CONVERTING RAW TRANSCRIPTS INTO AN ANNOTATED AND TURN-ALIGNED TEI-XML CORPUS: THE EXAMPLE OF THE CORPUS OF SERBIAN FORMS OF ADDRESS

Dolores LEMMENMEIER-BATINIĆ

Department of Slavonic Languages and Literatures, University of Zurich

Lemmenmeier-Batinić, D. (2021): Converting raw transcripts into an annotated and turn-aligned TEI-XML corpus: the example of the Corpus of Serbian Forms of Address. Slovenščina 2.0, 9(1): 123–144.

DOI: https://doi.org/10.4312/slo2.0.2021.1.123-144

This paper describes the procedure of building a TEI-XML corpus of spoken Serbian starting from raw transcripts. The corpus consists of semi–structured interviews, which were gathered with the aim of investigating forms of address in Serbian. The interviews were thoroughly transcribed according to GAT transcribing conventions. However, the transcription was carried out without tools that would control the validity of the GAT syntax, or align the transcript with the audio records. In order to offer this resource to a broader audience, we resolved the inconsistencies in the original transcripts, normalised the semi-orthographic transcriptions and converted the corpus into a TEI-format for transcriptions of speech. Further, we enriched the corpus by tagging and lemmatising the data. Lastly, we aligned the corpus turns to the corresponding audio segments by using a force-alignment tool. In addition to presenting the main steps involved in converting the corpus to the XML-format, this paper also discusses current challenges in the processing of spoken data, and the implications of data re-use regarding transcriptions of speech. This corpus can be used for studying Serbian from the perspective of interactional linguistics, for investigating morphosyntax, grammar, lexicon and phonetics of spoken Serbian, for studying disfluencies, as well as for testing models for automatic speech recognition and forced alignment. The corpus is freely available for research purposes.

Keywords: spoken Serbian, language biographical interviews, forms of address, data re-usability

1 INTRODUCTION

Serbian has long been an under-resourced language despite the long tradition of work on language corpora in the “West Balkans” (see Dobrić, 2012). Up until the past decade, there have been only two notable corpora of Serbian: Corpus of Serbian Language (Kostić, 2003) and SrpKor Corpus of Contemporary Serbian Language (Krstev and Vitas, 2005; Popović, 2010; Utvić, 2011). In the past decade, several corpora have been created in order to amend the lack of resources regarding the written data (Ljubešić and Klubička, 2014; Ljubešić et al., 2016; Miličević and Ljubešić, 2016; Batanović et al., 2018). However, although there has been a global increase in popularity of spoken language resources and tools (see Batinić et al., to appear), Serbian still lacks spoken language corpora. Considerable advances have been made regarding the Torlak dialect (Vuković, 2021), resources for automatic speech recognition and synthesis (Delić et al., 2013; Suzić et al., 2014), and specialised spoken corpora, such as the SCECL[1] corpus on early child language (Anđelković et al., 2001) and SrMaCo[2] corpus on language of Serbian minority in Hungary.

Creating corpora of spoken language demands not only field access in order to obtain recordings of spoken language data, but also intensive manual work to transcribe them. These two steps are usually the most time-consuming in the corpus creation, and prevent spoken corpora from growing at the same pace as written corpora (see Schmidt, 2016, pp. 127–128). Therefore, in order to address the lack of spoken language resources, it is convenient to start compiling spoken corpora from existing recordings and transcriptions. This paper presents a compilation of a corpus of Serbian forms of address, which has been created from an existing collection of interviews gathered for investigating Serbian forms of address (Ulrich, 2018). The interviewees were asked about forms (expressions) they use to address their relatives, friends, colleagues, neighbours, etc. The corpus contains 19 transcriptions of interviews amounting to a total of 171,552 tokens (19,5 hours of speech).

[1] *Serbian Corpus of Early Child Language (SCECL).* Available at: https://sla.talkbank.org/TBB/childes/Slavic/Serbian/SCECL.

[2] *Spoken corpus of the Serbian minority in Hungary (SrMaCo).* Available at: http://spokencorpus.eu/cms/bosco-2/.

While the first steps of the corpus compilation have been presented in Lemmenmeier-Batinić et al. (2020), this paper discusses them in more detail, and shows some additional steps that have been made since, such as evaluation of linguistic annotations, and integration of forced alignment. It also discusses the implications of data re-use for linguistic research, and encourages further sharing of high-quality transcripts of speech, while at the same time stressing the importance of using current transcription tools for facilitating not only one's own work, but also the future usability of collected material.

2 CORPUS OF SERBIAN FORMS OF ADDRESS

2.1 Recordings and metadata

The source data consists of transcriptions and audio-files of interviews with 19 participants (9 female, 10 male). The topic of the interviews are Serbian expressions that are used to address other people. The interview guidelines have four main parts: in the first part, the interviewer asks questions about forms of address interviewees use to address family members, friends, neighbors, colleagues, etc. In the second part, questions are asked about forms of address for people that have some particular profession or function. In the third part, the interviewer lists certain forms of address, and asks if participants use them. In the fourth part of the questionnaire, interviewees have the opportunity to elaborate on the topic of their attitudes and assessments about particular forms of address.[3] The interviews were recorded during 2008 and 2009. The interviewer (female) was aged 27 at the time of recording. With the exception of the interviewer, who acquired Serbian as a foreign language, all the interviewees are native speakers of Serbian. At the time of recording, participants were aged 27 to 64 years. Most of them resided in Belgrade and Niš, and had a university degree (see Table 1).

