

Gos 2: A New Reference Corpus of Spoken Slovenian

Darinka Verdonik, Kaja Dobrovoljc, Tomaž Erjavec, Nikola Ljubešić

Presented at LREC-Coling 2024

Introduction

- Evolution and enhancements to the corpus of spoken Slovenian - Gos
- Authentic linguistic data is important for improving speech technologies and understanding human communication

Background

- The Slovenian reference spoken corpus Gos first released in 2011, 113 hours of speech
- Remained relatively small, limiting linguistic research and technology development

Objective

- To extend and enhance the Gos corpus to provide a stronger empirical basis for linguistic investigations of spoken Slovenian
- Integrating data from other spoken resources, i.e. Gos Videolectures, Artur

Data selection

- Data sources used:
 - authentic spoken situations (no read texts)
 - manual transcription or transcription accuracy checks
 - freely accessible for public use
- Gos VideoLectures: 22 hours of academic speech recordings
- Artur:
 - 500 hours of read sentences (not used in Gos 2)
 - 200 hours of parliamentary speech (partially used in Gos 2)
 - 300 hours of public and private recordings, but only 100 hours with transcriptions and usable for the Gos 2

Data integration and unification

- Harmonization of metadata: speech events, speaker demographics
- Transcriptions standards: segmentation, orthography and punctuation
- Resulting corpus enhancement: a richer resource, allows for broader research

Data annotation

- Linguistic re-annotation using CLASSLA-Stanza NLP tool
- MULTEXT-East scheme
- Speech-specific models were developed to optimize annotation accuracy

Speech-to-text alignment

- Word-level alignment to synchronize transcriptions with recordings
- Improved usability for research requiring precise word-level analysis

Data release

- Text Encoding Initiative (TEI) Guidelines
- A single XML document comprising a top-level corpus file and individual files for subcorpus components
- CC BY-SA 4.0 license
- Audio:
 - Gos 1: restricted & limited to research
 - Gos VideoLectures: CC-BY-NC-ND 4.0
 - Artur: CC-BY-SA 4.0
- The CLARIN.SI repository of language resources and tools

lahko

Search | Standardized spelling



Basic forms

lahek 181

lahko 10,702

Discourse type

Public – informative and educational 6,399

Non-public – private 2,682





















Public – entertainment 973

Non-public – non-private 829

Medium

Personal contact 7,216

Showing 1-20 of **10,883** concordance lines. « 1 2 3 4 5 545 »

	je to pri vačah torej na našem kulturnem prostoru in	lahko	vidimo kako daleč nazaj sežejo prvi prebivalci /// [premor] /// tko no	
	to nedvomno potrdil ampak iz pač posrednih virov bi to	lahko	bila neka prva knjiga ki je v slovenščini nastala in	
	v potek preizkusa pozitivna zahteva kajti edino na ta način	lahko	zagotovimo da je preizkus res imanenten veljaven za naravno zavest	
	je lahko to kar je po vsebini nujno identično kako	lahko	to pride v nesovpadanje / kako se lahko tukaj sploh pojavi	
	eksplicitno v moment vednosti / tako da bi lahko tukaj si	lahko	naredite se prai da tisto kar ji j bilo prej	
	svojem predmetu čakajte da pogledam /// [premor] /// sama na sebi namreč	lahko	bi rekli izvaja sama na sebi izvaja sama s seboj	
	ima drugo ampak v drugem pride eee do sebe in	lahko	rečemo še drugače / eee eee zadeva se zaustavi v točki	
	hegel eee tako zato eee da bi bilo da bi	lahko	videla kaj je tam za zaveso kot tudi zato da	
	en dober modelni sistem / in z s pomočjo te spektroskopije	lahko	zlo natančno določamo vztrajnostne momente molekul eee in posledično skle	
	... / ... / ... / potem na da bi	lahko	obravnaval neravnotežno stanje k je karakterizirano s tremi rotacijskimi prostc	

Online concordancer

- User-friendly interface
- Search and navigation
- Concordance display
- Integration with other resources

Conclusion

- Gos 2: expanded size, improved annotation, alignment, and accessibility
- Continue to refine the corpus and support linguistic research and the development of speech technologies

Thank you!