EXTENDING THE SPOKEN SLOVENIAN TREEBANK

Kaja DOBROVOLJC,[1, 2]
[1]Faculty of Arts, University of Ljubljana
[2]Department for Artificial Intelligence, Jožef Stefan Institute

This paper presents a new version of the Spoken Slovenian Treebank (SST), a balanced and representative collection of transcribed spontaneous speech with manually annotated lemmas, part-of-speech tags, morphological features, and syntactic dependencies. The original version of the SST treebank was expanded with over 3,000 newly annotated utterances, and enhanced in terms of the consistency of transcriptions and the quality of the annotations. After a brief overview of the data sampling procedure and the semi-automatic morphological annotation, the core of the paper focuses on the the dependency annotation campaign, and the resolution of the discrepancies in sentence segmentation, capitalization and punctuation between the original and the newly added transcriptions. Finally, we summarize the contents of the new treebank with respect to its size and diversity, and evaluate it against the reference SSJ treebank of written Slovenian, highlighting the unique lexical and morphosyntactic characteristics of spoken communication.

Ključne besede: corpus annotation, dependency treebank, spontaneous speech, Slovenian language, Universal Dependencies

1 INTRODUCTION

Spoken language treebanks, i.e. syntactically annotated collections of transcribed speech, represent one of the fundamental language resources for data-driven spoken language research in both linguistics (e.g. Hinrichs & Kübler 2005; Pietrandrea and Delsart, 2019; van der Wouden et al. 2003) and nat-ural language processing (e.g. Braggaar & van der Goot (2021); Caines et al. 2017; Liu and Prud'hommeaux, 2021). Consequently, many spoken lan- guage treebanks have been developed over the recent decades, such as the Switchboard corpus for English (Godfrey et al., 1992), CGN for Dutch (van der Wouden et al., 2002), PDTSL for Czech (Hajič et al., 2008), NDC and LIA for Norweigian (Kåsen et al., 2022; Øvrelid et al., 2018), Rhapsodie for French

(Lacheret-Dujour et al., 2019), as well as the multilingual Verbmobil (Hinrichs et al., 2000) and CHILDES (MacWhinney, 2014) collections. Recently, many such treebanks have emerged as part of the expanding multilingual Universal Dependencies (UD) dataset (de Marneffe et al., 2021; Dobrovoljc, 2022).

For Slovenian, the Spoken Slovenian Treebank (SST) (Dobrovoljc & Nivre, 2016) has been the only language resource of this kind to date. To support computational and corpus linguistic research alike, the SST treebank was designed as a representative sample of the GOS reference corpus of spoken Slove-nian (Verdonik et al., 2013; Zwitter Vitez et al., 2021) and features manually an-notated transcriptions on the levels of lemmatization, MULTEXT-East morpho-logical tags and morphosyntactic annotations following the aforementioned UD annotation scheme, which includes cross-lingually comparable annotations of part-of-speech categories, morphological features and syntactic dependen-cies (Figure 1). As such, the treebank complements the SSJ reference treebank of written Slovenian (named after the *Sporazumevanje v slovenskem jeziku* project), which features identical annotations (Arhar Holdt et al., 2024; Dobrovoljc et al., 2017; Dobrovoljc & Ljubešić, 2022), and has already been used as the main data source for the development of specialized computational models for grammatical annotation of spoken Slovenian (Dobrovoljc & Martinc, 2018; Krsnik & Dobrovoljc, 2024; Verdonik et al., 2024).

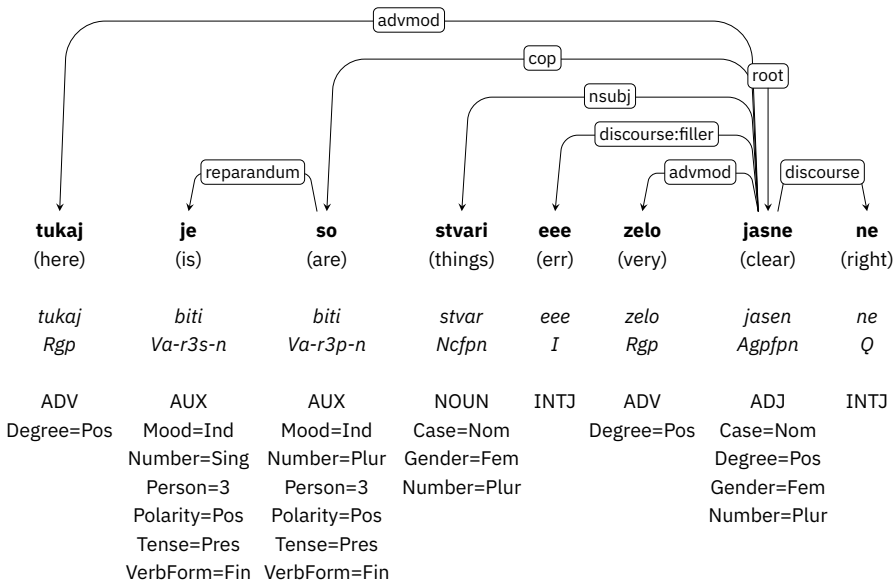
To alleviate the shortcomings of the original version of the SST treebank, such as its relatively small size (approximately 3,100 parsed utterances amounting to 30,000 annotated tokens), and diverse, but fragmented data (short samples of many speech events), the ongoing project SPOT (*Treebank-driven approach to the study of Spoken Slovenian*, ARIS grant no. Z6-4617),[1] aims at extending the treebank with a minimum of 50,000 new tokens. Consequently, the treebank was recently extended to more than triple its original size, by expanding some of the original data samples and adding completely new data from the recently expanded version of the reference corpus – GOS 2 (Verdonik et al., 2024).

We describe this major improvement of the SST treebank in the continuation of this paper by giving a brief summary of the data sampling procedure in Section 2 and describing the new data annotation and final dataset consolidation in

[1] https://spot.ff.uni-lj.si/

Section 3. We provide an overview of the resulting language resource in Section 4 and give details on its format and availability in Section 5. Finally, we present a comparison of the new SST treebank to the SSJ treebank of written Slovenian in Section 6 to exemplify its value for further empirical investigations of lexical and grammatical characteristics of Slovenian speech.

Figure 1: Example of a grammatically annotated utterance in the SST treebank featuring UD syntactic annotations (top), part-of-speech tags and morphological features (bottom), as well as MULTEXT-East lemmas and morphosyntactic tags (italics).



2 CORPUS EXTENSION

To address the aforementioned disadvantages of the original SST corpus, our aim was to extend the original SST treebank by a minimum of 50,000 new tokens while maintaining its representativeness with respect to the (updated) GOS 2 reference corpus of spoken Slovenian (Verdonik et al., 2024).

The data sampling procedure was designed in collaboration with the Mezzanine[2] project and is described in more detail by Verdonik et al. (2024). In summary, the sampling was conducted through a manual selection of specific speech events from the GOS 2 corpus (Verdonik et al. 2024)[3] and was performed in two steps. First, 22 samples from GOS 1 events in the original SST corpus were expanded with approximately 450 additional words per event, resulting in about 10,000 new words in total from the GOS 1 subset. Second, 57 entirely new speech events from the ARTUR subset were added, each contributing approximately 800 new words, totalling to around 40,000 new words from the ARTUR subset. The ex- act counts, which also account for the post-festum modifications of the data described in the following sections, are reported in Section 4 (Table 2).

From the perspective of subsequent syntactic annotation of the data, an important drawback of the sampled ARTUR subset was the presence of very short segments, with segment breaks introduced after each pause rather than after the completion of a semantically and syntactically complete unit of speech, as was the case for the utterance segmentation in GOS 1. To resolve this, the ARTUR subset was automatically resegmented based on the sentence-final punctuation markers available (for details and example see Verdonik et al. 2024)), which resulted in more coherent structures for subsequent syntactic analysis. The resegmentation was performed as part of the conver- sion of the newly sampled data (originally in XML TEI) to CONLL-U, which was also the file format we used in the continuation of our work presented below.

3 TREEBANK ANNOTATION

Following the data sampling and pre-processing steps presented above, the resulting new dataset was manually annotated for lemmas, morphological features and syntactic dependencies.

[2] https://mezzanine.um.si/

[3] The GOS 2 corpus consists of the original GOS 1 corpus (Zwitter Vitez et al., 2021), GOS VideoLectures corpus (Verdonik et al., 2021) and selected events from the ARTUR ASR database (Verdonik, Bizjak, Sepesy Maučec, et al., 2023).

