# ROG: A Multi-Layer Manually Annotated Corpus of Spoken Slovenian

**Kaja Dobrovoljc[1,2], Darinka Verdonik[3], Jaka Čibej[1,2], Peter Rupnik[2], Nikola Ljubešić[1,2]**
[1]University of Ljubljana, Slovenia
[2]Jožef Stefan Institute, Ljubljana, Slovenia
[3]Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia

**Abstract**

We present ROG, the first manually annotated spoken corpus of Slovenian to integrate morphosyntactic, prosodic, and interactional layers in a unified framework. Building on the pre-existing Spoken Slovenian Treebank (SST) and newly available recordings from the GOS 2 reference corpus, the resource combines over 75,000 words (10 hours) of annotated speech. The entire corpus features lemmatization, MULTEXT-East morphosyntax, and Universal Dependencies annotations, while approximately half includes additional layers for prosodic units, disfluencies, and dialogue acts. All annotation layers are systematically aligned and cross-referenced, enabling detailed multi-dimensional analyses of spoken language. We describe the corpus design, annotation workflow, data release, and baseline modeling results, showcasing the resource's value for both linguistic analysis and speech-aware NLP model development. All ROG transcriptions and annotations, along with half of the audio recordings, are freely available under CC-BY via CLARIN.SI repository: http://hdl.handle.net/11356/2062.

**Keywords:** spoken corpora, linguistic annotation, morphosyntax, prosody, disfluency, dialogue acts, speech processsing

## 1. Introduction

Manually annotated spoken corpora are essential for both linguistic research and speech technology (Ide and Pustejovsky, 2017). They provide gold-standard data for training and evaluating NLP systems (Bhat et al., 2023; Pupier et al., 2024; Si et al., 2023) and a key empirical basis for studying the structure and use of speech (Hinrichs and Kübler, 2005; Pietrandrea and Delsart, 2019). Consequently, numerous such corpora have been developed for different languages, communicative settings and annotation frameworks (e.g., Godfrey et al., 1992; Lacheret-Dujour et al., 2019; Kåsen et al., 2022; Hinrichs et al., 2000; MacWhinney, 2014, to name just a few).

For Slovenian, a South Slavic language with roughly two million speakers, the largest manually annotated resource to date has been the Spoken Slovenian Treebank (SST; Dobrovoljc and Nivre, 2016), which provided morphosyntactic annotation for samples from the reference GOS corpus (Verdonik et al., 2013). In parallel, other manually annotated corpora have emerged to address complementary dimensions, such as GORDAN (Verdonik, 2020) for dialogue acts and KOMET (Antloga and Donaj, 2022) for metaphoric meaning. While valuable within their respective domains, these resources remained limited in scope and based on heterogeneous, non-overlapping datasets. As a result, research on Slovenian spoken grammar and interaction has lacked a unified empirical foundation for systematic cross-dimensional linguistic analysis.

To bridge this gap, we combined efforts from two parallel projects—SPOT (*Treebank-Driven Approach to the Study of Spoken Slovenian*),[1] which aimed to extend the original SST by 50,000 new tokens, and MEZZANINE (*Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language*),[2] which aimed to develop a new 40,000-word corpus annotated for prosody, disfluencies, and dialogue acts. Recognizing their complementarity, the two initiatives were aligned toward a joint objective: the creation of ROG (*Ročno označeni govor* 'Manually Annotated Speech'), a multi-layer reference corpus of spoken Slovenian that would unify previous annotation efforts, integrate new linguistic layers, and ensure interoperability with existing Slovenian and multilingual resources.

In this paper, we present the design and development of the ROG corpus, providing the first comprehensive account of the resource as a whole. We outline its composition and annotation workflow (Sections 2–3), detail its formats and availability (Sections 4–5), and report baseline modeling results that demonstrate its early adoption and research potential (Section 6).

## 2. Data Selection and Consolidation

As outlined in the introduction, the ROG corpus extends the original SST treebank, which contained about 30,000 manually annotated words from the

[1]https://spot.ff.uni-lj.si/en/
[2]https://mezzanine.um.si/



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 5701–5710
11–16 May 2026. 

GOS 1.1 (Verdonik et al., 2013; Zwitter Vitez et al., 2021) reference corpus of spoken Slovenian. Our goal was to expand this dataset by approximately 50,000 new tokens, with two key objectives: improving audio quality and maintaining the demographic and genre balance of the reference corpus.

The expansion was based on the most recent GOS 2.1 corpus (Verdonik et al., 2023b), as described by Verdonik et al. (2024), which combines GOS 1.1 with new recordings from the open-access ARTUR ASR database (Verdonik et al., 2023a). Around 10,000 words were added by extending selected SST speech events from GOS 1.1 subset, mainly to better represent informal spontaneous conversations. The remaining 40,000 words were sampled from ARTUR subset, prioritizing high-quality audio and a balanced mix of speech types, including new ones such as parliamentary debates, round tables, and online meetings. All data in the GOS 2.1 and Artur corpora were previously anonymized with regard to personal information, including names and surnames.

