Gos 2: A New Reference Corpus of Spoken Slovenian

Darinka Verdonik, Kaja Dobrovoljc, Tomaž Erjavec, Nikola Ljubešić
University of Maribor, University of Ljubljana, Jožef Stefan Institute
Koroška 46, SI-2000 Maribor, Aškerčeva 2, SI-1000 Ljubljana, Jamova 39, SI-1000 Ljubljana
darinka.verdonik@um.si, kaja.dobrovoljc@ff.uni-lj.si, tomaz.erjavec@ijs.si, nikola.ljubesic@ijs.si

Abstract

This paper introduces a new version of the Gos reference corpus of spoken Slovenian, which was recently extended to more than double the original size (300 hours, 2.4 million words) by adding speech recordings and transcriptions from two related initiatives, the Gos VideoLectures corpus of public academic speech, and the Artur speech recognition database. We describe this process by first presenting the criteria guiding the balanced selection of the newly added data and the challenges encountered when merging language resources with divergent designs, followed by the presentation of other major enhancements of the new Gos corpus, such as improvements in lemmatization and morphosyntactic annotation, word-level speech alignment, a new XML schema and the development of a specialized online concordancer.

Keywords: spoken corpus, speech database, corpus compilation, corpus annotation, Slovenian language

1. Introduction

Spoken corpora, transcribed and annotated collections of spoken language recordings, are a vital resource for linguistic research and language technology development. They provide authentic, real-world linguistic data, essential for improving speech recognition, dialogue systems, and our understanding of human communication in general.

For Slovenian language in particular, a South Slavic language with 2 million speakers, the first reference corpus of spoken Slovenian, known as Gos, was initially released in 2011, containing approximately 113 hours (1 million transcribed words) of spontaneous speech in various everyday situations, carefully balanced according to discourse type, domain, channel and speaker demographics (Zwitter Vitez et al., 2013; Verdonik et al., 2013). Although Gos has played a pivotal role in advancing linguistic research of spoken Slovenian (Verdonik and Maučec, 2017; Dobrovoljc, 2017) and the development of language resources for Slovenian in general (e.g., Dobrovoljc and Nivre, 2016; Fišer et al., 2020; Verdonik, 2023), it has not undergone significant expansions or improvements over time, and remained relatively small when compared to reference spoken corpora available for other languages (Love et al., 2017; Dobrushina, 2022).

To address this gap and establish a stronger empirical basis for future linguistic investigations of spoken Slovenian, we have recently created an extended and enhanced version of the Gos corpus as part of the project Development of Slovenian in a Digital Environment (DSDE).[1] We present the results of this activity in the continuation of this paper by describing the newly added data (Section 2) and its unification (Section 3), the improvements in lemmatization and tagging (Section 4), addition of word-alignment information (Section 5), a new version of the TEI schema (Section 6), and the development of a new online concordancer (Section 7).

2. Data selection

Efforts to compile speech databases aimed at training automatic speech recognition systems (e.g. Kibria et al. 2022) often run in parallel with the creation of reference speech corpora intended for linguistic research (e.g. Love et. al., 2017; Schmidt, 2012; Kopřivová et al., 2017). Despite their separate objectives, both communities share the common goal of seeking high-quality speech recordings and transcriptions across diverse contexts. In line with this, there is a growing body of work leveraging the same data for both technology development and linguistic research (Schuppler et al.; 2017), especially in low-resource communities that are unable to meet substantial investments required to develop large scale speech corpora.

This was also the motivation behind expanding the Gos reference corpus Slovenian through the inclusion of data from two ASR-oriented speech databases: Gos VideoLectures (Verdonik, 2018) and Artur (Verdonik et al., 2022). Among other databases available for spoken Slovenian (Mihelič et al., 2003; Žgank et al., 2014), these were selected due to their adherence to our initial criteria to capture (1) authentic spoken situations (no read texts), which (2) have undergone manual transcription or transcription accuracy checks, and (3) are freely accessible for public use. The result of these efforts is a unified resource with improved lemmatization and tagging, which is made available to the linguistic community through a new web concordancer, and is therefore much easier to access and use than separate, diverse datasets.