3.1 Lemmatization and Morphology Annotation

In the first stage, the two new subsets presented in Section 2 have been semi-automatically annotated for lemmas and morphosyntactic tags in accordance with the MULTEXT-East annotation scheme (Erjavec, 2010; Holozan et al., 2023), which is the most widely used annotation scheme for Slovenian corpora. The process is described in detail by Čibej and Munda (2024), who also discuss the annotation issues related to the newly emerged speech-specific lexical and morphological phenomena.

The resulting morphologically annotated dataset was then converted to UD part-of-speech categories and morphological features using the jos2ud conver-sion pipeline (Dobrovoljc et al., 2017).[4] The conversion features a large num-ber of high-accuracy mapping rules and has previously been used for mapping MULTEXT-East tags to UD morphology in other reference resources for Slove-nian, such as the ssj500k (Krek et al., 2021) and SUK (Arhar Holdt et al., 2022) training corpora of standard written Slovenian, the Janes-Tag corpus of non-standard written Slovenian (Lenardič et al., 2022), and the Sloleks lexicon of inflected forms (Čibej et al., 2022).

In the second stage, the transcriptions have been syntactically parsed according to the UD annotation scheme through a semi-automatic procedure described below.

3.2 Automatic Dependency Parsing

Following the nowadays prevailing approach to manual data annotation, the transcriptions have first been pre-annotated using an automatic parser. To select the optimal tool for the task, several models have been developed and evaluated. For parsing spoken Slovenian in particular, the SLOKIT[5] project has recently produced a specialized model of the CLASSLA-Stanza tool (Ljubešić & Dobrovoljc, 2019; Terčon & Ljubešić, 2023). Following the findings by Dobrovoljc and Martinc (2018), the model was trained on a concatenation of spoken (SST) and written (SSJ) data and produced better results than the CLASSLA-Stanza parsing models trained on either written or spoken data alone

[4] https://github.com/clarinsi/jos2ud
[5] https://slokit.ijs.si/

(Verdonik et al., 2024), confirming the positive effect of the larger training set.

Given the recent technological advancements, we extended this work by producing three additional models using the Trankit transformer-based tool (Nguyen et al., 2021), trained on the written SSJ and spoken SST treebanks, as released in UD v2.12 (Zeman et al., 2023), and the combination of the two. Thus, five parsing models have been evaluated with respect to the standard evaluation metric of labelled-attachment score (LAS), which gives the percentage of tokens with correctly predicted parent node and the type of their relation:[6]

- CLASSLA-Stanza default model for written Slovenian (Terčon & Ljubešić, 2023), trained on SSJ
- CLASSLA-Stanza SLOKIT model for spoken Slovenian, trained on SSJ and SST
- Trankit model for written Slovenian (Krsnik & Dobrovoljc, 2023), trained on SSJ
- Trankit model for spoken Slovenian, trained on SST
- Trankit model for spoken Slovenian (Krsnik & Dobrovoljc, 2024), trained on SSJ and SST

Table 1 shows the models' performance on both written (SSJ) and spoken (SST) test set, featured in the same dataset release.[7] Our results confirm previous findings that, regardless of the tool, the performance of the standard models trained on written data drops significantly when confronted with transcribed speech, and increases significantly when spoken data is featured in the training (approx. +15pp LAS for both joint SSJ+SST models). However, the transformer-based Trankit models display a much higher performance overall (both in written and spoken testing scenarios). Therefore, for the use case at hand, the best-performing Trankit SSJ+SST model (81.26 LAS F1) was chosen for the automatic pre-annotation of the newly added SST data (Section 2).

[6] The SLOKIT and SST-only Trankit model have not been officially released, but are available directly from the authors.

[7] The evaluation is performed on pre-tokenized (gold) test sests to neutralize the impact of speech segmentation–a notoriously difficult task if no sentence-final punctuation is available in the transcripts (see Dobrovoljc and Martinc, 2018).

Table 1: LAS F1 performance of selected parsing models on the SSJ and SST test sets.

Model	SSJ-test (written)	SST-test (spoken)
CLASSLA-Stanza (written)	90.64	55.43
CLASSLA-Stanza Slokit (written+spoken)	88.64	70.58
Trankit SSJ (written)	95.39	66.36
Trankit SST (spoken)	74.83	79.84
Trankit SSJ+SST (written+spoken)	95.47	**81.26**

3.3 Manual Dependency Annotation

The automatically parsed dataset with manually revised lemmatization and morphology was then split into document-level files (79 in total), with 2–3 independent annotators assigned to each file. The annotation was performed in the Q-CAT annotation tool (Brank, 2023), which was upgraded for this particular campaign to also enable listening of audio files, provided the URLs to the audio files are given as part of the `# sound_url` comment line in the input CONLL-U file. Given that Q-CAT does not support comparison of annotations produced by different annotators, the curation process was carried out through the WebAnno annotation service maintained by CLARIN.SI (Yimam et al., 2013; Erjavec et al., 2016). Given the fact that the original SST was annotated by a single annotator and some annotation guidelines have been changed, the original SST was also manually revised.

3.4 Annotation Guidelines

In addition to the UD guidelines available online,[8] which mostly include robust language-independent definitions and a limited set of illustrative examples, especially for speech-specific phenomena, the annotators were instructed to use the stand-alone manual for UD annotation of Slovenian texts (Dobrovoljc & Terčon, 2023). This document was originally published within the DSDE project to document the annotation of the written SSJ UD dataset (Dobrovoljc et al., 2023; Dobrovoljc & Ljubešić, 2022) and was now upgraded to also document the guidelines for spoken data annotation. The latter are based on the (sparsely documented) annotation of the original SST treebank (Dobrovoljc &

[8] https://universaldependencies.org/guidelines.html

Nivre, 2016), as well as the more recent practices and discussions within the community (Dobrovoljc, 2022; Kahane et al., 2021).

Due to space limitations, we only describe here how the two most typical speech-specific phenomena are annotated: discourse markers (Section 3.4.1) and speech repairs (Section 3.4.2). For discussions of other speech-specific morphosyntactic phenomena, the readers are advised to refer to the full documentation in the aforementioned guidelines (Dobrovoljc & Terčon, 2023)[9] or the discussions in papers by Dobrovoljc and Nivre (2016) and Dobrovoljc (2022, 2024).

3.4.1 DISCOURSE MARKERS

According to the general UD guidelines, the *discourse* relation is used for interjections and other discourse particles and elements which are not clearly linked to the structure of the sentence, except in an expressive way. These include interjections (e.g. *oh*), fillers (e.g. *eee* 'err'),[10] and discourse markers in the narrow sense (*no* 'well', *a ne* 'right'). Figure 2 illustrates a tree involving two such typical expressions and shows that they attach to the head of the most relevant clause (usually the root predicate), even though they are not dependent of the predicates as such.

If an utterance consists of discourse elements only, the most prepositionally loaded marker (i.e. informative, content-rich) is chosen as the head node, as is the case with the feedback response *dobro* in Figure 3. If it is not possible to determine the most semantically salient expression, the first element in the sequence is treated as the head.

[9] The final version of the Slovenian UD guidelines for both written and spoken language annotation is planned to be published in September 2024 at https://wiki.cjvt.si/books/07-universal-dependencies-FPQ/page/annotation-guidelines.

[10] For filled pauses, we introduce a special *discourse:filler* label extension (relation sub-type), as illustrated by *eee* in Figure 2.

Figure 2: Annotation of discourse markers.

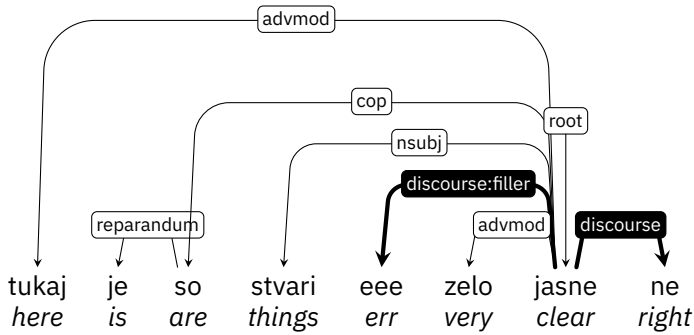
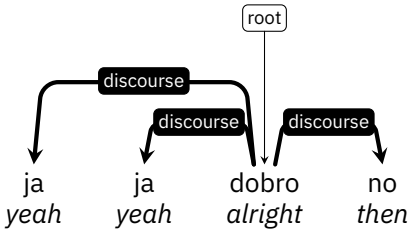


Figure 3: Annotation of a sequence of discourse markers.