Although all material originates from GOS 2.1, differences in transcription conventions required harmonization. All ROG data were standardized for capitalization and punctuation (lowercased except proper nouns, with written-like punctuation). ARTUR transcriptions were also automatically resegmented into sentence-like utterances to match the GOS 1.1 definition of 'prosodically, syntactically, and semantically complete units' (Verdonik et al., 2013). In addition, ROG preserves the original two-level transcriptions of GOS 2.1, comprising both normalized and pronunciation-based orthography.

The resulting corpus comprises nearly 100,000 tokens, providing a balanced and representative sample of speech across demographic groups and communicative settings (see Dobrovoljc, 2025 for detailed distribution plots by speech event type and channel, as well as speaker gender, age, education, first language, and place of residence). The final composition of the dataset is summarized in Table 1.

| Subset | Ev. | Spk. | Utt. | Tokens | Words | H |
|---|---|---|---|---|---|---|
| From GOS 1 | 287 | 594 | 4,139 | 50,072 | 37,340 | 5 |
| From ARTUR | 57 | 72 | 1,969 | 48,321 | 39,001 | 5 |
| **ROG (total)** | **344** | **676** | **6,108** | **98,393** | **76,341** | **10** |

Table 1: Overview of the ROG corpus, showing the number of events (Ev.), speakers (Spk.), utterances (Utt.), tokens and words, and the total length of audio recordings in hours (H).

## 3. Manual Annotation

Once the corpus data had been selected and consolidated, we proceeded with linguistic annotation across several independent campaigns to capture a wide range of speech-related phenomena at multiple levels of analysis. In the first stage, the existing morphosyntactic annotation of the original SST was extended to the entire dataset. In the second stage, the ARTUR subset of the corpus was enriched with additional annotation layers.

The following subsections describe each layer in more detail, and their overall scope is summarized in Table 2 (Section 4). An example of their alignment within a single utterance is shown in Figure 1.

### 3.1. Extension of morphosyntactic annotation

In the first stage, we extended the existing SST annotations, which are based on the MULTEXT-East morphosyntactic specifications (Erjavec, 2010) and the Universal Dependencies framework (de Marneffe et al., 2021). Together, these schemes represent the de facto standard for Slovenian language resources, ensuring both national and cross-linguistic interoperability.

#### 3.1.1. Lemmatization and MULTEXT-East Morphosyntactic Annotation

The corpus was first automatically annotated with the Slovene CLASSLA-Stanza 2.0 models for lemmatization (Terčon et al., 2023) and morphosyntactic annotation (Ljubešić et al., 2023). The tagset follows the *MULTEXT-East v6 Morphosyntactic Specifications for Slovene*,[3] in which the tags are finegrained and contain multiple linguistic characteristics in additon to the part-of-speech information. For instance, the tag *Somer* stands for: *samostalnik* (noun), *občni* (common), *moški* (masculine), *ednina* (singular), *rodilnik* (genitive).

All lemmas and tags were manually validated by two expert gold-standard annotators (one for lemmas and the other for the morphosyntactic tags). No double annotation was done because the expert annotators had been previously involved as final curators in the annotation of the *SUK Training Corpus of Slovene* (Arhar Holdt et al., 2024). However, lemmas and tags were first automatically cross-referenced with the *Sloleks Morphological Lexicon of Slovene* (Čibej et al., 2022), the largest open-source database with machine-readable information on Slovene words and their inflected forms, with approximately 100,800 manually validated lexemes and 2.8 million inflected forms (and their morphosyntactic tags).

The corrections were categorized into microtasks based on the likelihood of errors – for instance,

[3]MULTEXT-East v6 Morphosyntactic Specifications for Slovene: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html

the method took into account whether the tokens are homographs in the lexicon. The method clustered similar tagging problems and ensured that all tagging errors of the same type were corrected consistently throughout the corpus – the process is described in more detail in Čibej and Munda, 2024 and Čibej and Munda, upcoming . The lemmas were corrected first, followed by different sets of morphosyntactic tags. In total, only 1% of the tokens required lemma corrections, and 4.3% required corrections of morphosyntactic tags.

#### 3.1.2. Universal Dependencies

Universal Dependencies (UD)[4] is a cross-linguistic framework for morphosyntactic annotation that provides a harmonized set of part-of-speech tags, morphological features, and syntactic dependency relations across languages. The existing MULTEXT-East morphosyntactic annotation was automatically converted to UD POS tags and features using a rule-based conversion script, and parsed with a custom Trankit model (Nguyen et al., 2021) for Slovenian.