Most interviews were held in private homes. However, some of them were recorded in bars, restaurants or shopping malls, which often resulted in lower quality of audio-recordings. The interviews last about 61 minutes in average, and contain 171,552 tokens (10,045 types).[4] An overview over the size of each transcript in tokens and minutes is given in Table 2.

[3] See Ulrich (2018, pp. 338–341) for detailed interview guidelines.

[4] The token count includes full and truncated words.

Table 1: *Speaker metadata*

Id	Sex	Age	Origin	Residency	Education
S	f	27	CH	Zurich	university
F1	f	28	Belgrade	Belgrade	technical college
F2	f	27	Belgrade	Zurich	university student
F3	f	27	Niš	Niš, Kotor	university
F4	f	44	Lazarevo	Belgrade	university
F5	f	58	Belgrade	Belgrade	university
F6	f	55	Niš	Niš	university
F7	f	55	Skopje	Niš	high school
F8	f	64	Leskovac	Niš	high school
F9	f	60	Pirot	Niš	technical college
M1	m	28	Niš	Niš	university
M2	m	27	Niš	Niš, Kotor	university
M3	m	29	Niš	Niš	university
M4	m	27	Užice	Belgrade	university student
M5	m	33	Belgrade	Belgrade	university
M6	m	27	Belgrade	Belgrade	high school
M7	m	38	Belgrade	Belgrade	university
M8	m	44	Belgrade	Belgrade	high school
M9	m	54	Niš	Niš	university
M10	m	61	Belgrade	Belgrade	university

Table 2: *Transcript length and duration*

Transcript	Token count	Duration
F1	12,784	01:24:53
F2	8,463	01:12:12
F3	9,135	00:55:25
F4	5,995	00:38:26
F5	9,159	00:55:12
F6	7,365	00:40:40
F7	6,693	00:48:19
F8	5,408	00:44:21
F9	13,681	01:29:55
M1	9,140	00:58:33
M2	11,653	01:20:21

Transcript	Token count	Duration
M3	7,283	00:51:08
M4	10,445	01:11:46
M5	11,762	01:07:43
M6	9,836	01:18:54
M7	9,774	01:05:27
M8	6,485	00:45:44
M9	5,260	00:36:59
M10	11,231	01:29:12
Total	171,552	19:35:10

The participants originally agreed to their data being used for the project of investigating Serbian forms of address by Ulrich (2018). For securing the possibility of data re-use for other research projects as well, interviewees were retraced in 2020/2021 and they were asked to sign a data privacy agreement stating that their interviews can be used for research purposes.[5] The audio files were cut in order to match exactly with the start and the end of the corresponding transcripts prior to any other processing.

2.2 Transcripts

Although the aim of the data collection was a content analysis (see Ulrich, 2018), all the interviews were thoroughly transcribed following the GAT transcribing conventions (Selting et al., 1998, 2009), which were originally developed for purposes of conversation analysis and interactional linguistics. GAT differentiates between three levels of transcription granularity: *minimal* (Selting et al., 2009), *basic* and *fine-grained* (Selting et al., 1998, 2009). Ulrich's (2018) transcripts contain most features of basic transcripts (annotation of pauses, breathing, incidents, overlaps, vocal length, etc.), while some other features are omitted (such as segmenting turns in intonational phrases, and annotation of pitch movement) or sporadically applied (like focus accent annotation). Some features of fine-grained transcription conventions were used, out of which some were consistently applied in all transcripts, such as the annotation of pace and loudness (<<p>...>), and other were used only

[5] Three participants could not be retraced and two of them had passed away. We do not share the audio interviews of these participants.

occasionally, such as the annotation of pitch jumps (↑). Overlaps were marked with square brackets, as proposed in GAT, but they were not vertically aligned, so it is not always possible to reconstruct which segments overlap with which. An excerpt from one of the transcripts is given in Example 1.

Example 1: *Excerpt from an original transcript (transcript id: F8)*[6]

S: i: e: i samo (--) kako (--) e:: (.) kako VAs oslovljavaju na pijaci (-) kad vi: kupujete

K: ko kako (.) ko gospođo (-) ko (--) e: seko ko: (-) ženo (-) ko kako (.) kom kako <<lachend> padne napamet> ((lacht))

S: ↑e: da: (-) <<p> pa da (.) za= (-) primetila sam na pijaci (.) ima naj ((lacht)) zanimljivije [((lacht))]

K: [da (--) pa] pa pijaca je uopšte najzanimljivija

S: jeste

K: najzanimljivija i: (-) .h i ovo= ove (-) emisije kad gledamo preko televizije kad

S: aha

K: uglavnom se posećuju PIjace jer je tu nešto najinteresantnije [((lacht))]

[6] For reasons of clarity, some annotations are omitted in the English translation:
S: and e: and just (--) how (-) e:: (.) how do people address you at the market (-) when you are buying
K: it depends who (.) some say misses (-) some (--) e: sister some (-) women (-) it depends who (.) it depends how <laughing> it occurs to them> ((laughs))
S: oh yes (-) <<p> well yes (.) I noticed it's most ((laughs)) interesting at the market ((laughs))
K: yes (--) well the market is the most interesting of it all
S: yes it is
K: the most interesting and: (-) .h and this= those (-) shows we watch on television when
S: aha
K: they mostly visit the markets because there is something most interesting there ((laughs))
S: e (-) ((laughs)) yes (-) ((laughing)) exactly
[...]
S: mhm mhm (<<p> mhm) good> .hh e: so how would you (-) e: address a taxi driver for example
K: (2.5s) m exclusively with the polite form
S. mhm (--) mhm
K: exclusively with the polite form (-) .h I don't' use <<rall> sir> to address

S: [e (-) ((lacht)) da] (-) ((lächelnd)) baš tako

[...]

