XML-Encoding of a spoken Serbian corpus targeting forms of address

Dolores Lemmenmeier-Batinić[1], Nikola Ljubešić[2], Tanja Samardžić[3]

[1] Slavisches Seminar, University of Zurich
Plattenstrasse 43, 8032 Zurich
dolores.lemmenmeier@uzh.ch
[2] Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
nikola.ljubesic@ijs.si
[3] URPP Language and Space, University of Zurich
Freiestrasse 16, 8032 Zurich
tanja.samardzic@uzh.ch

1. Introduction

Serbian has long been an under-resourced language. However, following the recent global trend to represent languages with corpora, several projects have attempted to amend this lack of publicly available resources regarding the written register (Erjavec, 2012; Ljubešić and Klubička, 2014; Miličević and Ljubešić, 2016). Resources for spoken standard Serbian in interaction are still lacking, despite there being some work on Serbian dialectal corpora (Vuković et al., 2019) and repositories for speech recognition (Suzić et al., 2014). Meanwhile, tools and resources of spoken registers that are being created for similar South Slavic languages such as Croatian (Kuvač Kraljević and Hržica, 2016: hrAL) and Slovenian (Verdonik et al., 2013: GOS) are gaining in popularity.

Since collecting data of spoken interaction requires not only field research, but also intensive manual work, it is fruitful to start addressing this lack of resources by gathering existing material and arranging it in a way to make it useful for a larger audience. In this paper, we present the process of compiling a corpus of spoken Serbian starting from an existing collection of transcripts that have been gathered for investigating the use of forms of address in spoken Serbian (Ulrich, 2018). Given its relatively substantial size of 172,345 tokens and 10,061 types (in comparison, hrAL contains 250,000 tokens), a fine-grained transcription, and a complete documentation of metadata, the collection of interviews gathered by Ulrich (2018) presents a valuable linguistic resource for spoken Serbian. The current version of the corpus can be accessed at a SWITCHdrive public link.[1] For the long-term deposit, we plan to use CLARIN.Sl.

After describing the data, we show the procedure of detecting and correcting transcription inconsistencies, enriching the corpus with linguistic information, and converting this resource into a standard XML format. Our goal is to add value to this resource, underlining at the same time the need for data reuse and the need for sharing data from the start in similar future projects.

2. Data source

2.1. Recordings and metadata

The corpus consists of transcripts of audio-recorded biographical interviews with 19 participants (9 female, and 10 male) recorded in 2008 and 2009, lasting 63 minutes on average. The interviewees are asked about the forms of address they use in colloquial and in formal settings, and about attitudes and evaluations concerning particular forms of address. There is a set of questions asked in each interview, but the conversations are casual and often contain short anecdotal information. The majority of the participants at the time of recording resided in Niš and in Belgrade, and have a university degree. They predominantly speak in standard spoken Serbian, but sometimes also use regional varieties of particular forms.

2.2. Transcription

The records were transcribed according to the GAT transcription system (Selting et al., 1998; 2009), which differentiates between three levels of granularity in transcribing talk-in-interaction: minimal, basic and fine

[1] https://drive.switch.ch/index.php/s/oNpLQcsiRDojzuG (09.09.2020).

transcripts. Features belonging to all tree levels of granularity were used in Ulrich's (2018) transcriptions, although no convention has been adopted entirely. An excerpt of one of the transcripts is given in Example 1.

Example 1: Excerpt of an original transcript (transcript id: F2)

S: tako (-) .h kako osloviš članove (-) tvoje porodice
M: <<lachend> oslovljavaš> (--) to je pra= (e) perfektivni glagol
S: e (-) da
M: <<langsam> oslovljavaš>
S: dobro to (.) to sam još mislila (-) da li je možda (-) bolje (schreibt) .h (-) vljavaš (-) e (.) dobro (-) <<leise> (xxx)>
M: članove (-) ↑a (.) i možda bolje (--) čla= (-)
S: aha (.) SVOje porodice (-) ups (-) da
M: (-) ili tvoje (---) kako oslovljavaš [članove svoje po]
S: [pa ti: (--) svoje] dadada
M: svoje (-) (svoje) (-) mislim možeš i da kažeš tvoje nije tako strašno
S: <<p> dobr↑o>

Transcribing inconsistencies, mixing standards and typing errors were inevitable, since the transcripts were made without using transcription software that could control the syntax of GAT conventions.[2] Same information was occasionally annotated with different types of parentheses, and annotations for uncertain segments or for transcript gaps were sometimes used for marking comments, and vice versa. For instance, in Example 1, the same parenthesis type was used to annotate uncertain words "(svoje)" [English: "your own"] as well as non-verbal events "(schreibt)" [English: "writes"]. In rare cases, annotations that are not mentioned in GAT were used (for instance: * - <), and some GAT conventions were used for annotating other things than those described in the manual (for example, transcribing "Brankice=e" instead of "brankice:" for marking that the last syllable is long). Metalinguistic information was mostly given in German, but sometimes also in Serbian (for instance: "smeje se" [Serbian] and "lacht" [German] for annotating laughter).