Manual correction was performed by trained annotators in the Q-CAT tool (Brank, 2023), adapted to enable simultaneous text and audio inspection via embedded URLs. Each document was reviewed by two to three annotators, and discrepancies were resolved in a centralized WebAnno (Yimam et al., 2013) curation stage. The process followed an updated version of the Slovenian UD guidelines (Dobrovoljc and Terčon, 2024), which explicitly cover speech-specific phenomena such as self-repairs, discourse markers, and truncated utterances, consistent with best practices from comparable spoken UD treebanks (Kahane et al., 2021; Dobrovoljc, 2022).

Finally, the newly annotated data - combining MULTEXT-East (previous section) and UD annotations (this section) – was merged with the original SST and consolidated for transcription and annotation consistency (Dobrovoljc, 2024). As a result, the extended SST UD treebank represents one of the largest spoken UD treebanks currently available, with only the Naija NSC treebank exceeding it in size (140k tokens).

### 3.2. New Annotation Layers

In the second stage, we introduced additional annotation layers reflecting prosodic, discourse, and interactional dimensions of spoken communication, some of which were implemented for the first time in Slovenian. Given the limited resources available, these layers were initially applied only to the ARTUR subset of the corpus. The assessment of interannotator agreement for these newly introduced layers remains to be conducted.

#### 3.2.1. Prosodic units

Prosodic units—also referred to as intonation units, tone units, or intonational phrases—have been defined in different ways: semantically, as conveying a single idea (Chafe, 1994), syntactically and prosodically, as complete structures (Degand and Simon, 2009), or empirically, through prosodic cues such as pauses, pitch reset, tempo deceleration, syllable lengthening, and changes in intensity (Izre'el, 2020).

In the ROG corpus, prosodic units were annotated exclusively on the basis of prosodic cues as specified above, i.e., pauses, pitch reset, tempo deceleration, syllable lengthening, and loudness deceleration. Ideally, two or more of these phenomena were present to provide sufficient criteria for annotating a prosodic unit. The annotation was carried out manually using the speech analysis software Praat (Boersma and Weenink, 1992–2022) and subsequently imported into the Partitur Editor (Schmidt and Wörner, 2014), where further manual corrections were applied. The procedure consisted of the following steps: (1) students were trained to annotate prosodic units in Praat; (2) an experienced researcher revised the students' annotations in Praat; (3) the annotations were imported into the Partitur Editor format; and (4) a second experienced researcher revised the annotations in the Partitur Editor.

In total, 5,659 prosodic units have been manually identified in the ARTUR subset of the ROG corpus.

#### 3.2.2. Disfluencies

Disfluencies have been described as repetitions, false starts, filled or silent pauses (Maclay and Osgood, 1959), later refined in detailed taxonomies such as that proposed by Shriberg (1996). Recent approaches take a cross-linguistic and multimodal view that integrates prosodic and interactional features (Crible et al., 2024; Kosmala, 2024), reflecting a shift toward more comprehensive modeling of spontaneous speech.

The ROG disfluency annotation scheme builds on this tradition and closely follows Kosmala (2024). It was developed through corpus inspection, pilot testing, and expert revision, with annotation performed in the Partitur Editor (Schmidt and Wörner, 2014) by trained students and validated by experts. The hierarchical scheme distinguishes three tiers: (1) vocal disfluencies (silent and filled pauses, lengthening, blocks, non-linguistic sounds); (2) verbal disfluencies (repetitions, self-repair pronunciation, self-repair lexicogrammar, self-repair restart,