3.4.2 SELF-REPAIRS

The *reparandum* relation is used to annotate self-repairs in speech, i.e. instances where a speaker replaces previously uttered content with a new one, as illustrated in Figure 1, where the singular form of the copula verb *je* is replaced by the correct plural form *so*.

The repaired unit can be syntactically complete or incomplete, such as unfinished words, phrases or clauses. In case of shared dependants between the reparandum and its repair, such as modifiers applicable to both, the dependent is attached to the repair rather than the reparandum. This is illustrated in Figure 4 below, which shows the premodifier *kako* being attached to the noun *orožje* rather than the first, unfinished attempt of pronouncing the word (*orož-*).[11] In case of a sequence of self-repairs–for example, when a speaker

[11]This design principle enables the sub-trees spanning from *reparandum*-marked tokens (i.e. disfluencies) to be easily removed without causing the remaining tree to become ungrammatical or

keeps restarting their intended verbalization–all reparandums attach to the same head, i.e. the head of the final repair (Figure 5).

Figure 4: Annotation of self-repairs with shared dependants in SST.

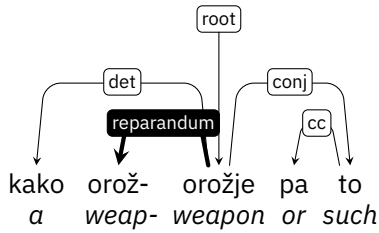
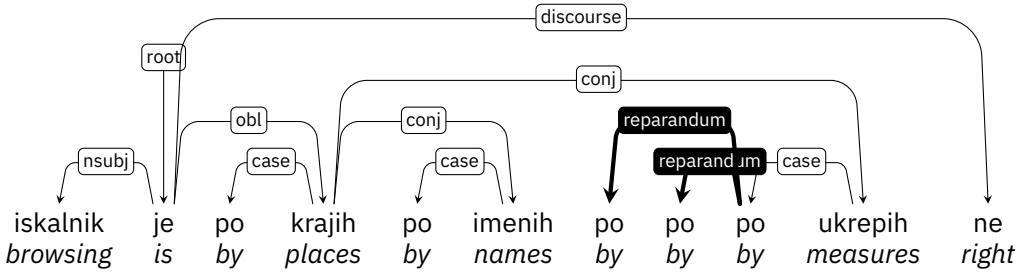


Figure 5: Annotation of a sequence of self-repairs in SST.



3.5 Final Subset Consolidation

Finally, both manually revised datasets (original SST and new data from GOS 2) have been merged and consolidated with respect to metadata formatting (see Section 5.2), but also transcription consistency, as different punctuation and casing principles have originally been adopted in GOS 1 and ARTUR corpora. Some corrections to the lemmatization and morphological layers have also been introduced.

semantically underspecified (e.g. *kako orožje pa to* in Figure 4 and *iskalnik je po krajih po imenih po ukrepih ne* in Figure 5).

3.5.1 ADDITION OF PUNCTUATION

While GOS 1 transcriptions include only sentence-final markers of question (?) and exclamation (!) intonation, ARTUR features written-like punctuation in both sentence-medial (e.g. commas) and sentence-final positions (e.g. full stops). To ensure dataset consistency across both subsets and comply with the general tendency to include punctuation in similar spoken language treebanks (Dobrovoljc, 2022), sentence-medial and sentence-final punctuation has been added to the GOS 1 subset. This was performed through a semi-automatic approach, in which the GOS 1 transcriptions were first automatically punctuated using the Slovene Punctuator[12] tool and then manually checked so as to conform to the punctuation principles of the ARTUR database (Verdonik & Bizjak, 2023). In total, 12,732 punctuation symbols have been added.

In parallel, GOS 1 transcriptions have also been stripped of non-lexical tokens (annotated as punctuation in the original SST treebank), such as *[audience:laughter]* and *[pause]*, which–with the exception of the latter–have not been transcribed in ARTUR. The new consolidated SST treebank contains transcriptions that are more similar to written text than those in the original treebank, as they include punctuation and exclude other markers of prosody. However, this change in the underlying data does not hinder the array of research applications, since non-lexical phenomena can still be accessed from the transcriptions of the reference GOS corpus if necessary.

3.5.2 CORRECTION OF TRANSCRIPTIONS

The process of final data consolidation also included the correction of the erroneously transcribed (standardized) tokens that were identified by the annotators or signalled as a mismatch in the data validation phase using the official UD validator.[13] This includes corrections of erroneous capitalisation at the beginning of the sentences, resulting from the automatic casing unification applied to the original GOS 2.1 (Verdonik et al., 2023), which aimed at lowercasing all words except for named entities. Transcription mistakes per-taining to tokenization, such as words that should either be split or merged,

[12] https://github.com/clarinsi/Slovene_punctuator
[13] https://github.com/UniversalDependencies/tools/

were not tackled in this iteration, as changes in the tokenization would impede the automatic mapping to the reference corpus and its derivatives.

3.5.3 CORRECTION OF MORPHOLOGY

The aforementioned data validation errors also highlighted some mistakes and inconsistencies in lemmatization and morphological annotation within both schemes, which were also resolved. In addition, some UD morphological annotations have been consolidated based on the final annotation guidelines, such as the categorization of colloquial expressions *kao* 'like' (PART), and *ene* 'about' (ADV), definite article *ta* 'the' (DET with no inflectional features), indefinite article *en* 'a' (DET) and anonymized names (PROPN with no inflectional features).

4 NEW SST TREEBANK OVERVIEW

This section presents the contents of the new SST treebank with respect to size (Section 4.1) and the diversity of the spoken data included (Section 4.2).

4.1 Treebank Size

As shown in Table 2, the resulting new, extended and revised, SST treebank based on approximately 10 hours of transcribed speech includes 344 unique speech events (documents) with a total of 6,108 utterances and 98,393 tokens. In comparison to the previous edition of the treebank (prior to the revisions presented in this paper),[14] the new SST treebank includes more than triple the number of transcribed tokens (+334%) and almost double the number of utter-ances (+196%), as well as a more varied set of events (+ 11%) and speakers (+ 11%). The average length of a (sampled) document has been extended from an average of 103 tokens per document to 286 tokens per document.

4.2 Data Diversity

At the same time, the new SST treebank remains representative with respect to the reference GOS 2.1 and, indirectly, to Slovenian speech in general, as

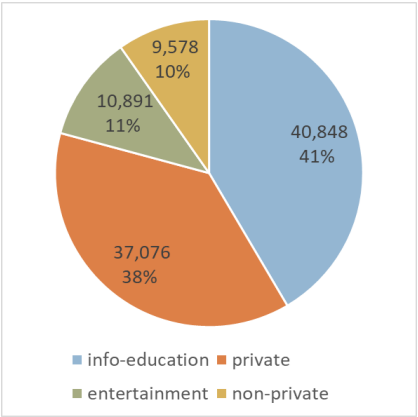
[14] The original version of the SST treebank (Dobrovoljc & Nivre, 2016) featured 287 events, 594 speakers, 3,188 utterances and 29,488 tokens.

Table 2: Overview of the new SST treebank and its subsets.

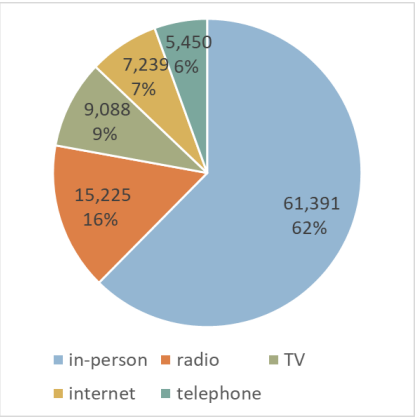
Subset	*Events*	*Speakers*	*Utterances*	*Tokens*
SST-2016-revised	287	594	2,903	36,960
New from GOS 1	22	61	1,236	13,112
New from ARTUR	57	72	1,969	48,321
SST-2024 (UD 2.15)	344	676	6,108	98,393

shown in Figures 6 to 9, which report the number of tokens per different types of speech events,[15] communication channels and speaker demographics.

Figure 6: Number of tokens in SST with respect to the nature of speech event.