S: mhm mhm (<<p> mhm) dobr↑o> .hh e: onda kako biste (-) e: oslovljavali vozača taksija naprimer

K: (2.5s) m isključivo sa vi

S: mhm (--) mhm

K: isključivo sa vi (-) .h <<rall> ne oslov>ljava= o=oslovljavam <<rall> gospodine>

The transcripts are very consistent, despite the fact that all interviews were transcribed without using any transcription software that would control the GAT syntax, and that the transcripts were originally not meant for re-distribution to a larger audience. However, with such a large amount of manual work, inconsistencies and typing errors are inevitable. For instance, different types of parenthesis ("(", "((", and "{") were occasionally used to annotate same information. Metalinguistic annotations were mostly written in German ("lacht" 'laughs'), but sometimes also in Serbian ("smeje se"). Rarely, symbols that are not proposed in GAT were used (* - <). The symbol "=" was, amongst other uses, frequently used for marking truncated (incomplete) words, which differs from its description in GAT, where it is proposed for marking fast continuation of new segments ("*latching*", Selting et al., 2009, p. 392; Selting et al., 1998, p. 31), or for marking contractions ("und=äh") and two syllabic reception signals such as "hm=hm" (only in the first GAT version, see Selting et al., 1998, p. 31). However, the frequent annotation of truncated words with "=" provided very valuable information, and was kept for further processing. Despite some inconsistencies, the transcriptions were accurate enough to permit a conversion into a standardised format such as XML, while including (most) annotations in the markup. Interviews were originally transcribed in Microsoft Word, and were converted to plain text files in order to allow for further data processing. The original files had a simple structure (one line for each speaker turn) and transporting them to plain text required no additional editing.

3 CONVERTING THE CORPUS TO TEI-XML

3.1 Preprocessing

Prior to XML-conversion, annotations of incidents, gaps, comments, pace, loudness, ambiguous segments ("je/i" 'it is/and') and occurrences of annotations with the equals sign ("=") were extracted, corrected, and made consistent. For instance, since the use of parentheses was not always consistent, all the parentheses were checked and marked with the corresponding label in the intermediate step (see Table 3).

Table 3: *Categorising comments in the preprocessing step (excerpt)*

Original annotation	Changes (intermediate step)
{Auslassung 14:58-15:53} *omission 14:58-15:53*	((gap:extent: 55s))
{Telefon klingelt} *the phone is ringing*	((incident: zvoni telefon))
((klopft auf den Tisch)) *knocks on the table*	((incident: kuca o sto))

In total, 707 unique annotations were checked, out of which 665 have been changed, and stored into intermediate (clean) transcript text files. Most corrections were related to the use of the equals sign, metalinguistic comments, and annotations of pace and speed that were set in the middle of words, which had to be reconstructed (for instance: "mla<<lachend> đi>" was changed to "mlađi" 'younger'; "po<imenu" was changed to "po imenu" 'by name'). The metalinguistic comments were translated into Serbian (see Table 3). Although they had to be adjusted in the preprocessing step, features of fine-grained transcription were not considered in further processing, because they were either seldom used in the transcripts (annotation of pitch jumps and focus accents) or because their conversion to TEI required prioritisation of overlapping annotations (in cases like "<<rall/p>...>"), and annotation of shifts on a sub-word level (like in "mla<<lachend> đi>" 'younger'). As shown in Section 3.3, we opted to keep the segmentation at word-level, and to provide a structure that makes XML-search and parsing of words as basic entities an undemanding task.

3.2 Normalisation

The interviews were transcribed based on their phonetic realisation, hence not always according to orthographic rules. In order to provide a corpus with normalised (standard) variants as well, tokens that did not occur in the Serbian lexicon srLex[7] (Ljubešić et al., 2016) were extracted and manually checked. Out of 387 types that were not present in srLex, 119 were correct (mostly rare words, proper names, or colloquialisms). The remaining 268 had to be normalised. Two types of normalised tokens were stored for further processing: corrections of transcriber's orthographic or typing errors (ex."označavaju" for "osnačavaju" 'they mark'), and standard variants of spoken forms (ex. "hoćete" for "oćete" 'you want'). The normalisation affected 4,055 tokens (2.4%) and 972 types (9.7%) in the corpus.

3.3 Marking up the corpus with TEI-annotations

Preprocessed transcripts have been converted into XML format following TEI conventions for transcriptions of speech.[8] Transcripts were segmented in speaker turns (<u>), and each turn was further segmented into full words: <w>, truncated words: <del>, unclear segments: <unclear>, gaps: <gap>, incidents: <incident>, vocalised non-lexical elements: <vocal>, and pauses: <pause>. Words that have been normalised to standard forms are stored in the @norm attribute. The original orthographic or transcription mistakes are stored as @orig. In addition to lemmatised and normalised forms, universal part-of-speech tags (@pos)[9] and MULTEXT-East Serbo-Croatian morphosyntactic specifications (@ana)[10] are provided (see Section 3.4). The attributes @start and @end point to the intervals in the audio-recordings defined in the <timeline> element (see Section 3.5).