2.1 Gos VideoLectures

Gos VideoLectures is a corpus of 22 hours of audio recordings of academic speech aimed at training continuous speech recognition in the Slovenian language, facilitating phonetic research, and supporting research on Slovenian academic discourse in general. As implied by the name, the

[1] https://rsdo.slovenscina.eu/

LREC-COLING 2024, pages 7825–7830

corpus was designed as a potential extension of the Gos 1 reference corpus, therefore all 55 lectures with transcriptions (Verdonik et al., 2019) and audio recordings (VideoLectures.NET, 2019) from the Gos VideoLectures corpus have been included in the Gos 2 extension.

2.2 Artur

In 2020, the DSDE project initiated an extensive recording campaign in Slovenia aimed at amassing a 1,000-hour database for automatic speech recognition purposes. This resulted in the Artur ASR database with transcriptions (Verdonik et al., 2023b) and recordings (Verdonik et al., 2023a) that comprise approximately 500 hours of read sentences, 200 hours of parliamentary speech with manual transcription, and 300 hours of recordings from public and private settings, 100 of which include manual transcriptions.

To ensure that Gos remains a balanced and representative reference corpus of spontaneous speech, only a selection of transcriptions was included in Gos 2, namely all 62 hours of transcribed recordings of public events (conferences, workshops, round tables, interviews, education videos etc.), all 61 hours of transcribed private events (monologue descriptions or explanations, free dialogue conversations), and a 62-hour subset of transcribed parliamentary speech.[2]

3. Data integration and unification

Data integration and unification was performed on transcriptions only. Audio files remained the same as in the original data releases (VideoLectures.NET, 2019; Verdonik et al., 2023a). For the Gos 1 subcorpus, audio files have legal restrictions for public use. They can be listened to through the web concordancer, but not freely downloaded as is the case with the Artur and the Gos Videolectures audio files.

As shown in Table 1, extending the original Gos 1 corpus with new data from the two resources described above resulted in a significant improvement of the corpus with respect to size and the diversity of speakers and speech events, which, in addition to the 2004-2010 recording period of the Gos 1 corpus, now also include recordings and topics from the period up to 2022.

As shown in Table 2, Gos 2 also maintains the speech event diversity and balance of the Gos 1 corpus with respect to public and non-public speech, however, with the inclusion of the new data from Gos VideoLectures and Artur, Gos 2 introduces a larger bias towards public speech settings. Authentic private dialogues and multilogues, as well as non-private interactions occurring at workplace, service-related contexts, and sales, remain unique to Gos 1 subset and thus represent a smaller share of Gos 2 corpus in comparison to its predecessor.

	Events	Speakers	Hours	Words
Gos 1	287	1,560	112	1,035,086
Gos VL	55	90	22	178,830
Artur	1,192	532	185	1,250,389
Gos 2	**1,534**	**2,182**	**319**	**2,464,305**

Table 1: The size of the new Gos 2 corpus and its subsets with respect to the number of speech events and speakers, the length of the recordings and the number of words transcribed.

	No. of words in thousands	%
Public	**1667**	**68%**
TV, informative broadcast	104	4%
TV, entertaining show	105	4%
Radio talk or show	111	5%
Radio interview	252	10%
Public event	295	12%
Lecture	222	9%
School lesson	111	5%
Parliament	462	19%
Non-public non-private	**162**	**7%**
Workplace or education	86	4%
Services	57	2%
Sales	19	1%
Non-public private	**632**	**26%**
Monologue	213	9%
Dialogue or multilogue	419	17%
Total	**2.4 mio**	100%

Table 2: The structure of Gos 2 according to interaction types.