(a) Event type



(b) Communication channel

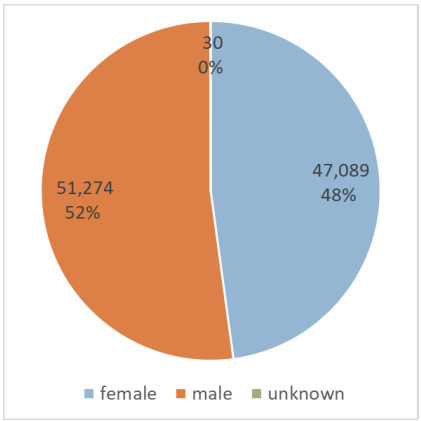
5 TREEBANK RELEASE

The new SST treebank is planned to be released as part of the official UD release v2.15 in November 2024.[16] It is freely available under the CC-BY license,

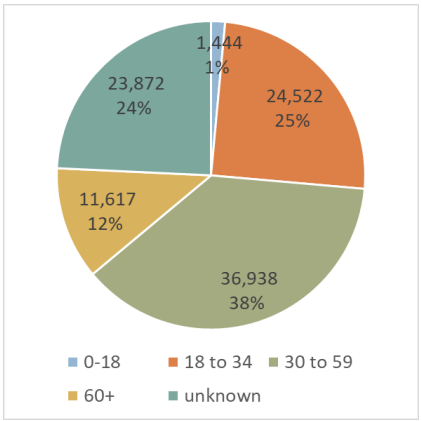
[15] Generally, all events feature spontaneous speech, i.e. unscripted verbal communication that occurs naturally in real-time, albeit with varying amounts of planning in public and non-public situations. A more detailed characterisation of speech events can be retrieved from the metadata available in the reference GOS 2 corpus.

[16] An interim version with extensions but no punctuation has already been published as part of UD release v2.14 in May 2024 (Zeman et al., 2024).

Figure 7: Number of tokens in SST with respect to speaker gender and age.

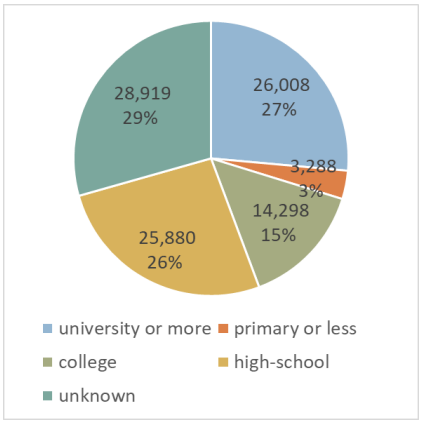


(a) Gender

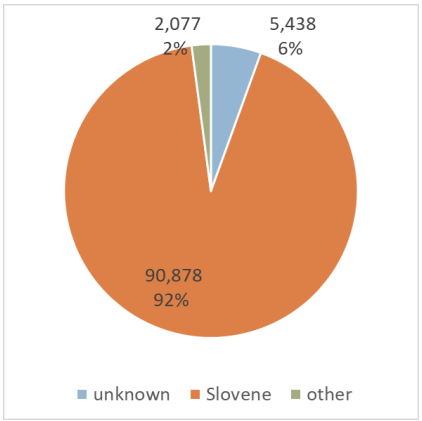


(b) Age

Figure 8: Number of tokens in SST with respect to speaker education and first language.



(a) Education



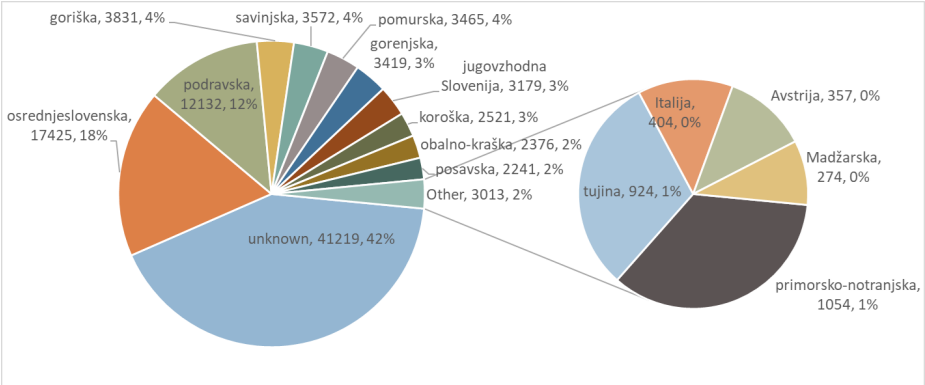
(b) First language

which is a less restrictive license in comparison to the CC-BY-NC license of the original version of the dataset, which prohibited commercial use.

5.1 Data Split

As required by the UD dataset release protocol, the treebank was split into training, development and test set with approximately 80%, 10% and 10% to-

Figure 9: Number of tokens in SST with respect to the region of speaker residence.



ken distribution in each. As with the original SST treebank, the split has been randomised on document-level, which ensures an equal distribution of the different event and speaker types (Section 4) across all three datasets. It was also ensured that the train, test and dev data from the original SST version was preserved in the same subset, so to enable fair model comparisons across different versions of the SST dataset.

5.2 Format

The treebank is encoded in the standard CONLL-U format,[17] illustrated in Figure 10, where each token in a sentence is represented on a single line with 10 fields: ID (token index), FORM (word form), LEMMA, UPOS (universal part-of-speech), XPOS (language-specific tag, i.e. MULTEXT-East), FEATS (morphological features), HEAD (index of the head token), DEPREL (dependency relation to the head), DEPS (enhanced dependency graph, not used in SST), and MISC (miscellaneous information).[18]

Speech-specific extensions of the format pertain to the comment lines, which include information on the document, sentence ID, speaker ID, and the audio

[17] https://universaldependencies.org/format.html

[18] Due to space limitations, the CONLL-U example in Figure 10 only shows the first feature in the FEATS column (but see the example in Figure 1) and omits the MISC column (e.g. `pronunciation=tuki|GOS2.1_token_id=GOS119.tok1104`).

URL,[19] as well as to the last (miscellaneous) column, which includes information on the pronunciation-based spelling of the word form (e.g. *tko* for the standardized word form *tako* 'such') and the token/segment IDs pertaining to the original GOS 2.1 corpus.

This ensures that all other types of metadata pertaining to the recorded event and speakers involved can easily be retrieved from the reference GOS 2.1 corpus via the persistent and traceable IDs. This includes retrieving all other relevant information omitted from the final SST treebank, such as the placement of pauses, non-vocal sounds or other types of transcribed but syntactically less relevant non-lexical phenomena.[20]

Figure 10: Example of an annotated utterance in the CONLL-U format.

```
# newdoc_id = GOS119
# sent_id = GOS119.s72
# speaker_id = Bm-gost-07155
# sound_url = https://nl.ijs.si/project/gos20/GOS119/GOS119.s72.mp3
# text = tukaj je so stvari eee zelo jasne ne
1   tukaj   tukaj   ADV     Rgp     Degree=Pos     7   advmod            _   _
2   je      biti    VERB    Va-r3s-n Mood=Ind...   3   reparandum        _   _
3   so      biti    AUX     Va-r3p-n Mood=Ind...   7   cop               _   _
4   stvari  stvar   NOUN    Ncfpn   Case=Nom...    7   nsubj             _   _
5   eee     eee     INTJ    I       _              7   discourse:filler  _   _
6   zelo    zelo    ADV     Rgp     Degree=Pos     7   advmod            _   _
7   jasne   jasen   ADJ     Agpfpn  Case=Nom...    0   root              _   _
8   ne      ne      PART    Q       Polarity=Neg   7   discourse         _   _
```

5.3 Online Access

In addition to the official SST dataset release in CONLL-U, which is also available on GitHub,[21] the SST treebank can also be accessed for browsing and

[19] For resegmented ARTUR-based data (see Section 2), the links in # sound_url point to a concatenation of the audio files available for the original GOS 2.1 segments. In the rare instance where an original ARTUR segment was split two SST segments, the original audio file appears in both concatenations. As a result, some linked audio files might include longer spans of speech than what is actually featured in the transcribed utterance.

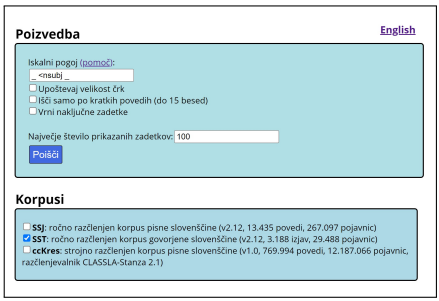
[20] This includes the audio recordings of the events, which are freely available under CC-BY for the ARTUR subset (Verdonik, Bizjak, Žgank, et al., 2023), and for research purposes for the GOS 1 subset.