[7] *Inflectional lexicon srLex 1.3*. Available at: Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1233.

[8] *TEI Guidelines Version 4.2.1 (Transcriptions of Speech)*. Available at: https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html.

[9] *Universal POS tags*. Available at: https://universaldependencies.org/u/pos/.

[10] *Serbo-Croatian MULTEXT-East Specifications*. Available at: http://nl.ijs.si/ME/V6/msd/html/msd-hbs.html. In the sixth and most recent MULTEXT-East release, Croatian, Serbian, and Bosnian specifications were replaced by Serbo-Croatian specifications, which cover the Croatian, Serbian, Bosnian and Montenegrin languages.

Example 2: *TEI version of the last turn shown in Example 1 (including the relevant lines in the element <timeline>)*

```
<TEI xmlns=“http://www.tei-c.org/ns/1.0” version=“4.1.0”>
  <text>
    <body>
      <timeline unit=“s” corresp=“../audio/F8.wav” origin=“#F8-u1-t0”>
        […]
        <when xml:id=“F8-u1-t0”/>
        <when xml:id=“F8-u366-t1” interval=“1024.2979996425654” since=“#F8-u1-t0”/>
        <when xml:id=“F8-u366-t2” interval=“1029.1793571238047” since=“#F8-u1-t0”/>
        […]
      </timeline>
      […]
      <u who=“F8” xml:id=“F8-u366” start=“#F8-u366-t1” end=“#F8-u366-t2”>
        <w xml:id=“F8-u366-w1” lemma=“isključivo” pos=“ADV” ana=“mte:Rgp”>isključivo</w>
        <w xml:id=“F8-u366-w2” lemma=“sa” pos=“ADP” ana=“mte:Si”>sa</w>
        <w xml:id=“F8-u366-w3” lemma=“vi” pos=“PRON” ana=“mte:Pp2-pn”>vi</w>
        <pause type=“short” xml:id=“F8-u366-p4”/>
        <vocal>
          <desc xml:id=“F8-u366-v5”>inhale (short)</desc>
        </vocal>
        <w xml:id=“F8-u366-w6” lemma=“ne” pos=“PART” ana=“mte:Qz”>ne</w>
        <w xml:id=“F8-u366-w7” lemma=“oslovljavati” pos=“VERB” ana=“mte:Vmr3s”>oslovljava</w>
        <del type=“truncation” xml:id=“F8-u366-w8”>o</del>
        <w xml:id=“F8-u366-w9” lemma=“oslovljavati” pos=“VERB” ana=“mte:Vmr1s”>oslovljavam</w>
        <w xml:id=“F8-u366-w10” lemma=“gospodin” pos=“NOUN” ana=“mte:Ncmsv”>gospodine</w>
      </u>
      […]
    </body>
  </text>
</TEI>
```

3.4 Lemmatisation and morphosyntactic annotations

3.4.1 TAGGER

The normalised corpus was tagged with the tagger for Serbian and other South-Slavic languages CLASSLA-StanfordNLP (Ljubešić and Dobrovoljc, 2019), which is a fork of the StanfordNLP tagger.[11] The estimate of the accuracy on standard data for Serbian is 97.89 F1 for lemmatisation, and 95.23 F1 for morphosyntactic annotations. As in the first version (Lemmenmeier-Batinić et al., 2020), the corpus was tagged with a model trained on a set of all available training data for Serbian and Croatian: SETimes.SR 1.0 corpus of

[11] *Classla 1.0.0* (CLASSLA Fork of Stanza for Processing Slovenian, Croatian, Serbian, Macedonian and Bulgarian). Available at: https://pypi.org/project/classla/.

newspaper texts (Batanović et al., 2018)[12], the hr500k Croatian reference training corpus (Ljubešić et al., 2016)[13], the ReLDI-NormTagNER, corpus of Serbian and Croatian tweets (Miličević and Ljubešić, 2016)[14,15], and the RAPUT corpus of Croatian non-professional writing (Štefanec et al., 2016). While in the first version of this corpus the tagger erroneously tagged several Ekavian words with Ijekavian lemmas (for instance, "hteo" 'wanted' was lemmatised as "htjeti" instead of "hteti" 'to want'), this feature was corrected in the second version, as the tagger was set to prefer Ekavian instead of Ijekavian variants.[16]

3.4.2 Evaluation of the TAGGER output

The accuracy of the tagger on our data was evaluated by checking the annotation of the first 500 tokens in one transcript.[17] The lemmatiser performed well with an accuracy of 98.2 F1. However, having both Serbian and Croatian corpora in the training set occasionally caused lemmatisation errors, since some word forms were annotated with lemmas characteristic of the Croatian, rather than the Serbian standard variety (such as the lemma "netko" [hr.] instead of "neko" [sr.] for the word form "neko" 'somebody').[18] The accuracy of morphosyntactic tags amounted to 92.2, which is, as expected, lower than the estimated accuracy for standard language data. Tagging errors are likely due to spoken

[12] *Training corpus SETimes.SR 1.0*. Available at: Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1200.

[13] *Training corpus hr500k 1.0*. Available at: Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1183.

[14] *Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.1*. Available at: Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1240.

[15] *Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.1*. Available at: Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1241.