3 Corpus compilation

3.1. Preprocessing

In order to convert the raw text files to an XML format, we first normalised the white space inconsistencies, and deleted the symbols that are irrelevant for further processing steps, such as the various codes for writing quotations marks (‘ ' ’ , , „ ” “). As shown in section 2.2., the inconsistencies were unsystematic, and correcting them automatically would have led to more errors in further processing steps. We wanted to ensure the consistency of the annotation in order to be able to process and categorise it in the next steps. Therefore, we used regular expressions to extract all the unique occurrences of annotations of a particular type, and saved them in separate files. We corrected and categorised the annotations in order to convert them more easily to TEI[3] conventions in the following steps (see Table 1).

Original annotation	English translation	Changes (intermediate step)
{Auslassung 14:58-15:53}	*omission 14:58-15:53*	((gap:extent: 55s))
((Exkurs über Mathe-Lehrerin nicht transkribiert))	*digression about the math teacher was not transcribed*	((gap:reason: deo o učiteljici matematike nije transkribovan))
{Telefon klingelt}	*the phone is ringing*	((incident: zvoni telefon))
((klopft auf den Tisch))	*knocking on the table*	((incident: kuca o sto))

Table 1: Categorising comments in the preprocessing phase (excerpt)

Once each annotation has been checked, we replaced the original data with the corrections. Totally, 693 annotations were checked, out of which 604 have been changed. The most common corrections regarded annotations for metalinguistic comments and the use of the equals sign "=".

[2] See FOLKER transcription software: https://exmaralda.org/de/folker-de/ (09.09.2020).
[3] https://tei-c.org/release//doc/tei-p5-doc/en/html/TS.html (09.09.2020).

Since the interviews were meticulously transcribed, we aimed to keep as many annotations as possible for further processing. However, we decided not to consider the conventions of the "fine transcript" (Selting et al., 2009) in the first release of the corpus because they were either sporadically used (annotation of pitch peaks and accentuation) or required extensive manual corrections (annotations of loudness and speed). As shown in Table 1, in order to make sure that comments and tokens are written in the same language, we also translated transcriber's comments from German to Serbian.

3.2. XML conversion

We converted the preprocessed files into XML format following the TEI conventions for transcriptions of speech. The transcripts were segmented for each turn, and each word in a turn was segmented as well. In addition to lemmatised and normalised forms (@lemma, @norm), we provided MULTEXT-East morphosyntactic specifications (@msd) and universal pos tags (@pos), as shown in Example 2. We marked unclear segments, deletions, gaps, incidents, vocal elements, and pauses.

Example 2: XML version of a part of the excerpt shown in Example 1

```
<u who="#F2" xml:id="F2-u6">
   <w lemma="oslovljavati" "  pos="VERB" msd="Vmr2s" xml:id="F2-u6-w1">oslovljavaš</w>
   <pause length="middle" xml:id="F2-u6-p2"/>
   <w lemma="taj" pos="DET"  msd="Pd-nsn"  xml:id="F2-u6-w3">to</w>
   <w lemma="biti" pos="AUX"  msd="Var3s"  xml:id="F2-u6-w4">je</w>
   <del type="truncation" xml:id="F2-u6-w5">pra</del>
   <unclear>
      <w lemma="e" pos="INTJ"  msd="I"  xml:id="F2-u6-w6">e</w>
   </unclear>
   <w lemma="perfektivan" pos="ADJ"  msd="Agpmsny" xml:id="F2-u6-w7">perfektivni</w>
   <w lemma="glagol" pos="NOUN"  msd="Ncmsn" xml:id="F2-u6-w8">glagol</w>
</u>
```

3.3. Normalisation

In order to normalise the transcribed data, we detected and automatically replaced all the tokens in which standard forms were reduced (for instance, "išo" instead of "išao" [English: "went"]; "kolko" instead of "koliko" [English: "how much"], etc.) by comparing the transcribed tokens with the tokens in Serbian lexicon srLex (Ljubešić et al., 2016). We stored the normalised tokens in the @norm attribute. Then, we extracted a list of all the tokens that did not occur in srLex and checked them manually. Out of 387 types, 119 were correct, although they were not present in our lexicons (mostly uncommon words, proper names, or slang expressions). The remaining 268 types were either (lowercase) proper names or reduced standard forms, which we corrected and marked as @norm, or they were due to orthographic and typing errors like "osnačavaju" instead of "označavaju" [English: "they mark"], which we marked as original transcriptions (@orig). In total, 3,824 tokens (2.2%) and 962 types (9.6%) were affected by the normalisation.