The process of merging the data from the three different resources, however, was far from trivial, as several differences between the three corpora had to be considered. First, the three databases adopted divergent and only partially overlapping approaches to text categorization with respect to metadata related to the speech events (e.g. speech type, time and place of the recording) and speaker demographics (e.g. gender, education, region, age), for which a special mapping procedure was devised to map the metadata categories and values from all three subcorpora to a new, unified taxonomy (Verdonik et al., 2022).

To our advantage, both Gos VideoLectures and Artur transcriptions closely adhered to the guidelines for tokenization and orthography used in the original Gos 1 corpus, including the two-level transcription approach encompassing pronunciation-based spelling and its normalized alternative in standard orthography. However, the Artur database, designed

[2] Parliamentary speech to be included in Gos 2 was limited to max. 4,000 words per speaker to avoid the parliamentary data predominating in the new edition of the corpus.

for specific end-applications, introduced some variation. It features shorter, prosody-driven utterance segmentation and also include punctuation and capitalization – elements absent from the Gos 1 and Gos VideoLectures datasets. With the exception of casing, which was resolved automatically by removing the casing information from the beginnings of utterances in Artur datasets, we opted not to resolve segmentation and punctuation-related differences in this initial release to allow for user-driven suggestions on the optimal transcription representations.

Finally, the encoding of the three datasets had to be unified, as the Gos 1 and Gos VideoLectures followed older versions of the TEI XML schema. The unification included the already mentioned typology but also e.g. harmonization of text titles, their source and setting descriptions, coding of disfluencies, and different treatments of word capitalization.

4. Data annotation

Given that different tools have been used for automatic linguistic annotation of Gos 1 and Gos VideoLectures corpora, and no annotations were featured in the Artur database, we decided to follow a complete reannotation of the new Gos 2 corpus using the CLASSLA-Stanza NLP tool (Ljubešić and Dobrovoljc, 2019; Terčon and Ljubešić, 2023) to annotate transcriptions with lemmas, part-of-speech tags and other morphosyntactic features according to the MULTEXT-East annotation scheme (Erjavec, 2010), commonly used in the reference corpora of Slovenian.

However, rather than applying the default models trained on standard and non-standard written data, which have been shown to perform suboptimally when confronted with speech transcriptions (Dobrovoljc and Martinc, 2018), we developed speech-specific models based on a series of experiments combining the available manually annotated datasets for Slovenian, i.e. the ssj500k training corpus of standard written Slovenian (Krek et al., 2021), the JANES-tag corpus of non-standard written Slovenian (Erjavec et al., 2019), and the SST treebank of spoken Slovenian, a manually annotated sample of the original Gos 1 corpus (Dobrovoljc and Nivre, 2016).

As shown in Table 1,[3] which gives the performance of the different CLASSLA-Stanza v1.2 models when evaluated on the SST test set, the tool performs better if SST spoken data is featured in the training procedure. Specifically, best performance for the morphosyntactic tagging was achieved using the combination of all three datasets (F1 94.02), while the best lemmatization performance was achieved by a model trained on a combination of spoken and standard written data alone (F1 98.33). Therefore, these two CLASSLA-Stanza models have been used in the final annotation of Gos 2 corpus.

Training data	Tag	Lemma
classla_standard (ssj500k)	82.02	97.24
classla_non-standard (JANES)	84.21	93.68
SST	91.74	97.87
SST+ssj500k	93.85	**98.33**
SST+JANES	92.63	98.22
SST+ssj500+JANES	**94.02**	98.27

Table 3: CLASSLA-Stanza performance (micro F1) on the spoken SST test set using different training scenarios for lemmatization and PoS tagging.

5. Speech-to-text alignment

Gos 2 transcriptions are aligned with the speech recordings on the level of utterances and words. While utterance-level alignment was already performed manually as part of the compilation of the original Gos 1, as was the case with Gos VideoLectures and Artur corpora, there was no word-level alignment available.