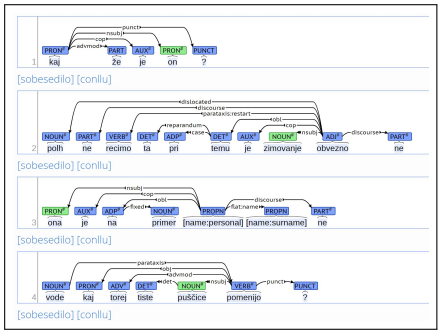
[21] https://github.com/UniversalDependencies/UD_Slovenian-SST

analysis through the numerous tools that support querying and visualising UD treebanks worldwide. Online services with regular data updates include Grew-match,[22] maintained by INRIA Nancy, and INESS,[23] maintained by CLARINO. An important advantage of the former is the fact that it also supports listening to audio recordings in treebanks featuring spoken data.

The latest version of the SST treebank has also been uploaded to the locally developed Drevesnik treebank-querying service (Štravs & Dobrovoljc, 2024),[24] which is based on the open-source dep_search tool (Luotolahti et al., 2017). In addition to featuring other manually and automatically parsed UD corpora for Slovenian, the main advantage of the service (illustrated in Figure 11) from the perspective of Slovenian users is that it features a powerful and easy-to-use query language (documented in both English and Slovenian), enables regex-supported querying of the popular MULTEXT-East tags (XPOS column), ran-domisation of the results and their limitation to short sentences only (useful for illustrative or didactic purposes).

Figure 11: Drevesnik online service for querying Slovenian dependency treebanks (left: query interface, right: results interface).





The SST treebank also represents the backbone of the emerging ROG training corpus of spoken Slovenian (Verdonik, Dobrovoljc, Čibej, et al., 2024), which will feature additional annotation layers for disfluencies, dialogue acts, and prosody boundaries for some of the transcribed events, and will be encoded in other formats as well.

[22] https://universal.grew.fr/
[23] https://clarino.uib.no/iness
[24] https://orodja.cjvt.si/drevesnik/

6 COMPARISON WITH THE SSJ TREEBANK OF WRITTEN SLOVENIAN

Finally, we compare the new SST treebank with its written counterpart, the SSJ UD treebank of written Slovenian (Dobrovoljc et al., 2017), which has been annotated using the same annotation scheme and thus enables direct comparison of annotations on various levels. To neutralize the effect of punctuation tokens, adopting different functions in the representation of both modalities, the comparison is based on treebanks excluding punctuation. The results thus reflect the analysis of all uttered phenomena rather than all transcribed phenomena.

6.1 Vocabulary

The comparison of the vocabulary in Table 3 shows that, despite the spoken SST treebank being much smaller than its written counterpart, there are as many as 5,242 unique words (39.5% of all word types in SST) and 2,293 (30.1%) unique lemmas featured in the SST treebank that do not occur in the written SSJ treebank, confirming previous findings on the unique lexical characteristics of spoken Slovenian (Verdonik & Maučec, 2016; Dobrovoljc, 2018).[25]

Table 3: Comparison of vocabulary diversity in spoken and written treebank.

	SST (spoken)	SSJ (written)
Words	76,341	227,619
Word types	13,268	48,570
Unique word types	5,242	40,544
Lemma types	7,617	25,352
Unique lemma types	2,293	20,028

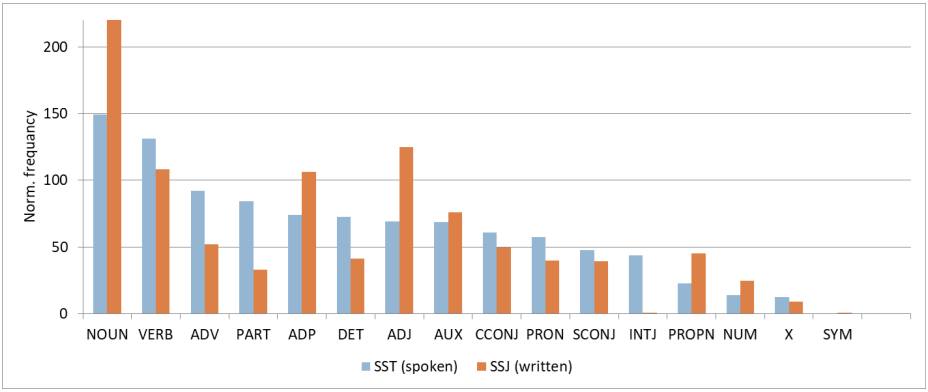
6.2 Part-of-Speech Categories

The comparison of part-of-speech tag frequencies per thousand words shown in Figure 12 reveals that the two modalities also differ with respect to the type of vocabulary used. For instance, spoken language exhibits a much higher fre-

[25] Examples of most frequent unique lemmas in SST include filled pauses (e.g. *eee*), response tokens (e.g. *aja*), anonymized names (e.g. *[name:personal]*), and colloquial expressions (e.g. *ke*), while most frequent unique lemmas in SSJ include roman numbers (e.g. *2*), abbreviations (e.g. *dr.*), acronyms (e.g. *EU*) and culturally obsolete vocabulary (e.g. *tolar*).

quency of word classes pertaining to interaction, subjectivity, deixis and modification, such as particles (PART), adverbs (ADV), interjections (INTJ), determiners (DET) and pronouns (PRON). The higher frequency of verbs (VERB) in spoken language also suggests a more dynamic narrative style, while a higher frequency of nouns (NOUN, PROPN), adjectives (ADJ) and prepositions (ADP) in written communication suggests a denser information structure and more descriptive content. Our findings confirm that spoken and written communication exhibit distinct tendencies towards nominal and verbal styles, aligning with Douglas Biber's seminal work on register variation (Biber, 1988; Biber et al., 2010).

Figure 12: Comparison of the distribution of POS categories in spoken (SST) and written (SSJ) treebank.



6.3 Dependency Relations

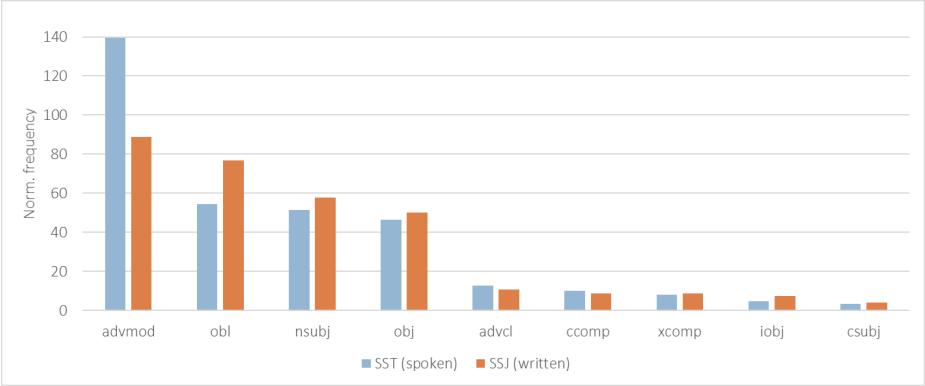
Finally, we compare the distribution of the dependency relations (syntactic functions of words) across the two datasets.

6.3.1 CORE DEPENDANTS OF PREDICATES

Figure 13 shows the comparison of the distribution of the predicate arguments, namely the nominal or clausal subjects (*nsubj*, *csubj*), objects (*obj*, *iobj*, *ccomp*) and adjuncts (*advmod*, *obl*, *advcl*). Interestingly, there are no major differences observed in the distribution of core arguments within each treebank,

confirming that similar clause pattern strategies are used in both modalities. However, the notable differences in the frequency of some relations in both treebanks confirm the aforementioned nominal-heavy nature of written communication, i.e. more nominal subject (*nsubj*), objects (*obj*, *iobj*) and adjuncts (*obl*) in the written SSJ treebank. At the same time, the clauses in spoken language contain a much higher percentage of adverbial modification (*advmod*),[26] which could be explained by the abundance of modal adverbials, which speakers use to express stance, convey attitude, and balance the interaction.

Figure 13: Comparison of core predicate arguments in the spoken (SST) and written (SSJ) treebank.