[16] The Proto-Slavic jat-vowel (ѣ in Cyrillic) has three different pronunciations in today's Shtokavian dialects: Ekavian (cf. first "e" in "vreme" 'time'), Ijekavian (cf. "ije" in "vrijeme") and Ikavian (cf. "i" in "vrime"). Standard Serbian has two variants: the Ekavian, which is spoken in most of Serbia, and the Ijekavian, which is spoken in south-west Serbia, but also in Croatia, Bosnia and Herzegovina and Montenegro. Since the corpus represents Serbian spoken by speakers using the Ekavian pronunciation (and living in Serbia), the tagger was set to prefer the Ekavian variants.

[17] The evaluation of tagger's performance on this dataset was made in order to examine challenges related to tagging spoken language data. An elaborate evaluation of the tagger model would require a bigger and more diversified sample.

[18] For this reason, in future versions we will test the tagger trained only on Serbian data.

character of the data, sometimes having a different word order, and extra-sentential elements that are rare in written (or standard) language data. One of the common tagging errors is the affirmative particle “da” (‘yes’, tag: “Qr”), which is frequently erroneously tagged as subordinating conjunction (“Cs”). Other erroneously tagged tokens are relative pronouns (“Pr”) such as “koji” (‘who’) that are tagged as indefinite pronouns (“Pi”), as well as the interrogative particle “kako” (‘how’), which is tagged as a subordinating conjunction (“Cs”).[19]

Since Serbo-Croatian MULTEXT-East specifications do not propose tags for discourse particles, annotations that were compatible with the current MULTEXT-East specifications were regarded as correct in the evaluation process. For example, “znači” (literally: ‘it means’) was not counted as an error when it was tagged as verb, although it was used as discourse marker instead (see Halupka-Rešetar and Radić-Bojanić, 2014 on “znači” as discourse marker). Since specifications are missing for other discourse markers as well, they were also regarded as correct if their tags corresponded to the proposed MULTEXT-East specifications (for instance, “pa” ‘well’ was regarded as correct if it was tagged as coordinating conjunction “Cc”). However, in order to capture the peculiarities of spoken language, morphosyntactic specifications should ideally be extended to include discourse particles, hesitation signals, tag questions and other recurrent phenomena of the spoken register. Some examples of tagsets that were adapted to spoken language are STTS 2.0 for German (Westpfahl et al., 2017), and VOICE tagset (2014) for English. Extending Serbo-Croatian morphosyntactic specifications to suit spoken language phenomena would not only be of advantage for linguists interested in their use, but also for researchers developing other tools for processing spoken data.[20]

3.5 Aligning the corpus with audio segments

The transcripts were not originally aligned with the respective audio segments. This made searching for particular transcript segments in the audio

[19] Specifications for tagging relative pronouns and interrogative particles are insufficiently documented in the MULTEXT-East specifications for Serbo-Croatian, which might have resulted in them being erroneously tagged not only in this, but also in other Serbian and Croatian corpora as well (see srWaC and hrWaC).

[20] See Dobrovoljc and Martinc (2018) on the impact of discourse markers on spoken language dependency parsing for Slovene.

file an arduous task. In order to obtain alignments for each speaker turn, two forced alignment tools were tested: *aeneas*[21], and the model proposed by Plüss et al. (2020), using the Google Cloud STT Serbian ASR model. While *aeneas* offers support for aligning Serbian data, the model by Plüss et al. (2020) is not specifically tailored for Serbian, but requires an external ASR model.

For the first evaluation, we examined the difference in turn onset within the first minute in 9 different transcripts (88 turns). A comparison of turn beginnings produced by these two forced alignment tools against manual alignments showed that the model by Plüss et al. (2020) performs convincingly better than *aeneas* on our data (see Table 4). An assessment of the accuracy of alignment of 200 consecutive turns (17.5 minutes) is shown in Table 5.

Table 4: *Average absolute difference between turn beginnings calculated by forced alignment tools compared to manual alignment (measured in seconds)*

Absolute difference in turn onset $\lvert$**turn start**$_{\text{forced alignment}}$ **–turn start**$_{\text{reference alignment}}\rvert$	**Plüss et al. (2020)**	***aeneas***
mean	1.17	10.32
median	0.58	2.75
standard deviation	1.88	15.14

Table 5: *Comparison of aeneas and the model by Plüss et al. (2020) regarding the accuracy of turn alignment in the transcript F1 (including non-lexical backchannels and affirmative particles)*

	Erroneously aligned turns	**Turns corresponding to the audio segments to a certain extent**			**Total**
		Partially correct	**Predominantly correct**	**Fully correct**	
Model by Plüss et al. (2020)	90 (45.0%)	21 (10.5%)	53 (26.5%)	36 (18.0%)	200 (100.0%)
aeneas	96 (48.0%)	27 (13.5%)	34 (17.0%)	43 (21.5%)	200 (100.0%)

At first glance in Table 5, both tools seem to produce unsatisfactory results: they both generate a high amount of erroneously aligned turns. *Aeneas* outputs more 'fully correct' alignments, but also more misalignments than the

[21] *Aeneas*. Available at: https://www.readbeyond.it/aeneas/.

model by Plüss et al. (2020). The high amount of errors is due to a high rate of turns consisting only of affirmative particles ("da" 'yes') and non-lexical backchannels such as "mhm", or "aha", which are frequently misaligned (respectively, not-aligned) by both tools. [22] However, when turns consisting only of non-lexical backchannels and affirmative particles (n=66), are omitted, it becomes evident that the model by Plüss et al. (2020) outputs better alignments on our data than *aeneas* (see Table 6).