[4] https://universaldependencies.org/

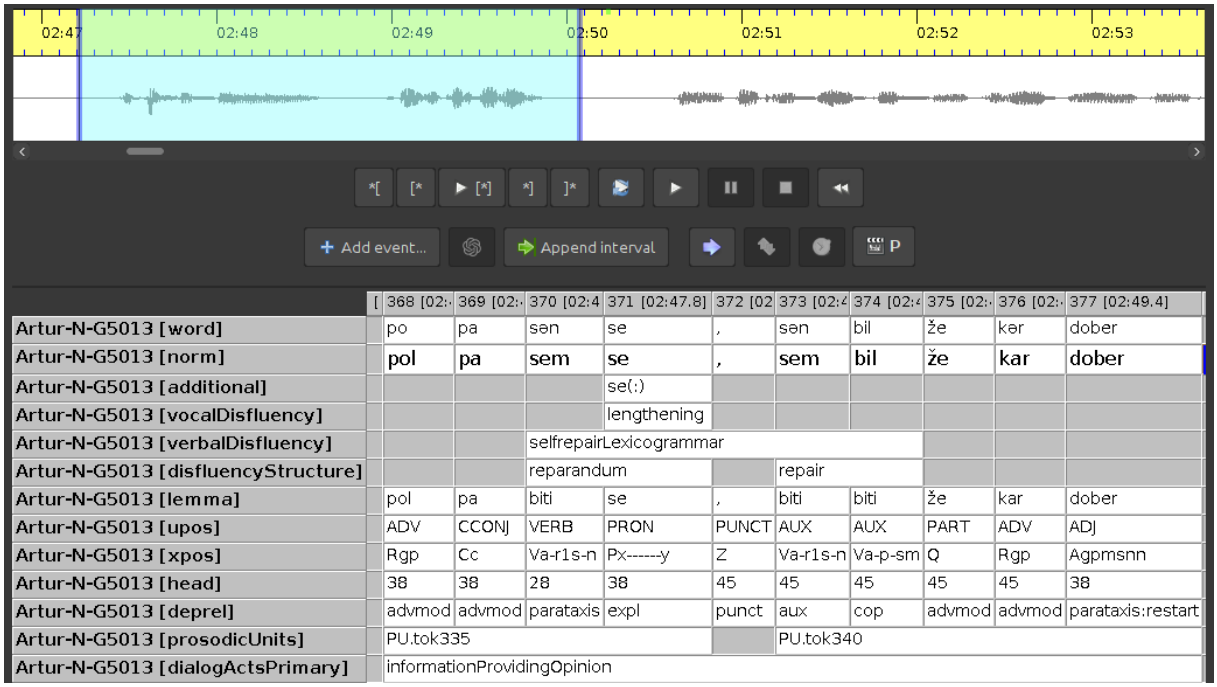


Figure 1: Example of an annotated ROG utterance in the EXMARaLDA tool (roughly translated as '*and then I had, I was pretty good*'), showing the two-level transcription and aligned layers for lemmatization, morphology (upos, xpos), dependency syntax (head, deprel), prosody, disluencies, and dialogue acts.

self-repair complex, unrepaired pronunciation, unrepaired structure, abandoned, disfluency marker, disfluency comment); and (3) self-repair structure (reparandum, interregnum, repair).

In total, 6,234 disfluencies have been manually identified in the ARTUR subset of the ROG corpus. Additional information are available in (Verdonik et al., 2025).

#### 3.2.3. Dialogue Acts

For dialogue annotation, we thoroughly revisited four generic dialogue act annotation schemes: AMI, DART, ISO 24617-2, and SWBD-DAMSL, surveyed by Verdonik (2023). None of these schemes proved fully optimal, as each exhibited certain limitations. We therefore adopted a highly simplified seven-category annotation scheme for basic dialogue acts, comprising: (1) information providing fact, (2) information providing opinion, and (3) information providing sentiment (all derived from ISO information-providing functions); (4) a single tag for information seeking acts; (5) a single tag for directive and commissive functions, termed as action discussion; (6) a single tag for social obligations management; and (7) a single tag for metadiscourse acts, which cover ISO feedback, turn management, time management, discourse structuring, and own- and partner-management functions. All annotations were carried out manually in the Partitur Editor and placed into a separate tier. Developing a more detailed annotation scheme remains a task for future research.

## 4. Formats and License

Reflecting the two-stage annotation design described above, the ROG corpus is distributed in two complementary parts that differ in scope, format and audio availability (Table 2). The first, ROG-SST,[5] covers the entire dataset and includes all morphosyntactic layers described in Section 3.1, provided in CoNLL-U format. Each sentence also contains links to utterance-level audio segments, while complete recordings of individual speech events are available from the reference GOS 2.1 corpus (Verdonik et al., 2024).

The second part, ROG-Art, adds annotation layers for prosody, disfluencies, and dialogue acts (Section 3.2). It is distributed in EXMARaLDA format (.EXB) with accompanying files – .EXS and .coma for EXAKT concordancing, .TRS for Transcriber, and .TextGrid for Praat, and an ISO-TEI representation – allowing flexible viewing and editing across tools (e.g., Figure 1). Audio recordings

[5]The working name *ROG-SST* denotes that this part of the ROG corpus was also released as an updated version of the SST UD treebank within the UD release v2.14 (Zeman et al., 2024).

are available in single-channel .WAV format.

Both corpus parts share consistent token and utterance identifiers, allowing direct cross-referencing across layers and linking to the original GOS 2.1 metadata. A standardized 80-10-10 train–development–test split supports reproducible experiments and fair model comparison, with the ROG-Art split fully nested within the ROG-SST split. The splits were prepared manually on the level of recordings, which allowed balancing recording metadata across splits while preventing speaker leakage between them. With the exception of the GOS 1.0 audio, all data, documentation, and metadata are freely available through the CLARIN.SI repository under a CC-BY-SA 4.0 license (Verdonik et al., 2026).

## 5. Availability

In addition to the data release described above, the corpus is accessible through several tools and services supporting data browsing and analysis, though coverage of annotation layers differs across platforms.