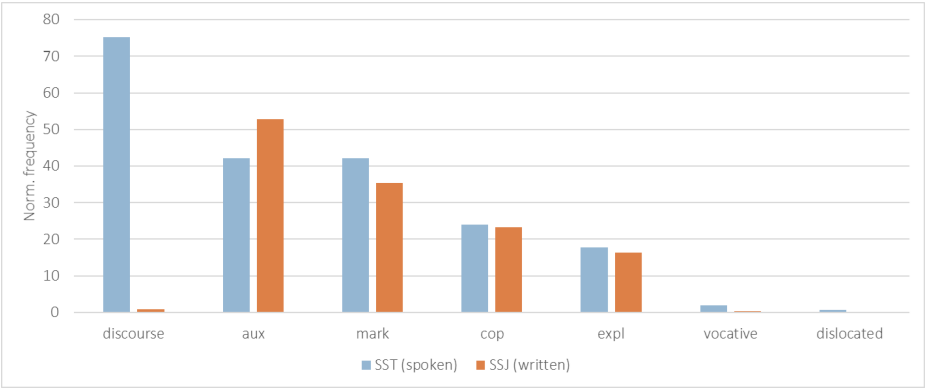
6.3.2 OTHER DEPENDANTS OF PREDICATES

In contrast to the much higher number of discourse elements (*discourse*), vocatives (*vocative*), and fronted or postponed elements (*dislocated*) in SST, which only rarely occur in written data, the differences in the distribution of other dependants of predicates are less pronounced, with two exceptions. First, spoken communication seems to show a preference for simple verbs phrases in the present tense (i.e. less auxiliary verbs marked with *aux*). Second, despite the very similar frequency of subordinate clauses in both modalities (*csubj*, *ccomp* and *advcl* in Figure 13 and *acl* in Figure 15), spoken data exhibits a higher num-

[26] The *advmod* relation is used both for modification of predicates (e.g. *Pride jutri.*) but also for modification of other modifier words, such as adjectives (e.g. *zelo umazana posoda*), so the number reflects both.

ber of subordinate conjunctions (*mark*). This might be explained by the frequency of insubordinate clauses used as independent utterance to respond or to build upon a previous utterance on context (e.g. replying *Ker dežuje.* 'Because it is raining.' to a question on why an event was cancelled).

Figure 14: Comparison of the non-core predicate arguments in the spoken (SST) and written (SSJ) treebank.



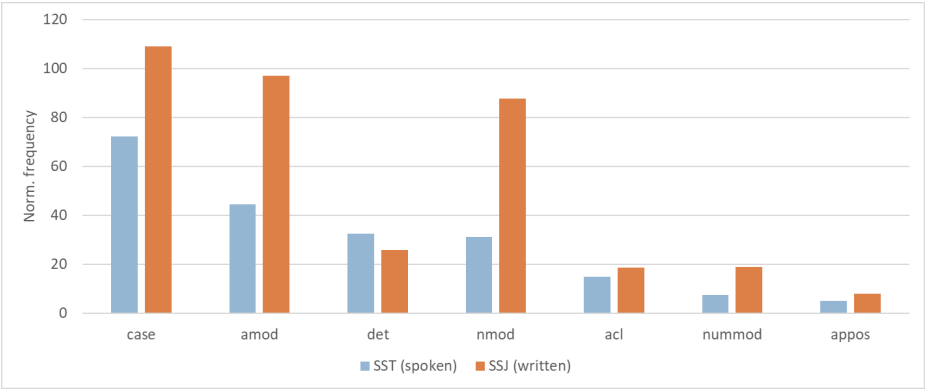
6.3.3 DEPENDANTS OF NOMINALS

The comparison of the distrubtion of the relations pertaining to the dependents of nominals (e.g. noun phrase constituents) in Figure 15 shows a lower frequency of modifiers of nouns, such as adjectival (*amod*), nominal and prepositional (*nmod*, *case*), numerical (*nummod*), clausal (*acl*) and appositional (*appos*) modifiers. This is in line with the aforementioned lower number of nominal phrases in speech (Figure 12), but also suggests an overall simpler structure of such phrases (i.e. less pre- and post-modification of nouns). The only exception to this rule is the higher frequency of determiners (*det*) in SST, which can be explained by the frequent use of demonstrative pronouns and other context-grounding deictical premodifiers in speech.

6.3.4 OTHER RELATIONS

Last, Figure 16 shows the comparison of the distribution for all other types of dependency relations that do not fall into any of the main syntactic categories

Figure 15: Comparison of the dependents of nominals in the spoken (SST) and written (SSJ) treebank.



mentioned above. Naturally, the biggest differences between both modalities can be observed for the (*reparandum*) relation pertaining to speech repairs, which only occur in the spoken treebank.

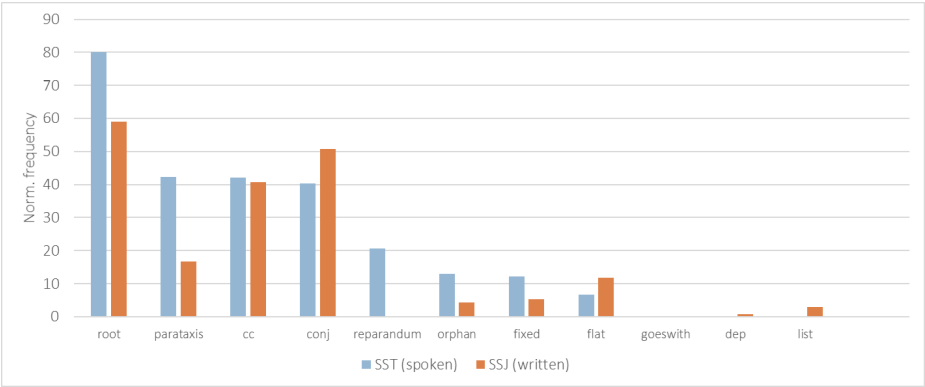
The second important observation is that sentences in speech are generally much shorter than in writing. This is not only reflected by the difference in the average number of words per utterance/sentence (i.e. the frequency of *root* elements in a treebank),[27] but also by the higher frequency of *parataxis* relation, which is used for run-on clauses with no linking conjunction.

Our results also confirm the elliptical nature of spoken communication, with SST exhibiting a higher frequency of *orphan* relations, which are used to mark core arguments in cases of predicate ellipsis. We can also observe that speech features a higher number of coordinating conjunctions (*cc*) in relation to the number of coordinating conjuncts (*conj*); however, the cause might be attributed to various reasons, such as a higher number of discourse-structuring devices in speech in general (see the higher frequency of subordinating conjunctions labeled as *mark* in Figure 14) or longer coordination phrases in writing (i.e. multiple conjuncts).

[27] Average sentence length without punctuation is 12.5 tokens per utterance in SST and 17 tokens per sentence in SSJ.

Last, SST treebank also features a larger number of *fixed* multi-word expressions, which is in line with previous findings on the formulaic nature of this type of communication (Dobrovoljc, 2018). On the other hand, flat multi-word expressions (mainly encompassing personal names and foreign named entities) occur less often in speech.

Figure 16: Comparison of all other relations in the spoken (SST) and written (SSJ) treebank.



7 CONCLUSION

In this paper, we presented the recent extension of the Spoken Slovenian Treebank with more than 3,000 new manually parsed utterances, resulting in a new, balanced and representative, version of the corpus to be used in linguistic, computational and other empirical investigations of spoken communication in Slovenian. We made a first step in this direction by comparing it to the SSJ treebank of written Slovenian, which revealed the unique lexical and morphosyntactic characteristics of spoken communication in comparison to writing. These findings relate to the interactive and situation-related nature of this type of language modality and further highlight the importance of integrating spoken language data into the Slovenian language resource landscape.

Short-term goals for future work include the integration of the treebank into the emerging multi-layer ROG corpus of spoken Slovenian, as well as the retraining and evaluation of state-of-the-art parsing models trained on the new

dataset. Most importantly, the new SST treebank is planned to be used as the main data-source for a corpus-driven analysis of speech-specific syntactic patterns within the SPOT project, which will complement the robust SSJ-SST comparison presented in this paper with a more sophisticated analysis of syntactic (sub-)trees encountered in both treebanks, by using the STARK tool (Krsnik et al., 2024). Finally, our long-term goal is also to ensure a continuous incremental improvement of the quality of this richly annotated corpus, as well as to promote and facilitate its usage in Slovenian corpus linguistics.

8 ACKNOWLEDGMENTS

This work was financially supported by the Slovenian Research and Innovation Agency through the research project *Treebank-Driven Approach to the Study of Spoken Slovenian* (Z6-4617) and the research program *Language Resources and Technologies for Slovene* (P6-0411). In addition to the collaborators from the Mezzanine project (J7-4642) who have been involved with the data sampling and morphological annotation (Jaka Čibej, Tina Munda, Nikola Ljubešić, Peter Rupnik, Darinka Verdonik), we also wish to thank the data annotators (Nives Hüll, Karolina Zgaga, Luka Terčon, Matija Škofljanec) and the technical collaborators who have contributed to data pre-annotation (Luka Krsnik), punctuation insertion (Iztok Lebar Bajec) and audio resegmentation (Janez Križaj, Simon Dobrišek, Tomaž Erjavec).