Table 6: *Comparison of aeneas and the model by Plüss et al. (2020) regarding the accuracy of turn alignment in the transcript F1 (excluding non-lexical backchannels and affirmative particles)*

	Erroneously aligned turns	**Turns corresponding to the audio segments to a certain extent**			**Total**
		Partially correct	**Predominantly correct**	**Fully correct**	
Plüss et al. (2020)	24 (17.9%)	21 (15.7%)	53 (39.5%)	36 (26.9%)	134 (100.0%)
aeneas	54 (40.3%)	25 (18.7%)	24 (17.9%)	31 (23.1%)	134 (100.0%)

Misalignments produced by the model by Plüss et al. (2020) are fewer (17.9% in comparison to 40.3% by *aeneas*), and they always consist of short speaker turns, whereas *aeneas* frequently misaligns longer turns as well. Therefore, the corpus has finally been aligned with the model proposed by Plüss et al. (2020).[23] With the help of turn alignments, users can navigate the transcripts while being able to hear the respective turns in the same time (or detect their approximate location in the audio segment in case they are not fully correct). The alignments are provided for each turn in the TEI version of the corpus (see attributes @start and @end in Example 2).

[22] *Aeneas* has the advantage of *sometimes* producing correct alignments for these turns. However, the model by Plüss et al. (2020) has the advantage of pointing at empty alignments for these turns, so that they don't stand out as false positives during a manual inspection of alignments with transcription editors. The failed alignment of short and non-lexical backchannels is likely due to the fact that their transcription does not exactly correspond to their vocal realisation. A possible solution would be to add these alignments using transcription editors such as Partitur Editor (EXMARaLDA). However, this would require extensive manual adjustments, since non-lexical backchannels are frequent in our corpus (a search of all "aha", "hm", and "mhm" returns 5028 occurrences).

[23] Only one transcript (id: F2) could not be aligned with the audio segments with either of the two tools, probably due to the low quality of the recording.

4 DATA SHARING

The corpus is available on CLARIN.SI.[24] In addition to the TEI-XML version of the corpus presented in this paper, we also provide raw transcripts including all annotations. The work in progress is documented at the GitLab repository of ZuCoSlaV corpora (Zurich Corpora of Slavic Varieties).[25] In accordance with the data privacy agreement, audio files are available on request. The corpus is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA).[26]

5 POSSIBLE APPLICATIONS

The corpus presents a valuable resource for researchers interested in interactional linguistics, since it contains long fragments of natural language in interaction transcribed in great level of detail. The length of the transcripts, averaging to one hour of conversation, additionally allows one to study speaker-related peculiarities and different types of disfluencies produced in spontaneous conversation (pauses, truncations, self-repetitions, etc.). The almost equal number of male and female speakers allows for gender comparisons regarding content, as well as form-related phenomena. The corpus can be used for studying prosodic, lexical and morphosyntactic patterns of spoken Serbian. For instance, it is currently being used for investigating the use of simple past tenses and auxiliary omission in Serbian (Escher and Sonnenhauser, in preparation).

By providing semi-orthographic transcripts, this corpus may contribute to the development of tools for automatic speech recognition and forced alignment. Lastly, the XML encoding and annotation of the corpus also facilitates the study of forms of address, which are now normalised, lemmatised and tagged, and can be examined more easily by a quantitative approach.

[24] *Corpus of Serbian Forms of Address 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1422.

[25] *ZuCoSlaV: Zurich Corpora of Slavic Varieties*. Available at: https://gitlab.uzh.ch/uzh-slavic-corpora.

[26] Licence details are available at: https://creativecommons.org/licenses/by-nc-sa/4.0/.

6 DISCUSSION

The Corpus of Serbian Forms of Address represents a significant step towards filling the gap of missing linguistic resources for spoken Serbian. While converting existing transcriptions requires substantial amount of manual work in the preprocessing step, in our case, the gain was worth the effort, since the interviews are long, the speaker metadata is provided, and the corpus has been meticulously and relatively consistently transcribed. Therefore, it cost less effort to clean and convert the corpus to a TEI-format and include all annotations, than it would to collect and transcribe new data of spoken Serbian from scratch.

The processing steps presented in this paper are useful for other researchers wanting to re-use existing material to create annotated corpora, and thereby enhance the study of spoken language. However, before starting the work on converting existing transcripts to a standardised format such as TEI-XML, it is important to carefully examine the quality of the transcripts, given that, depending on transcription consistency, the length of the corpus, or data formatting issues, it might take more time to preprocess the data than to transcribe it again with recent transcription tools. Transcription tools (such as for instance, FOLKER[27]) can control the syntax of transcribing conventions and align text with audio/video segments. Using these tools would not only assist the transcriber him/herself, but it would also significantly reduce the amount of work invested in enabling data re-use on the part of any third parties.

Another important issue that would facilitate data re-use is resolving possible data-privacy issues from start by ensuring that participants are willing to permit data re-use for general research purposes (and not only for one specific project they are originally taking part of). Making own transcripts available to a larger audience guarantees the transparency of research, and enables development of further work based upon it. Hopefully, considerations discussed in this paper will encourage data sharing of further collections of transcripts, and assist other researchers in converting existing transcript collections into annotated corpora of transcriptions of speech.

[27] *FOLKER*. Available at: https://exmaralda.org/de/folker-de/.