3.4. Tagging the corpus

For automatic annotation of the (normalised) corpus with morphosyntactic and lemma information we used the state-of-the-art tagger for Serbian and other South-Slavic languages CLASSLA-StanfordNLP (Ljubešić and Dobrovoljc, 2019)[4] - a fork of the StanfordNLP tagger with a series of improvements, especially on the lemmatisation level. The estimate of the accuracy on standard data of this tagger for Serbian is 95.23 F1 for morphosyntax and 97.89 F1 for lemmatisation.[5] However, given that this corpus consists of spoken data transcriptions, we annotated the corpus with a novel model trained on a union of all available training data for Serbian and Croatian, namely the SETimes.SR corpus of newspaper texts (Batanović et al., 2018), the hr500k Croatian reference training corpus (Ljubešić et al., 2016), the ReLDI-NormTagNER corpus of Serbian and Croatian tweets (Miličević and Ljubešić, 2016), and the RAPUT corpus of Croatian non-professional writing (Štefanec et al., 2016).

[4] https://github.com/clarinsi/classla-stanfordnlp (09.09.2020).
[5] https://github.com/clarinsi/babushka-bench (09.09.2020).

4. Conclusion and future work

We presented a number of processing steps needed for turning a collection of raw transcripts into a standardised spoken language corpus accessible to a wider community. These steps consisted in a) resolving the inconsistencies in the original transcripts, b) converting the transcripts into XML, c) enriching the data with normalisation and part-of-speech annotation, and d) publishing the corpus. The corpus can be used for investigating peculiarities of spoken Serbian in talk-in-interaction, for studying disfluencies in spontaneous speech, as well as for studying Serbian spoken as native and foreign language. We are currently working on the alignment of corpus turns with the respective audio-segments. The original transcripts contain some additional mark-up that has been left out in the current XML version, such as the annotation of loudness and speed. The integration of these items, as well as the correction of automatic annotations, is left for future work.

5. References

Vuk Batanović, Nikola Ljubešić and Tanja Samardžić. 2018. SETimes.SR – A Reference Training Corpus of Serbian. In: *Proceedings of the Conference on Language Technologies & Digital Humanities 2018 (JT-DH 2018)*, pages 11-17, Ljubljana, Slovenia.

Tomaž Erjavec. 2012. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Lang Resources & Evaluation*, 46(1):131–142.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC - Web Corpora of Bosnian, Croatian and Serbian. In: *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29-35, Gothenburg, Sweeden.

Nikola Ljubešić, Filip Klubička, Željko Agić and Ivo-Pavao Jazbec. 2016. New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4264-4270, Portorož, Slovenia.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. In: *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29-34, Florence, Italy.

Jelena Kuvač Kraljević and Gordana Hržica. 2016. Croatian Adult Spoken Language Corpus (HrAL). *Fluminensia: Journal for philological research*, 28(2):87-102.

Maja Miličević and Nikola Ljubešić. 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.

Margret Selting, Peter Auer, Birgit Barden, Jörg Bergmann, Elizabeth Couper-Kuhlen, Susanne Günthner, Uta Quasthoff, Christoph Meier, Peter Schlobinski and Susanne Uhmann. 1998. Gesprächsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte* 173, 91-122.

Margret Selting, Peter Auer, Dagmar Barth-Weingarten, Jörg Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, Peter Gilles, Susanne Günthner, Martin Hartung, Friederike Kern, Christine Mertzlufft, Christian Meyer, Miriam Morek, Frank Oberzaucher, Jörg Peters, Uta Quasthoff, Wilfried Schütte, Anja Stukenbrock and Susanne Uhmann. 2009. Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion,* (10):353-402.

Siniša Suzić, Stevan Ostrogonac, Edvin Pakoci and Milana Bojanić. 2014. Building a Speech Repository for a Serbian LVCSR System. *Telfor Journal,* 6(2):109-114.

Vanja Štefanec, Nikola Ljubešić and Jelena Kuvač Kraljević. 2016. Croatian Error-Annotated Corpus of Non-Professional Written Language. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016),* pages 3220-3226, Portorož, Slovenia.

Darinka Verdonik, Iztok Kosem, Anna Zwitter Viter, Simon Krek and Marko Stabej. 2013. Compilation, transcription and usage of a reference speech corpus: The case of the Slovene corpus GOS. *Language Resources and Evaluation,* 47(4):1031-1048.

Teodora Vuković, Nora Muheim, Olivier Winistörfer, Ivan Šimko, Anastasia Makarova and Sanja Bradjan. 2019. Corpora and Processing Tools for Non-Standard Contemporary and Diachronic Balkan Slavic. In: *Proceedings of the Student Research Workshop (RANLPStud 2019)*, pages 62-68, Varna, Bulgaria.

Sonja Ulrich. 2018. *Anredeformen im Serbischen*. Slavistische Beiträge (508), Wiesbaden.