4.0 licence. Audio files in WAV format, in which the untranscribed sections have been masked, are published separately (in the CLARIN.SI repository), under the same licence. At the moment, the corpus text and the recordings cannot be searched in parallel using the corpus search platform, but they can be viewed using the *EXMARaLDA* software.

4.2. Search Options

NoSkE is a very friendly environment for end-users, with an intuitive graphical interface that makes it possible to exploit most of the corpus metadata and to carry out metadata-based queries. It also handles different layers of linguistic annotation well, enabling numerous search options, including complex ones, as well as the possibility of subcorpora creation. It implements several types of search, most of which include regular expressions. The most useful option is the use of *Corpus Query Language* (*CQL*; Jakubíček, Kilgarriff, McCarthy, Rychlý 2010; Evert et al. 2019). The *NoSkE* search interface is shown in Figure 3.

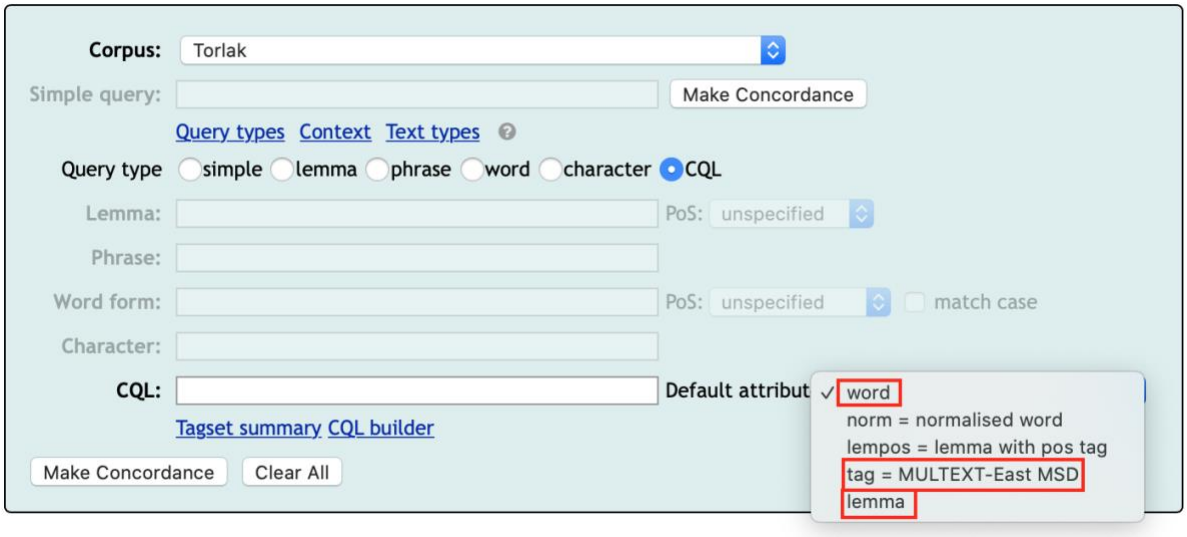

Figure 3. The *NoSketch Engine* search interface.

The attributes marked in red in the bottom right corner (*word, tag, lemma*), concern the information shown in the different columns in Table 5 above: *word* refers to information from the first column (the original text), *tag* refers to the morphosyntactic descriptions, and *lemma* to the base forms. After normalization is performed, *norm* will refer to the column containing normalized forms. It is important to note that all information is simultaneously present in the corpus and can be used individually or combined. Figure 4 illustrates the results of a CQL query using the *lemma* attribute, [lemma="doći"].

Figure 4. Example results of a CQL query.

Among other options, by combining the *word* and *tag* attributes, it is possible to disambiguate homophones (e.g., for the multifunctional pronoun *mi* 'we/to me'). By using the *norm* attribute, it will be possible to obtain standard and dialect forms in a single query (e.g., for *bio* 'was' the result would be the forms *bio* and *bil*). This will make it much easier to study the degree of presence of dialect forms, either in the entire corpus or in geographically defined subcorpora. In addition, the 'Text types' option allows the creation of subcorpora based on the metadata entered in the corpus; metadata make it possible to restrict queries based on attributes assigned to speakers and speech events, e.g., to look for instances of post-positive article use in a specific geographic region. Overall, it is possible to perform complex queries combining words, lemmas and parts-of-speech, restricting them on the basis of the specific characteristics of speech events and the speakers who participated in them.