The morphosyntactically parsed layers of the corpus (ROG-SST) can also be queried through existing UD-compatible services such as INESS (Rosén et al., 2012),[6] PML Tree Query (Štěpánek and Pajas, 2010),[7], Drevesnik (Štravs et al., 2025), [8] and Grew-match (Guillaume, 2021),[9] with the latter also supporting utterance-level audio playback.

In the near future, we plan to integrate the full corpus into the CLARIN.SI installations of noSketchEngine and Kontext concordancers,[10] while continuing our search for an optimal platform that would support user-friendly visualisation and analysis across all annotation layers.

## 6. Baseline Modeling

Finally, to demonstrate the potential of the corpus for computational linguistic research and automatic processing of spoken Slovenian, we trained several baseline models on key text- and speech-based tasks, including grammatical annotation, prosodic unit identification, and filled-pause detection.

[6] https://clarino.uib.no/iness-prod/treebanks

[7] https://lindat.mff.cuni.cz/services/pmltq

[8] https://orodja.cjvt.si/drevesnik/en

[9] https://universal.grew.fr/?corpus=UD_Slovenian-SST@2.16

[10] https://www.clarin.si/info/concordances/

| Annotation layer | ROG-SST (full corpus) | ROG-Art (subset) |
|---|---|---|
| Lemmatization | x | x |
| MULTEXT-East | x | x |
| Univ. Dependencies | x | x |
| Prosodic units | – | x |
| Disfluencies | – | x |
| Dialogue acts | – | x |
| **Formats** | .conllu | .exb, .xml |
| **Text license** | CC-BY | CC-BY |
| **Audio license** | research only | CC-BY |

Table 2: Overview of annotation layers, formats, and licenses for the ROG corpus.

### 6.1. Grammatical annotation on text

Two annotation pipelines, Trankit (Nguyen et al., 2021), and CLASSLA-Stanza (Ljubešić and Dobrovoljc, 2019; Terčon et al., 2025), were trained on combined spoken (ROG-SST) and written Slovenian data (Dobrovoljc et al., 2017; Arhar Holdt et al., 2024) and evaluated on the ROG corpus for lemmatization, part-of-speech tagging (UPOS), full morphosyntactic MULTEX-East tagging (XPOS), and dependency parsing (LAS).

As summarized in Table 3, both systems perform strongly on morphological processing, but parsing spoken data remains substantially more difficult. The Trankit transformer technology shows to beat the older RNN-style Stanza approach, especially on syntactic parsing, with only lemmatization still having the upper hand in CLASSLA-Stanza due to an included inflectional lexicon. A more extensive discussion of these findings, including a comparison with standard Slovenian models for written language processing, is provided by Dobrovoljc (2025) and Terčon et al. (2025).

| Model | Lemmas | UPOS | XPOS | LAS |
|---|---|---|---|---|
| Trankit | 98.85 | 98.97 | 98.02 | 87.93 |
| CLASSLA | 99.23 | 98.15 | 96.76 | 81.91 |

Table 3: Performance (micro F1) of Trankit and CLASSLA-Stanza spoken models on the ROG test set.

### 6.2. Prosodic unit detection from speech

We developed a prosodic unit detector by fine-tuning a Wav2VecBert2 model (Barrault et al., 2023) with an audio frame classification head on the training portion of the ROG dataset. The audio frame classification head allows us to predict for each 20 ms audio frame whether a prosodic unit boundary should be placed there or not.

The development set of the ROG dataset allowed us to do hyperpararameter optimization, investigating learning rates of $3\times10^{-5}$, $1\times10^{-6}$, and $8\times10^{-6}$, total training duration of 10 and 20 epochs, and

batch sizes of 4 and 16. The final system was trained with learning rate $3 \times 10^{-5}$ for 20 epochs in batches of 16.

We used the test portion of ROG to evaluate the final system. The evaluation aims at comparing gold prosodic unit spans with predicted prosodic unit spans, aligning best-fit prosodic units between gold and predicted as true positives, while the remaining prosodic units are either considered false positives (non-aligned prosodic unit from predicted) or false negatives (non-aligned prosodic unit from gold). This evaluation protocol resulted in precision of 0.9464 and recall of 0.8260, and an F1 of 0.8821. Given that recall is significantly lower than precision, this points that the model more often identifies only one prosodic unit where human annotators annotated two or more. Nevertheless, we consider both evaluation metrics to be reasonably high, and the system very useful in pre-annotating data, and thereby speeding up manual annotation. The model is available through HuggingFace[11].

### 6.3. Filled pause detection from speech

The filled pause detection system (Rupnik et al., 2024) was developed in a very similar way to the above prosodic unit identifier, by fine-tuning a Wav2VecBert2 model with an audio frame classifier, and the hyperparameter search on the dev dataset resulting in the same optimal hyperparameters.