To increase the usefulness of the Gos corpus for word-level research, we have performed word alignment of the transcripts with the speech recordings by following the Kaldi (Povey et al., 2011) recipe for forced alignment via an acoustic model based on parliamentary speech and transcript data. We have encoded the word alignment information in form of time offsets of the beginning of each word only if the alignment procedure offered an alignment for every word in a segment. With this procedure, 79% of segments received alignment information. The reason for a rather low percentage of successful alignments is primarily to be followed to varying quality of the recordings inside the Gos 1 section of the Gos 2 corpus. While 98% of the segments in the Artur section of the Gos 2 corpus were successfully aligned, the number plummets to around 20% for Gos 1 and around 40% for Gos VideoLectures.

The alignment of the speech signal was very successful in the Artur section of the Gos 2 corpus due to the high quality of recordings, as well as low percentage of overlapping speech. Given that the word alignment of the Gos Videolectures and the Gos 1 sections of the Gos 2 corpus, the usefulness of the only partial word alignment information is questionable and requires for stronger word alignment techniques to be applied in the future.

6. Data release

Gos 2 is encoded in XML following the TEI Guidelines[4] using the TEI parameterization as recommended by the CLARIN.SI research infrastructure[5]. The corpus is encoded as one XML

[3] Full results available at: https://github.com/clarinsi/classla-spoken.

[4] https://tei-c.org/release/doc/tei-p5-doc/en/html

[5] https://github.com/clarinsi/TEI-schema, see also https://www.clarin.si/repository/xmlui/page/data#tei.

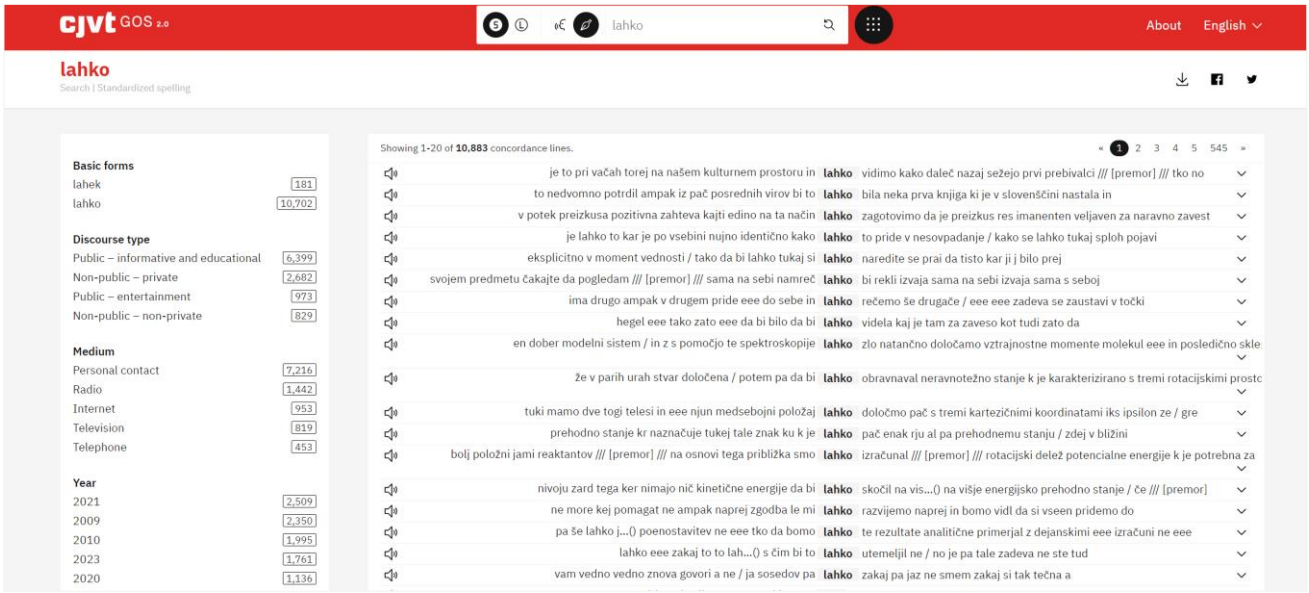

Figure 1: List of concordances for the word ‘lahko’ in the new Gos 2 online interface.

document composed of the top-level corpus file and the individual files of the five subcorpora, as Artur data was split into the private, public and parliamentary subsets. The top-level file contains the TEI header giving the common metadata for the corpus as well as the list of all speakers with their metadata, while the 1,534 files of the subcorpora contain the annotated speeches.