REFERENCES

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Erjavec, T., Gantar, P., Krek, S., Munda, T., Robida, N., Terčon, L., & ... Žitnik, S. (2024). SUK 1.0: A New Training Corpus for Linguistic Annotation of Modern Standard Slovene. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on com-putational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 15428–15435). ELRA and ICCL. https://aclanthology.org/2024 .lrec-main.1340

Arhar Holdt, Š., Krek, S., Dobrovoljc, K., Erjavec, T., Gantar, P., Čibej, J., Pori, E., Terčon, L., Munda, T., Žitnik, S., Robida, N., Blagus, N., Može, S., Ledinek, N., Holz, N., Zupan, K., Kavčič, T., Škrjanec, I., ... Zajc, A. (2022). *Training corpus SUK 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1747

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511621024

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2010). The Longman Grammar of Spoken and Written English. De Gruyter Mouton. https://www.degruyter.com/database/COGBIB/entry/cogbib.1245/html

Braggaar, A., & van der Goot, R. (2021). Challenges in Annotating and Parsing Spoken, Code-switched, Frisian-Dutch Data. In *Proceedings of the second work- shop on domain adaptation for nlp* (pp. 50–58). Association for Computational Linguistics. https://aclanthology.org/2021.adaptnlp-1.6

Brank, J. (2023). *Q-CAT Corpus Annotation Tool 1.5.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1844

Caines, A., McCarthy, M., & Buttery, P. (2017). Parsing transcripts of speech. In *Proceedings of the workshop on speech-centric natural language processing* (pp. 27–36). Association for Computational Linguistics. http://aclweb.org/anthology/W17-4604

Čibej, J., Gantar, K., Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L., & Robnik-Šikonja, M. (2022). Morphological lexicon Sloleks 3.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1745

de Marneffe, M. C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308. doi: 10.1162/COLI_a_00402

Dobrovoljc, K. (2018, December). Formulaičnost v slovenskem jeziku. *Slovenščina 2.0: empirical, applied and interdisciplinary research*, *6*(2), 67–95. http://dx.doi.org/10.4312/slo2.0.2018.2.67-95 doi: 10.4312/slo2.0.2018.2.67-95

Dobrovoljc, K. (2022). Spoken Language Treebanks in Universal Dependencies: an Overview. In N. Calzolari et al. (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 1798–1806). Marseille, France: European Language Resources Association. https://aclanthology.org/2022.lrec-1.191

Dobrovoljc, K. (2024). Skladenjska drevesnica govorjene slovenščine: stanje in perspektive. In *Stanje in perspektive uporabe govornih virov v raziskavah govora* (p. 41–62). Univerza v Mariboru, Univerzitetna založba. http://dx.doi.org/10.18690/um.ff.4.2024.3 doi: 10.18690/um.ff.4.2024.3

Dobrovoljc, K., Erjavec, T., & Krek, S. (2017). The Universal Dependencies Tree- bank for Slovenian. In T. Erjavec, J. Piskorski, L. Pivovarova, J. Šnajder, J. Steinberger, & R. Yangarber (Eds.), *Proceedings of the 6th workshop on Balto-Slavic natural language processing* (pp. 33–38). Valencia, Spain: Association for Computational Linguistics. https://aclanthology.org/W17-1406 doi: 10.18653/v1/W17-1406

Dobrovoljc, K., & Ljubešić, N. (2022). Extending the SSJ Universal Dependen- cies Treebank for Slovenian: Was It Worth It? In S. Pradhan & S. Kuebler

(Eds.), *Proceedings of the 16th linguistic annotation workshop (law-xvi) within lrec2022* (pp. 15–22). European Language Resources Association. https://aclanthology.org/2022.law-1.3

Dobrovoljc, K., & Martinc, M. (2018). Er … well, it matters, right? On the role of data representations in spoken language dependency parsing. In M.-C. de Marneffe, T. Lynn, & S. Schuster (Eds.), *Proceedings of the second workshop on universal dependencies (UDW 2018)* (pp. 37–46). Association for Computational Linguistics. https://aclanthology.org/W18-6005 doi: 10.18653/v1/W18-6005

Dobrovoljc, K., & Nivre, J. (2016). The Universal Dependencies Treebank of Spoken Slovenian. In N. Calzolari et al. (Eds.), *Proceedings of the tenth inter- national conference on language resources and evaluation (LREC'16)* (pp. 1566– 1573). European Language Resources Association (ELRA). https://aclanthology.org/L16-1248

Dobrovoljc, K., & Terčon, L. (2023). *Universal Dependencies: Smernice za označevanje besedil v slovenščini. Različica 1.3. .* Center za jezikovne vire in tehnologije Univerze v Ljubljani. https://wiki.cjvt.si/attachments/66

Dobrovoljc, K., Terčon, L., & Ljubešić, N. (2023). Universal Dependen- cies za slovenščino: Nove smernice, ročno označeni podatki in razčlenjevalni model. *Slovenščina 2.0: empirical applied and interdisciplinary research*, *11*(1), 218–246. http://dx.doi.org/10.4312/slo2.0.2023.1.218-246 doi: 10.4312/slo2.0.2023.1.218-246

Erjavec, T. (2010). MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and C orpora. In N. Calzolari et al. (Eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC'10).* European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/pdf/138_Paper.pdf

Erjavec, T., Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Fišer, D., Laskowski, C., & Zupan, K. (2016). Annotating CLARIN. SI TEI corpora with WebAnno. In *Proceedings of the clarin annual conference* (pp. 1–5).

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone Speech Corpus for Research and Development. In *Proceedings of the 1992 ieee international conference on acoustics, speech and signal processing - volume 1* (pp. 517–520). IEEE Computer Society. http://dl.acm.org/citation.cfm?id=1895550.1895693

Hajič, J., Cinková, S., Mikulová, M., Pajas, P., Ptácek, J., Toman, J., & Uresová, Z. (2008). PDTSL: An annotated resource for speech reconstruction. In *Proceedings of the 2008 ieee workshop on spoken language technology* (pp. 93–96). IEEE.

Hinrichs, E., Bartels, J., Kawata, Y., Kordoni, V., & Telljohann, H. (2000). The Tübingen treebanks for spoken German, English, and Japanese. In W. Wahlster (Ed.), *Verbmobil: Foundations of speech-to-speech translation* (p. 550-574). Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-662-04230-4_40 doi: 10.1007/978-3-662-04230-4_40

Hinrichs, E., & Kübler, S. (2005). *Treebank profiling of spoken and written German*. Universitätsbibliothek Johann Christian Senckenberg.

Holozan, P., Krek, S., Pivec, M., Rigač, S., Rozman, S., Velušček, A., Pori, E., & Arhar Holdt, Š. (2023). *Specifikacije za učni korpus: lematizacija in MSD. Različica 2.0.* https://wiki.cjvt.si/attachments/21

Kahane, S., Caron, B., Strickland, E., & Gerdes, K. (2021). Annotation guide- lines of UD and SUD treebanks for spoken corpora: A proposal. In D. Dakota, K. Evang, & S. Kübler (Eds.), *Proceedings of the 20th international workshop on treebanks and linguistic theories (tlt, syntaxfest 2021)* (pp. 35–47). Association for Computational Linguistics. https://aclanthology.org/2021 .tlt-1.4

Kåsen, A., Hagen, K., Nøklestad, A., Priestly, J., Solberg, P. E., & Haug, D. T. T. (2022). The Norwegian Dialect Corpus Treebank. In N. Calzolari et al. (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 4827–4832). European Language Resources Association. https://aclanthology.org/2022.lrec-1.516

Krek, S., Dobrovoljc, K., Erjavec, T., Može, S., Ledinek, N., Holz, N., Zupan, K., Gantar, P., Kuzman, T., Čibej, J., Arhar Holdt, Š., Kavčič, T., Škrjanec, I., Dafne, M., Jezeršek, L., & Zajc, A. (2021). *Training corpus ssj500k 2.3.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1434

Krsnik, L., & Dobrovoljc, K. (2023). *The Trankit model for linguistic processing of standard Slovenian.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1870

Krsnik, L., & Dobrovoljc, K. (2024). *Trankit model for linguistic processing of spoken Slovenian.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1909

Krsnik, L., Dobrovoljc, K., & Robnik-Šikonja, M. (2024). *Dependency tree extraction tool STARK 3.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1958

Lacheret-Dujour, A., Kahane, S., & Pietrandrea, P. (2019). *Rhapsodie: A prosodic and syntactic treebank for spoken French* (Vol. 89). John Benjamins Publishing Company.