Evaluation on the ROG test data, comparable to the evaluation of prosodic units, resulted in precision of 0.914 and recall of 0.973. A manual analysis of human and model disagreements by a phonetician showed that the model actually misses fewer filled pauses than human annotators, but sometimes predict filled pauses erroneously, which is consistent with the high observed recall and lower precision. More details can be found and on the HuggingFace model card[12].

## 7. Conclusion

We presented ROG, a new open-source, multi-layer corpus of spoken Slovenian designed to support linguistic research and computational modelling of Slovenian speech. The corpus consolidates existing resources, extends them with new annotation layers, and provides standardized formats and splits for reproducible modeling.

At present, the ROG corpus is being expanded with newly collected data comprising private, conversational speech — a type of data that is least represented in the existing ROG corpus, yet poses the greatest challenges for both linguistic research and the development of language technologies. The long-term objective of this work is to harmonize the newly collected and existing data across multiple levels of annotation and to extend the fully open-access ROG corpus from five to ten hours of recorded speech. We will also work toward unifying the corpus across annotation layers and formats, extending its coverage with additional data and annotations, and fostering its use in both computational and linguistic research. The latter also includes integrating the resource into infrastructures that enable seamless access, visualization, and analysis of all available layers.

[11] https://huggingface.co/classla/wav2vecbert2-prosodicUnit

[12] https://huggingface.co/classla/wav2vecbert2-filledPause

## 8. Acknowledgements

We thank all annotators, collaborators, and contributors involved in the development of the ROG corpus.

The research presented in this paper was conducted within the research projects “Treebank-Driven Approach to the Study of Spoken Slovenian“ (Z6-4617), “Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language” (J7-4642), “Large Language Models for Digital Humanities” (GC-0002), the Research Infrastructure DARIAH-SI (I0-E007), and within the research programmes “Language resources and technologies for Slovene” (P6-0411) and “Advanced Methods of Interaction in Telecommunications” (P2-0069), all funded by the Slovenian Research and Innovation Agency (ARIS).

Generative AI tools were used to assist with language editing.

## 9. Bibliographical References

Špela Arhar Holdt, Jaka Čibej, Kaja Dobrovoljc, Tomaž Erjavec, Polona Gantar, Simon Krek, Tina Munda, Nejc Robida, Luka Terčon, and Slavko Žitnik. 2024. SUK 1.0: A New Training Corpus for Linguistic Annotation of Modern Standard Slovene. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15428–15435, Torino, Italia. ELRA and ICCL.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean

Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Vineet Bhat, Preethi Jyothi, and Pushpak Bhattacharyya. 2023. DISCO: A large scale human annotated corpus for disfluency correction in Indo-European languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12833–12857, Singapore. Association for Computational Linguistics.

Wallace Chafe. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press.

Jaka Čibej and Tina Munda. 2024. Metoda polavtomatskega popravljanja lem in oblikoskladenjskih oznak na primeru učnega korpusa govorjene slovenščine rog. In *Proceedings of the Conference on Language Technologies & Digital Humanities 2024 (JTDH2024)*, pages 66–86, Ljubljana, Slovenia.

Jaka Čibej and Tina Munda. upcoming. Leveraging a morphological lexicon for a semi-automatic approach to correcting lemmas and morphosyntactic tags. *Contributions to Contemporary History*.

Ludivine Crible, Ivana Didirková, Christelle Dodane, and Loulou Kosmala. 2024. Towards an inclusive system for the annotation of (dis)fluency in typical and atypical speech. *Clinical Linguistics & Phonetics*, 38(4):381–398. PMID: 36205188.

Marie Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Liesbeth Degand and Anne Catherine Simon. 2009. On identifying basic discourse units in speech: theoretical and empirical issues. *Discours*, (4).

Kaja Dobrovoljc. 2022. Spoken Language Treebanks in Universal Dependencies: an Overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.

Kaja Dobrovoljc. 2024. Extending the spoken slovenian treebank.

Kaja Dobrovoljc. 2025. Treebanking spoken slovenian: New data, models, and lessons learned. *Contributions to Contemporary History*, 65(3).

Kaja Dobrovoljc, Tomaž Erjavec, and Simon Krek. 2017. The Universal Dependencies Treebank for Slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 33–38, Valencia, Spain. Association for Computational Linguistics.

Kaja Dobrovoljc and Joakim Nivre. 2016. The Universal Dependencies Treebank of Spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1566–1573, Portorož, Slovenia. European Language Resources Association (ELRA).

Kaja Dobrovoljc and Luka Terčon. 2024. Universal Dependencies: Smernice za označevanje besedil v slovenščini. Version 1.7. .

Tomaž Erjavec. 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 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*, ICASSP'92, pages 517–520, Washington, DC, USA. IEEE Computer Society.