It should also be emphasized that different search options treat accented letters differently. The 'simple query' option is accent-neutral, allowing users to type in non-accented queries and returning results with forms with the accent in all possible positions. Searches by lemma and tag are by definition accent-independent, as these attributes are coded in layers added to the original text. The remaining options require words to be typed in with accents.

Another type of information that can be obtained from the corpus are frequency data for specific words, lemmas, or parts of speech. Queries can also be adapted to allow the extraction of syntactic constructions with components that can be defined through single words or PoS tags (e.g., the passive). More options are of course available to users who

download the entire corpus and use their own scripts for extracting information (see Vuković et al., this volume, for examples of this approach).

Another technical challenge that will be relevant for future work is that *NoSkE* does not have in-built facilities for text-audio/video alignment. However, it allows their integration through external tools (see e.g., Bernardini et al. 2018).

5. Conclusion

Despite the fact that a great deal of work has been dedicated to the best practices in spoken corpora creation (see e.g., Haugh et al. 2014; Raso, Mello 2014), each corpus presents its own specific problems. In this paper, we attempted to provide as complete an account as possible of the steps involved in designing (in both the conceptual and technical sense) and building sociolinguistic spoken dialect corpora of Torlak, and in particular of a subcorpus from Timok. We described the process of data transcription and the transformation of the transcribed speech samples into a corpus that is not only searchable, but also sustainable and reusable.

In the next steps, the annotation of the Lužnica subcorpus will be finalized, complementing the already available Timok subcorpus, which will in turn be re-annotated with an updated version of the tagger, trained on a larger and more varied manually checked dataset. The work on the normalization layer in linguistic annotation will be continued, and we will also consider solutions for the alignment of textual transcripts with their audio sources. The information needed for alignment is already coded in the transcripts, so the work will mostly involve finding technical solutions for allowing end users to consult the aligned corpus, given that, as noted by Raso and Mello (2014: 3), "[…] corpora which do not foresee any form of text-to-sound alignment, through any of the various software now available, are completely insufficient for a realistic analysis of the structuring of speech".

The work on the Bulgarian data will also continue, as will the work on the data from SAOSWB, with a joint Serbian-Bulgarian corpus in mind as a post-*TraCeBa* objective. This will pose additional challenges related to unifying the transcripts in terms of format and transcription conventions. An acceptable approach to the annotation of Bulgarian data will also need to be identified, taking into account the availability of the Serbian-based tagger for Torlak on the one hand, and the influence of Standard Bulgarian on the other.

Last but not least, in order to attract as many researchers as possible to use the Torlak corpus, we plan to prepare tutorials aimed at making the corpus appeal to dialectologists, ethnolinguists and anthropological linguists, who typically conduct qualitative studies on manually annotated data, often focusing on larger texts, and who might show reluctance towards having to formulate corpus queries and explore results in concordance format. Other issues to be addressed will be the necessarily less-than-perfect automatic annotation, which might be perceived as problematic by researchers coming from a more traditional paradigm, and the difficulty of extracting many (morpho)syntactic phenomena automatically (e.g., the analytical future forms, whose elements do not have fully fixed positions or order). These issues can possibly be addressed through consultations with dialectologists on the most serious tagging errors, but also through explanations of how to make use of the raw corpus files, which are available for download. Even though such liaising activities are still being developed, our hope is that some of them will be realized jointly or soon after the release of the version 2.0 of the Serbian corpus, containing data from Timok and Lužnica.

University of Belgrade and University of Bologna — Miličević Petrović Maja, maja.milicevic2@unibo.it

University of Zurich — Vuković Teodora, teodora.vukovic2@uzh.ch

Institute of Balkan Studies, Serbian Academy of Sciences and Arts — Mirić Mirjana, mirjana.miric@bi.sanu.ac.rs

Russian Academy of Sciences, Institute for Linguistic Studies — Konior Daria V., dariakonior@iling.spb.ru

University of Zurich — Escher Anastasia, anastasia.escher@uzh.ch

References

Anderwald, Lieselotte; Wagner, Susanne. 2007. "FRED – The Freiburg English Dialect Corpus: Applying Corpus-Linguistic Research Tools to the Analysis of Dialect Data." In Joan Beal; Karen Corrigan, Hermann Moisl (eds). *Creating and Digitizing Language Corpora. Volume 1: Synchronic Databases*, 35–53. Basingstoke: Palgrave Macmillan.