Gos 2 transcriptions are openly available under the CC BY-SA 4.0 license in the CLARIN.SI repository of language resources and tools (Verdonik et al., 2023c). Because of the restricted copyright status of the original Gos 1 audio recordings, which are only accessible for research purposes, only the audio recordings from the Gos VideoLectures and Artur subsets can be freely accessed under the CC-BY-NC-ND 4.0 and CC-BY-SA 4.0 licenses, respectively.

7. Online Concordancer

In addition to the CLARIN.SI concordancers aimed at researchers and other expert users, a specialized web interface[6] was developed to promote and facilitate the use of the Gos 2 corpus among general public and other language professionals. The interface, illustrated in Figure 1, follows the design and the functionalities of the concordancer of the Gigafida reference corpus of written Slovenian (Krek et al., 2020; Arhar Holdt et al., 2019), and extends it with several new speech-specific features.

Specifically, users can browse the corpus by typing in a word or a phrase based on either standardized or pronunciation-based spelling (simple search) or by specifying its grammatical and contextual features (advanced search) to obtain a list of concordances and associated utterance recordings (Figure 1) with frequencies across different metadata categories, which can also be used for further filtering of the results by event type, speaker demographics and other relevant criteria. Clicking on a specific concordance displays the preceding and the following utterance, more information on the event and the speaker, as well as information on grammatical features of the tokens in the utterance under investigations.

In addition to the ‘List’ function that produces a frequency list of the word forms matching the input criteria (e.g. a list of all pronunciation variations of a given standardized form), users can also benefit from a function that triggers identical queries in online interfaces of other reference language resources for Slovenian, such as the Gigafida written corpus, the Sloleks morphological dictionary, the Dictionary of Collocations and the Thesaurus.

The source code of the new concordancer is freely available on the CLARIN.SI GitHub repository and consists two main parts: the web application (.ASP NET Core 6) and the console application (.NET 6).[7]

8. Conclusion

We have presented Gos 2, a new version of the Gos reference corpus of spoken Slovenian, which has recently been extended to more than double the original size and improved with respect to linguistic annotation, speech-to-text alignment, data formatting and online accessibility.

Although this work represents an important and much anticipated addition to the Slovenian language resource landscape, especially in the realm of corpus linguistic research, it is merely the initial stride in the

[6] https://viri.cjvt.si/gos/

[7] https://github.com/clarinsi/rsdo_gos

ongoing evolution of the corpus. Most notably, we aim to consolidate the structure and the content of the Gos corpus based on the insights and computational tools generated through the MEZZANINE project (2022-2025) focusing on foundational research for the development of spoken language resources and speech technologies for Slovenian.

9. Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful suggestions. The data integration and unification and the development of the new online concordancer for the Gos 2 corpus was performed within the project Development of Slovenian in a Digital Environment, financed by the Republic of Slovenia and European Regional Fund. Data annotation and text-to-speech alignment was performed within the research project titled Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language (J7-4642) and the research programme titled Language resources and technologies for Slovene (P6-0411), both financed by the Slovenian Research and Innovation Agency (ARIS).

Bibliographical References

Arhar Holdt, Š., Dobrovoljc, K., & Logar, N. (2019). Simplicity matters: user evaluation of the Slovene reference corpus*. Language resources and evaluation*, 53(1), 173-190.

Dobrovoljc, K. (2017). Multi-word discourse markers and their corpus-driven identification: The case of MWDM extraction from the reference corpus of spoken Slovene*. International journal of corpus linguistics,* 22(4), 551-582.