Bruno Guillaume. 2021. Graph matching and graph rewriting: Grew tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175.

Erhard Hinrichs, Julia Bartels, Yasuhiro Kawata, Valia Kordoni, and Heike Telljohann. 2000. The Tübingen treebanks for spoken German, English, and Japanese. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech*

*Translation*, Artificial Intelligence, pages 550–574. Springer Berlin Heidelberg.

Erhard Hinrichs and Sandra Kübler. 2005. *Treebank profiling of spoken and written German*. Universitätsbibliothek Johann Christian Senckenberg.

Nancy Ide and James Pustejovsky. 2017. *Handbook of Linguistic Annotation*, 1st edition. Springer Publishing Company, Incorporated.

Shlomo Izre'el. 2020. *Chapter 2. The basic unit of spoken language and the interfaces between prosody, discourse and syntax: A View from spontaneous spoken Hebrew*, pages 77–106. John Benjamins Publishing Company.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021. Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, Syntaxfest 2021)*, pages 35–47. Association for Computational Linguistics.

Andre Kåsen, Kristin Hagen, Anders Nøklestad, Joel Priestly, Per Erik Solberg, and Dag Trygve Truslew Haug. 2022. The Norwegian Dialect Corpus Treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4827–4832, Marseille, France. European Language Resources Association.

Loulou Kosmala. 2024. *Beyond Disfluency*. John Benjamins.

Anne Lacheret-Dujour, Sylvain Kahane, and Paola Pietrandrea. 2019. *Rhapsodie: A prosodic and syntactic treebank for spoken French*, volume 89. John Benjamins Publishing Company.

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. Association for Computational Linguistics.

Howard Maclay and Charles E. Osgood. 1959. Hesitation phenomena in spontaneous english speech. *WORD*, 15(1):19–44.

Brian MacWhinney. 2014. *The Childes Project*. Psychology Press.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Paola Pietrandrea and Aline Delsart. 2019. Chapter 16. macrosyntax at work. In *Studies in Corpus Linguistics*, pages 285–314. John Benjamins Publishing Company, Amsterdam.

Adrien Pupier, Maximin Coavoux, Jérôme Goulian, and Benjamin Lecouteux. 2024. Growing trees on sounds: Assessing strategies for end-to-end dependency parsing of speech. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 225–233, Bangkok, Thailand. Association for Computational Linguistics.

Victoria Rosén, Koenraad De Smedt, Paul Meurer, and Helge Dyvik. 2012. An open infrastructure for advanced treebanking. In *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29. Hajič, Jan.

Thomas Schmidt and Kai Wörner. 2014. Exmaralda. In *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.

Elizabeth Shriberg. 1996. Disfluencies in Switchboard. In *Proceedings of International Conference on Spoken Language Processing*, volume 96, pages 11–14.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. In *Advances in Neural Information Processing Systems*, volume 36, pages 39088–39118. Curran Associates, Inc.

Jan Štěpánek and Petr Pajas. 2010. Querying diverse treebanks in a uniform way. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Luka Terčon, Kaja Dobrovoljc, and Nikola Ljubešić. 2025. Classla-stanza: The next step for linguistic processing of south slavic languages. *Contributions to Contemporary History*, 65(3).

Darinka Verdonik. 2023. Annotating dialogue acts in speech data. *International Journal of Corpus Linguistics*, 28(2):144–171.

Darinka Verdonik, Kaja Dobrovoljc, Tomaž Erjavec, and Nikola Ljubešić. 2024. Gos 2: A new reference corpus of spoken Slovenian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7825–7830, Torino, Italia. ELRA and ICCL.

Darinka Verdonik, Iztok Kosem, Ana Zwitter Vitez, 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.

Darinka Verdonik, Peter Rupnik, and Nikola Ljubešić. 2025. Disfluencies in public and private speech. *Language and Speech*. PMID: 41214903.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart De Castilho, and Chris Biemann. 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*, pages 1–6.

## 10. Language Resource References

Antloga, Špela and Donaj, Gregor. 2022. *Corpus of metaphorical expressions in spoken Slovene language G-KOMET 1.0*. Slovenian language resource repository CLARIN.SI.

Arhar Holdt, Špela and Krek, Simon and Dobrovoljc, Kaja and Erjavec, Tomaž and Gantar, Polona and Čibej, Jaka and Pori, Eva and Terčon, Luka and Munda, Tina and Žitnik, Slavko and Robida, Nejc and Blagus, Neli and Može, Sara and Ledinek, Nina and Holz, Nanika and Zupan, Katja and Kuzman, Taja and Kavčič, Teja and Škrjanec, Iza and Marko, Dafne and Jezeršek, Lucija and Zajc, Anja. 2024. *Training corpus SUK 1.1*. Slovenian language resource repository CLARIN.SI.