Batanović, Vuk; Ljubešić, Nikola; Samardžić, Tanja. 2018. "SETimes.SR – A Reference Training Corpus of Serbian." In Darja Fišer; Andrej Pančur (eds). *Proceedings of the Conference on Language Technologies & Digital Humanities 2018 (JT-DH 2018)*, 11–17. Ljubljana: Ljubljana University Press, Faculty of Arts.

Bernardini, Silvia; Ferraresi, Adriano; Russo, Mariachiara; Collard, Camille; Defrancq, Bart. 2018. "Building Interpreting and Intermodal Corpora: A *How-to* for a Formidable Task." In Mariachiara Russo; Claudio Bendazzoli; Bart Defrancq (eds). *Making Way in Corpus-Based Interpreting Studies*, 21–42. Singapore: Springer.

Ćirković, Svetlana. 2018. "Uvod." In Svetlana Ćirković (ed.) *Timok. Folkloristička i lingvistička terenska istraživanja 2015-2017*, 7–15. Knjaževac: Narodna biblioteka "Njegoš", Beograd: Udruženje folklorista Srbije.

Ćirković, Svetlana; Konior, Daria V.; Sobolov, Andrey N.; Mirić, Mirjana. (this volume). "Challenges for Torlak Data Collection."

Erjavec, Tomaž. 2019. "Serbo-Croatian Specifications." In Tomaž Erjavec (ed.). *MULTEXT-East Morphosyntactic Specifications Version 6*. http://nl.ijs.si/ME/V6/msd/html/msd-hbs.html, accessed October 14, 2021.

Evert, Stefan; CWB Development Team. 2019. *The IMS Open Corpus Workbench (CWB) CQP Query Language Tutorial*. http://cwb.sourceforge.net/files/CQP_Tutorial/, accessed October 14, 2021.

Haugh, Michael; Ruhi, Sükriye; Wörner, Kai; Schmidt, Thomas (eds). 2014. *Best Practices for Spoken Corpora in Linguistic Research*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Jakubíček, Miloš; Kilgarriff, Adam; McCarthy, Diana; Rychlý, Pavel. 2010. "Fast Syntactic Searching in Very Large Corpora for Many Languages." In Ryo Otoguro; Kiyoshi Ishikawa; Hiroshi Umemoto; Kei Yoshimoto; Yasunari Harada (eds). *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, 741–7. Sendai: Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Ljubešić, Nikola; Klubička, Filip; Agić, Željko; Jazbec, Ivo-Pavao. 2016. "New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian." In Nicoletta Calzolari; Khalid Choukri; Thierry Declerck; Sara Goggi; Marko Grobelnik; Bente Maegaard; Joseph Mariani; Helene Mazo; Asuncion Moreno; Jan Odijk; Stelios Piperidis (eds). *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 4264–70. Portorož: European Language Resources Association (ELRA).

Machálek, Tomáš. 2020. KonText: Advanced and Flexible Corpus Query Interface. In Nicoletta Calzolari; Frédéric Béchet; Philippe Blache; Khalid Choukri; Christopher Cieri; Thierry Declerck; Sara Goggi; Hitoshi Isahara; Bente Maegaard; Joseph Mariani; Hélène Mazo; Asuncion Moreno; Jan Odijk; Stelios Piperidis (eds). *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. 7003–8. Marseille: European Language Resources Association (ELRA).

Machálek, Tomáš; Křen, Michal. 2013. "Query Interface for Diverse Corpus Types." In Katarína Gajdošová; Adriána Žáková (eds). *Natural Language Processing, Corpus Linguistics, E-learning*, 166–173. Lüdenscheid: RAM Verlag.

Makarova, Anastasija; Konior, Darja. V.; Vukovič, Teodora; Sobolev, Andrej N.; Winistörfer, Olivier. 2020. "Avtomatičeskij metod jazykovogo profilirovanija nositelja dialekta (na materiale vostočnoserbskogo idioma sela Berčinovac)." *Acta Linguistica Petropolitana* 16.2: 160–80.