7 CONCLUSION

Spoken language has long been overlooked not only when it comes to corpus resources, but also in regard to annotation conventions and development of models for automatic language processing. In addition to assessing the implications of data re-usability, and presenting a new resource for spoken Serbian, this paper addressed some unresolved issues regarding part-of-speech tags for spoken language phenomena, which are often left unspecified in the tagset specifications. An important step for further development of Serbian spoken language corpora would be to define the specifications for phenomena that are particular for the spoken register, such as discourse markers, non-lexical backchannels, hesitation markers, etc. The evaluation of forced alignment tools showed that there is also place for improvement regarding the implementation of Serbian models within current forced alignment tools. Using the approach of Plüss et al. (2020) via an open-domain ASR system for Serbian and resolving the issue of misaligned response tokens in future work would be a promising development for processing spoken Serbian data.

Acknowledgments

I would like to thank to Sonja Ulrich for sharing the transcripts and recordings she collected for her PhD thesis, and Tanja Samardžić (URPP Language and Space, Zurich) and Barbara Sonnenhauser (Department of Slavonic Languages and Literatures, Zurich) for enabling work on this corpus. I am also thankful to Nikola Ljubešić for tagging the corpus and for his insightful suggestions, Michel Plüss for aligning the corpus with his forced alignment model, and Miro Rodin, Petra Abramović and Luka Jovanović for their assistance in the evaluation of the automatic tools used for creating this corpus.

REFERENCES

Corpora, tools and tagsets

Aeneas. Retrieved from https://www.readbeyond.it/aeneas/

Classla 1.0.0 (CLASSLA Fork of Stanza for Processing Slovenian, Croatian, Serbian, Macedonian and Bulgarian). Retrieved from https://pypi.org/project/classla/

Croatian Twitter training corpus ReLDI-NormTagNER-hr 2.1. Retrieved from Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1241

FOLKER. Retrieved from https://exmaralda.org/de/folker-de/

Inflectional lexicon srLex 1.3. Retrieved from http://hdl.handle.net/11356/1233

Serbian Corpus of Early Child Language (SCECL). Retrieved from https://sla.talkbank.org/TBB/childes/Slavic/Serbian/SCECL

Serbian Twitter training corpus ReLDI-NormTagNER-sr 2.1. Retrieved from Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1240

Serbo-Croatian MULTEXT-East Specifications. Retrieved from http://nl.ijs.si/ME/V6/msd/html/msd-hbs.html

Spoken corpus of the Serbian minority in Hungary (SrMaCo). Retrieved from http://spokencorpus.eu/cms/bosco-2/

TEI Guidelines Version 4.2.1 (Transcriptions of Speech). Retrieved from https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html

Training corpus hr500k 1.0. Retrieved from Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1183

Training corpus SETimes.SR 1.0. Retrieved from Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1200

Universal POS tags. Retrieved from https://universaldependencies.org/u/pos/

ZuCoSlav: Zurich Corpora of Slavic Varieties. Retrieved from https://gitlab.uzh.ch/uzh-slavic-corpora

Other

Anđelković, D., Ševa, N., & Moskovljević, J. (2001). Serbian Corpus of Early Child Language. Laboratory for Experimental Psychology, Faculty of Philosophy, and Department of General Linguistics, Faculty of Philology, University of Belgrade.

Batanović V., Ljubešić, N., & Samardžić, T. (2018). SETimes.SR – A Reference Training Corpus of Serbian. *Proceedings of the Conference on Language Technologies & Digital Humanities 2018 (JT-DH 2018)* (pp. 11–17). Ljubljana, Slovenia.

Batinić, J., Frick, E., & Schmidt, T. (in press). Accessing spoken language corpora: An overview of current approaches. *Corpora. Edinburgh University Press.*

Delić V., Sečujski, M., Jakovljević, N., Pekar, D., Mišković, D., Popović, B., Ostrogonac, S., Bojanić, M., & Knežević, D. (2013). Speech and Language Resources within Speech Recognition and Synthesis Systems for Serbian and Kindred South Slavic Languages. In M. Železný, I. Habernal, A. Ronzhin (Eds.), *Speech and Computer. SPECOM 2013. Lecture Notes in Computer Science: Vol. 8113* (pp. 319–326). Springer, Cham. doi: 10.1007/978-3-319-01931-4_42

Dobrić N. (2012). Language Corpora in The West Balkans – History, Current State and Future Perspective. *Slavistična revija, 60*(4), 677–692.

Dobrovoljc, K., & Martinc, M. (2018). Er ... well, it matters, right? On the role of data representations in spoken language dependency parsing. *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)* (pp. 37–46). Brussels, Belgium.

Escher, A., & Sonnenhauser, B. (in press). Simple Past Tenses in the Timok dialect.

Halupka-Rešetar, S., & Radić-Bojanić. B. (2014). The discourse marker *znači* in Serbian: An analysis of semi-formal academic discourse. *Pragmatics, 24*(4), 785–798.

Kostić, A. (2003). Đorđe Kostić electronic corpus of the Serbian language. In *Zbornik Matice srpske za slavistiku: Vol. 64* (pp. 260–264).

Krstev, C., & Vitas, D. (2005). Corpus and Lexicon – Mutual Incompleteness. In *Proceedings of the Corpus Linguistics Conference*, 14–17 July 2005, Birmingham. United Kingdom (hal-01108218).