Lenardič, J., Čibej, J., Arhar Holdt, Š., Erjavec, T., Fišer, D., Ljubešić, N., Zupan, K., & Dobrovoljc, K. (2022). CMC training corpus janes-tag 3.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1732

Liu, Z., & Prud'hommeaux, E. (2021). Dependency Parsing Evaluation for Low-resource Spontaneous Speech. In *Proceedings of the second workshop on domain adaptation for nlp* (pp. 156–165). Association for Computational Linguistics. https://aclanthology.org/2021.adaptnlp-1.16

Ljubešić, N., & Dobrovoljc, K. (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* (pp. 29–34). Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-3704 doi: 10.18653/v1/ W19-3704

Luotolahti, J., Kanerva, J., & Ginter, F. (2017). dep_search: Efficient search tool for large dependency parsebanks. In *Proceedings of the 21st nordic conference on computational linguistics* (pp. 255–258).

MacWhinney, B. (2014). *The childes project*. Psychology Press. https://doi.org/10.4324/9781315805641 doi: 10.4324/9781315805641

Nguyen, M. V., Lai, V. D., Pouran Ben Veyseh, A., & Nguyen, T. H. (2021). Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Lan- guage Processing. In D. Gkatzia & D. Seddah (Eds.), *Proceedings of the 16th conference of the european chapter of the association for computational linguis- tics: System demonstrations* (pp. 80–90). Online: Association for Computational Linguistics. https://aclanthology.org/2021.eacl-demos.10 doi: 10.18653/v1/ 2021.eacl-demos.10

Øvrelid, L., Kåsen, A., Hagen, K., Nøklestad, A., Solberg, P. E., & Johannessen, J. B. (2018). The LIA Treebank of Spoken Norwegian Dialects. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)* (pp. 4482–4488). European Language Resources Asso- ciation (ELRA). https://aclanthology.org/L18-1710

Pietrandrea, P., & Delsart, A. (2019). Chapter 16. macrosyntax at work. In *Stud- ies in corpus linguistics* (pp. 285–314). John Benjamins Publishing Company.

Štravs, M., & Dobrovoljc, K. (2024). *Service for querying dependency treebanks Dreves-nik 1.1.* Slovenian language resource repos- itory CLARIN.SI. http://hdl.handle.net/11356/1923

Terčon, L., & Ljubešić, N. (2023). *The CLASSLA-Stanza model for UD dependency pars-ing of standard Slovenian 2.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1769

Terčon, L., & Ljubešić, N. (2023). *CLASSLA-Stanza: The Next Step for Linguistic Process-*

ing of South Slavic Languages.

van der Wouden, T., Hoekstra, H., Moortgat, M., Renmans, B., & Schuurman, I. (2002). Syntactic analysis in the Spoken Dutch Corpus (CGN). In *Proceedings of the third international conference on language resources and evaluation, LREC 2002, may 29-31, 2002, las palmas, canary islands, spain.* http://www.lrec-conf.org/proceedings/lrec2002/pdf/71.pdf

van der Wouden, T., Schuurman, I., Schouppe, M., & Hoekstra, H. (2003). Harvesting Dutch Trees: Syntactic Properties of Spoken Dutch. In *Computational lin- guistics in the netherlands 2002* (p. 129–141). BRILL. http://dx.doi.org/10.1163/9789004334441_011 doi: 10.1163/9789004334441_011

Verdonik, D., & Bizjak, A. (2023). *Pogovorni zapis in označevanje govora v govorni bazi Artur projekta RSDO.* https://dk.um.si/IzpisGradiva.php?lang=slv&id=85198

Verdonik, D., Bizjak, A., Sepesy Maučec, M., Gril, L., Dobrišek, S., Križaj, J., Strle, G., Bajec, M., Lebar Bajec, I., Jelovšek, T., Lokovšek, J., Trojar, M., Erjavec, T., Bernjak, M., Žganec Gros, J., Čakš, P., Pucer, M., Cvetko, M., Pavlič, J., ... & Dretnik, N. (2023). *ASR database ARTUR 1.0 (transcriptions).* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/ 1772

Verdonik, D., Bizjak, A., Žgank, A., Bernjak, M., Antloga, Š., Majhenič, S., Čakš, P., Pucer, M., Cvetko, M., Zelenik, M., Pavlič, J., Dobrišek, S., Križaj, J., Strle, G., Ivanovska, M., GRm, K., Bajec, M., Lebar Bajec, I., Jelovšek, T., ... Bordon, D. (2023). ASR database *ARTUR 1.0 (audio).* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1776

Verdonik, D., Dobrovoljc, K., Erjavec, T., & Ljubešić, N. (2024). Gos 2: A New Reference Corpus of Spoken Slovenian. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 7825–7830). ELRA and ICCL. https://aclanthology.org/2024 .lrec-main.691

Verdonik, D., Dobrovoljc, K., Čibej, J., Ljubešić, N., & Rupnik, P. (2024). Izbor in urejanje gradiv za učni korpus govorjene slovenščine - ROG. In *Zbornik konference jezikovne tehnologije in digitalna humanistika 2024.*

Verdonik, D., Kosem, I., Vitez, A. Z., Krek, S., & Stabej, M. (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. http://dx.doi.org/10.1007/s10579-013-9216-5 doi: 10.1007/s10579-013-9216-5

Verdonik, D., & Maučec, M. S. (2016). A speech corpus as a source of lexical information. *International Journal of Lexicography*, *30*(2), 143-166. https://doi.org/10.1093/ijl/ecw004 doi: 10.1093/ijl/ecw004

Verdonik, D., Potočnik, T., Sepesy Maučec, M., Erjavec, T., Majhenič, S., & Žgank, A. (2021). *Spoken corpus Gos VideoLectures 4.2 (transcription).* Slovenian language resource repository CLARIN.SI. http://hdl.handle .net/11356/1444

Verdonik, D., Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., Erjavec, T., ...

Rupnik, P. (2023). *Spoken corpus Gos 2.1 (transcriptions).* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/ 11356/1863

Yimam, S. M., Gurevych, I., De Castilho, R. E., & Biemann, C. (2013). Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st annual meeting of the association for computational linguistics: System demonstrations* (pp. 1–6).

Zeman, D., et al. (2023). *Universal Dependencies 2.12.* LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-5150

Zeman, D., et al. (2024). *Universal Dependencies 2.14.* LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-5502

Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., & Erjavec, T. (2021). *Spoken corpus Gos 1.1.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1438

Čibej, J., & Munda, T. (2024). Metoda polavtomatskega popravljanja lem in oblikoskladenjskih oznak na primeru učnega korpusa govorjene slovenščine ROG. In *Zbornik konference jezikovne tehnologije in digitalna humanistika 2024.*

RAZŠIRITEV DREVESNICE GOVORJENE SLOVENŠČINE SST

V prispevku predstavljamo novo različico drevesnice govorjene slovenščine SST (angl. *Spoken Slovenian Treebank*), uravnoteženega in reprezentativnega korpusa transkribiranega govora z ročno označenimi lemami, besednimi vrstami, oblikoslovnimi lastnostmi in skladenjskimi odvisnostmi med besedami. Izvorno različico drevesnice SST smo razširili z več kot 3.000 novimi izjavami in jo izboljšali z vidika poenotenja načel zapisovanja govora ter zanesljivosti ročno pripisanih oznak. Po kratki predstavitvi vzorčenja novih podatkov iz referenčnega korpusa govorjene slovenščine GOS 2 ter polavtomatskega oblikoslovnega označevanja v jedru prispevka opisujemo proces skladenjskega razčlenjevanja novih besedil ter poenotenja med prvotnimi in novo dodanimi transkripcijami, ki so se razlikovale na ravni segmentacije govora, rabe ločil in velikih začetnic. V drugem delu vsebino nove različice drevesnice SST povzamemo z vidika velikosti in raznolikosti podatkov in predstavimo rezultate njene primerjave z referenčno drevesnico pisne slovenščine SSJ, ki razkriva unikatne leksikalne in skladenjske lastnosti govorjenega jezika.

Keywords: korpusno označevanje, odvisnostna drevesnica, govorjeni jezik, spontani govor, Universal Dependencies