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

Dobrovoljc, K., & Nivre, J. (2016). The Universal Dependencies treebank of spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1566-1573).

Dobrushina, N., & Sokur, E. (2022). Spoken Corpora of Slavic Languages. *Russian Linguistics*, 46(2), 77-93.

Erjavec, T. (2012). MULTEXT-East: morphosyntactic resources for Central and Eastern European languages*. Language resources and evaluation*, 46, 131-142.

Fišer, D., Ljubešić, N., & Erjavec, T. (2020). The Janes project: language resources and tools for Slovene user generated content*. Language resources and evaluation,* 54, 223-246.

Kibria, S., Samin, A. M., Kobir, M. H., Rahman, M. S., Selim, M. R., & Iqbal, M. Z. (2022). Bangladeshi Bangla speech corpus for automatic speech recognition research. *Speech Communication*, 136, 84-97.

Kopřivová, M., Komrsková, Z., Lukeš, D., & Poukarová, P. (2017). Korpus ORAL: sestavení, lemmatizace a morfologické značkování. *Korpus–gramatika–axiologie,* 15, 47-67.

Krek, S., Holdt, Š. A., Erjavec, T., Čibej, J., Repar, A., Gantar, P., ... & Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 3340-3345).

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).

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), 319-344.

Mihelič, F., Gros, J., Dobrišek, S., Žibert, J., & Pavešić x0107, N. (2003). Spoken language resources at LUKS of the University of Ljubljana*. International Journal of Speech Technology*, 6, 221-232.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF).* IEEE Signal Processing Society.

Schmidt, T. (2014). The database for spoken German-DGD2. In *LREC* (pp. 1451-1457).

Schuppler, B., Hagmüller, M., & Zahrer, A. (2017). A corpus of read and conversational Austrian German. *Speech Communication*, 94, 62-74.

Terčon, L., & Ljubešić, N. (2023). CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages. arXiv preprint arXiv: 2308.04255.

Verdonik, D. (2018). Korpus in baza Gos Videolectures. In V *Zbornik 11. konference Jezikovne tehnologije in digitalna humanistika* (pp. 265-268).

Verdonik, D. (2023). Annotating dialogue acts in speech data: Problematic issues and basic dialogue act categories. *International Journal of Corpus Linguistics*, 28(2), 144-171.

Verdonik, D., & Maučec, M. S. (2017). A speech corpus as a source of lexical information*. International Journal of Lexicography*, 30(2), 143-166.

Verdonik, D., Bizjak, A., Žgank, A., & Dobrišek, S. Metapodatki o posnetkih in govorcih v govornih virih: primer baze Artur. In *Zbornik conference Jezikovne tehnologije in digitalna humanistika.*

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, 1031-1048.

Žgank, A., Vitez, A. Z., & Verdonik, D. (2014, May). The Slovene BNSI Broadcast News database and reference speech corpus GOS: Towards the uniform guidelines for future work. In *LREC* (pp. 2644-2647).

10. Language Resource References

Erjavec, Tomaž; et al., 2019, CMC training corpus Janes-Tag 2.1, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1238.

Krek, Simon; et al., 2021, Training corpus ssj500k 2.3, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1434.

Verdonik, Darinka; et al., 2021, Spoken corpus Gos VideoLectures 4.2 (transcription), Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1444.

Verdonik, Darinka; et al., 2023a, ASR database ARTUR 1.0 (audio), Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1776.

Verdonik, Darinka; et al., 2023b, ASR database ARTUR 1.0 (transcriptions), Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1772.

Verdonik, Darinka; et al., 2023c, Spoken corpus Gos 2.1 (transcriptions), Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1863.

VideoLectures.NET, 2019, Spoken corpus Gos VideoLectures 4.0 (audio), Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1222.

Zwitter Vitez, Ana; Zemljarič Miklavčič, Jana; Krek, Simon; Stabej, Marko and Erjavec, Tomaž, 2013, Spoken corpus Gos 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1040.