Paul Boersma and David Weenink. 1992–2022. *Praat: doing phonetics by computer [Computer program]*. Version 6.4.45, retrieved 12 October 2025 from https://praat.org.

Brank, Janez. 2023. *Q-CAT Corpus Annotation Tool 1.5*. Slovenian language resource repository CLARIN.SI.

Čibej, Jaka and Gantar, Kaja and Dobrovoljc, Kaja and Krek, Simon and Holozan, Peter and Erjavec, Tomaž and Romih, Miro and Arhar Holdt, Špela and Krsnik, Luka and Robnik-Šikonja, Marko. 2022. *Morphological lexicon Sloleks 3.0*. Slovenian language resource repository CLARIN.SI.

Ljubešić, Nikola and Terčon, Luka and Čibej, Jaka. 2023. *The CLASSLA-Stanza model for morphosyntactic annotation of standard Slovenian 2.0*. Slovenian language resource repository CLARIN.SI.

Rupnik, Peter and Ljubešić, Nikola and Porupski, Ivan and Verdonik, Darinka. 2024. *wav2vecbert2-filledPause (Revision 5e75061)*. Hugging Face.

Štravs, Miha and Dobrovoljc, Kaja and Bezgovšek, Luka. 2025. *Service for querying dependency treebanks Drevesnik 1.2*. Slovenian language resource repository CLARIN.SI.

Terčon, Luka and Čibej, Jaka and Ljubešić, Nikola. 2023. *The CLASSLA-Stanza model for lemmatisation of standard Slovenian 2.0*. Slovenian language resource repository CLARIN.SI.

Verdonik, Darinka. 2020. *Dialogue act annotated spoken corpus GORDAN 1.0 (transcription)*. Slovenian language resource repository CLARIN.SI.

Verdonik, Darinka and Bizjak, Andreja and Žgank, Andrej and Bernjak, Mitja and Antloga, Špela and Majhenič, Simona and Čakš, Peter and Pucer, Matevž and Cvetko, Mitja and Zelenik, Marijana and Pavlič, Jani and Dobrišek, Simon and Križaj, Janez and Strle, Gregor and Ivanovska, Marija and Grm, Klemen and Bajec, Marko and Lebar Bajec, Iztok and Jelovšek, Tjaša and Lokovšek, Jure and Longyka, Jure and Trojar, Mitja and Žganec Gros, Jerneja and Mihelič, Aleš and Vesnicer, Boštjan and Dretnik, Naum and Bordon, David. 2023a. *ASR database ARTUR 1.0 (audio)*. Slovenian language resource repository CLARIN.SI.

Darinka Verdonik, Kaja Dobrovoljc, Peter Rupnik, Nikola Ljubešić, Simona Majhenič, Jaka Čibej, Thomas Schmidt, and Jasna Vidinić. 2026. Training corpus of spoken slovenian ROG 1.1. Slovenian language resource repository CLARIN.SI.

Verdonik, Darinka and Zwitter Vitez, Ana and Zemljarič Miklavčič, Jana and Krek, Simon and Erjavec, Tomaž and Potočnik, Tomaž and Bizjak, Andreja and Žgank, Andrej and Bernjak, Mitja and Antloga, Špela and Majhenič, Simona and Čakš, Peter and Pucer, Matevž and Cvetko, Mitja and Pavlič, Jani and Dobrišek, Simon and Križaj, Janez and Bajec, Marko and Lebar Bajec, Iztok and Jelovšek, Tjaša and Trojar, Mitja and

Dretnik, Naum and Bordon, David and VideoLectures.NET and Križaj, Janez. 2024. *Spoken corpus Gos 2.1 (audio, video)*. Slovenian language resource repository CLARIN.SI.

Verdonik, Darinka and Zwitter Vitez, Ana and Zemljarič Miklavčič, Jana and Krek, Simon and Stabej, Marko and Erjavec, Tomaž and Potočnik, Tomaž and Sepesy Maučec, Mirjam and Majhenič, Simona and Žgank, Andrej and Bizjak, Andreja and Gril, Lucija and Dobrišek, Simon and Križaj, Janez and Bajec, Marko and Lebar Bajec, Iztok and Jelovšek, Tjaša and Trojar, Mitja and Bernjak, Mitja and Dretnik, Naum and Strle, Gregor and Dobrovoljc, Kaja and Ljubešić, Nikola and Rupnik, Peter. 2023b. *Spoken corpus Gos 2.1 (transcriptions)*. Slovenian language resource repository CLARIN.SI.

Zeman, Daniel and others. 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.

Zwitter Vitez, Ana and Zemljarič Miklavčič, Jana and Krek, Simon and Stabej, Marko and Erjavec, Tomaž. 2021. *Spoken corpus Gos 1.1*. Slovenian language resource repository CLARIN.SI.