OLA. 1994. *Obščeslavjanskij lingvističeskij atlas. Vstupitel'nyj vypusk*. 2nd Edition. Moskva: Nauka.

Raso, Tommaso; Mello, Heliana (eds). 2014. *Spoken Corpora and Linguistic Studies*. Amsterdam: John Benjamins.

Raso, Tommaso; Mello, Heliana. 2014. "Introduction." In Tommaso Raso; Heliana Mello (eds). *Spoken Corpora and Linguistic Studies*, 1–24. Amsterdam: John Benjamins.

Rychlý, Pavel. 2007. "Manatee/Bonito-A Modular Corpus Manager." In Petr Sojka; Aleš Horák (eds). *RASLAN 2007: Recent Advances in Slavonic Natural Language Processing*, 65–70. Brno: Masaryk University.

Schmidt, Thomas. 2009. "Creating and Working with Spoken Language Corpora in EXMARaLDA." In Verena Lyding (ed.). *Lesser Used Languages & Computer Linguistics II*, 151–64. Bolzano: Eurac Research.

Sobolev, Andrey. 1998. *Sprachatlas Ostserbiens und Westbulgariens*. Marburg, Lahn: Biblion-Verlag.

Sobolev, Andrey N.; Gorlov, Nikita G. 2021. *Digital Linguistic Atlas of Eastern Serbia and Western Bulgaria*. Saint Petersburg: Russian Academy of Sciences, Institute for

Linguistic Studies. https://nggorlov.shinyapps.io/saoswb/, accessed October 14, 2021.
Sobolev, Andrey N.; Mirić, Mirjana; Konior, Daria V.; Ćirković, Svetlana. (this volume). "Torlak: Areal Embedding and Linguistic Characteristics."
Šantić, Danica. 2019. "Uticaj migracija stanovništva na demografski razvitak knjaževačkog kraja." In Mikica Sibinović; Vladana Stojadinović; Danijela Popović Nikolić (eds). *Knjaževački kraj - potencijali, stanje i perspektive razvoja*, 62–77. Knjaževac: Narodna biblioteka "Njegoš".
Vuković, Teodora. 2021. "Representing Variation in a Spoken Corpus of an Endangered Dialect: The Case of Torlak." *Language Resources and Evaluation* 55: 731–56.
Vuković, Teodora. 2020. "Torlak Specifications." In Tomaž Erjavec (ed.) *MULTEXT-East Morphosyntactic Specifications Version 6*. http://nl.ijs.si/ME/V6/msd/html/msd-sr-tor.html, accessed October 14, 2021.
Vuković, Teodora. 2016. *Torlak Transcription Guidelines*. Unpublished manuscript. Zurich: University of Zurich.
Vuković, Teodora; Mirić, Mirjana; Escher, Anastasia; Ćirković, Svetlana, Miličević Petrović, Maja; Sobolev, Andrey N.; Sonnenhauser, Barbara. (this volume). "Under the Magnifying Glass. Dimensions of Variation in the Contemporary Timok Variety."

Создание торлакского социолингвистического корпуса

В статье описывается процесс создания двух корпусов, нацеленных на документирование торлакской речи в Восточной Сербии и Западной Болгарии. Корпусы будут отражать базовые торлакские языковые признаки с учетом их географической и социолингвистической вариативности. Создание корпусов описывается от этапа выборки записанных в поле интервью до этапа их транскрибирования (для болгарского корпуса) и от этапа выборки до кодирования метаданных, морфосинтаксической аннотации и публикации (для находящегося в открытом доступе Тимокского подкорпуса сербского корпуса объемом в 498,021 словоформ). Охарактеризованы проблемы, с которыми столкнулись разработчики на различных этапах создания корпусов, обоснованы пути их решения, причем основное внимание уделено лингвистической стороне дела. Даны разъяснения о сбалансированном применении собственно лингвистических и социолингвистических критериев при отборе образцов речи, подлежащих транскрибированию, о решениях в связи с трактовкой диалектных явлений в транскрипции и аннотации, о компьютерных методах аннотирования. Намечены также дальнейшие шаги, преимущественно в части расширения в близком будущем сербского корпуса Лужницким подкорпусом, а также о планах объединения сербских и болгарских материалов в едином "Устном торлакском диалектном комплексе" (англ. *Spoken Torlak Dialect Corpus*).