Lemmenmeier-Batinić, D., Ljubešić, N., & Samardžić, T. (2020). XML-Encoding of a spoken Serbian corpus targeting forms of address. In D. Fišer in T. Erjavec (Eds.), *Proceedings of the Conference on Language Technologies & Digital Humanities* (pp. 127–130). Ljubljana: Institute of Contemporary History.

Ljubešić N., & Klubička. F. (2014). {bs,hr,sr}WaC – Web Corpora of Bosnian, Croatian and Serbian. *Proceedings of the 9th Web as Corpus Workshop (WaC-9)* (pp. 29–35). Gothenburg, Sweden.

Ljubešić, N., Klubička, F., Agić, Ž., & Jazbec. I. (2016). New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4264–4270). Portorož, Slovenia.

Ljubešić, N., & Dobrovoljc, K. (2019). What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 29–34). Florence, Italy.

Miličević, M., & Ljubešić. N. (2016). Tviterasi, tviteraši or twitteraši? Producing and analysing a normalised dataset of Croatian and Serbian tweets. *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research, 4*(2), 156–188.

Plüss, M., Neukom, L., & Vogel, M. (2020). Swiss Parliaments Corpus, an Automatically Aligned Swiss German Speech to Standard German Text Corpus. Retrieved from https://arxiv.org/abs/2010.02810

Popović, Z. (2010). Taggers Applied on Texts in Serbian. *INFOtheca, 11*(2), 21–38.

Schmidt, T. (2016). Construction and Dissemination of a Corpus of Spoken Interaction – Tools and Workflows in the FOLK project. *Corpus linguistic software tools, 31*(1), 127–154.

Selting, M., Auer, P., Barden, B., Bergmann, J., Couper-Kuhlen, E., Günthner, S., Quasthoff, U., Meier, C., Schlobinski, P., & Uhmann, S. (1998). Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte 173*, 91–122.

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., Hartung, M., Kern, F., Mertzlufft, C., Meyer, C., Morek, M., Oberzaucher, F., Peters, J., Quasthoff, U., Schütte, W., Stukenbrock, A., & Uhmann, S. (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion,* (10), 353–402.

Suzić, S., Ostrogonac, S., Pakoci, E., & Bojanić. M. (2014). Building a Speech Repository for a Serbian LVCSR System. *Telfor Journal*, *6*(2), 109–114.

Štefanec, V., Ljubešić, N., & Kuvač Kraljević. J. (2016). Croatian Error-Annotated Corpus of Non-Professional Written Language. *Proceedings of the*

Tenth International Conference on Language Resources and Evaluation (LREC 2016) (pp. 3220–3226). Portorož, Slovenia.

Ulrich, S. (2018). *Anredeformen im Serbischen*. Wiesbaden.

Utvić, M. (2011). Annotating the Corpus of Contemporary Serbian. *INFOtheca 12*(2), 36–47.

VOICE (2014). Part-of-Speech Tagging and Lemmatization Manual. With assistance of Barbara Seidlhofer, Stefan Majewski, Ruth Osimk-Teasdale, Marie-Luise Pitzl, Michael Radeka, Nora Dorn. The Vienna-Oxford International Corpus of English. Retrieved from http://www.univie.ac.at/voice/documents/VOICE_tagging_manual.pdf

Vuković, T. (2021). Representing variation in a spoken corpus of an endangered dialect: the case of Torlak. *Language Resources and Variation*. Springer Nature. doi: 10.1007/s10579-020-09522-4

Westpfahl, S., Schmidt, T., Jonietz, J., and Borlinghaus, A. (2017). STTS 2.0. Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS). Working paper. Mannheim: Institut für Deutsche Sprache.

PRETVORBA ZBIRKE SUROVIH ZAPISOV V ANOTIRAN IN SPREMENJEN TEI-XML KORPUS: PRIMER KORPUSA SRBSKIH OBLIK NASLAVLJANJA

V prispevku je opisan postopek gradnje TEI-XML korpusa govorjenega srbskega jezika, začenši s surovimi prepisi. Korpus sestavljajo polstrukturirani intervjuji, ki so bili zbrani z namenom raziskati oblike naslavljanja v srbščini. Intervjuji so bili temeljito prepisani v skladu s konvencijami o prepisovanju GAT. Prepis pa je bil izveden brez orodij, ki bi nadzorovala veljavnost sintakse GAT ali poravnala prepis z zvočnimi zapisi. Da bi ta vir ponudili širši publiki, smo odpravili nedoslednosti v izvirnih prepisih, normalizirali polortografske prepise in korpus pretvorili v format TEI za prepise govora. Nadalje smo korpus obogatili z označevanjem in lematizacijo podatkov. Nazadnje smo z orodjem za prisilno poravnavo v korpusu poravnali govore posameznih govorcev s pripadajočimi segmenti govornega signala. Ta članek poleg predstavitve glavnih korakov pri pretvorbi korpusa v format XML razpravlja tudi o trenutnih izzivih pri obdelavi govorjenih podatkov ter o implikacijah ponovne uporabe podatkov pri prepisih govora. Korpus srbskih oblik naslavljanja lahko uporabimo za preučevanje srbščine z vidika interakcijske lingvistike, za raziskovanje morfosintakse, leksike in fonetike govorjenega srbskega jezika, za preučevanje disfunkcij ter za preizkušanje modelov za samodejno prepoznavanje govora in prisilno poravnavo. Korpus je prosto dostopen za raziskovalne namene.

Ključne besede: govorjena srbščina, jezikovni biografski intervjuji, oblike naslavljanja, ponovna uporabnost podatkov