<..> mezzanine

Govorjeni jezik med raziskovanjem in tehnologijo

Zbornik povzetkov

<..> uredila Darinka Verdonik in Nikola Ljubešić

Univerzitetna založba
Univerze v Mariboru

Univerza v Mariboru

Fakulteta za elektrotehniko,
računalništvo in informatiko

Govorjeni jezik med raziskovanjem in tehnologijo

Zbornik povzetkov

Urednika
Darinka Verdonik
Nikola Ljubešić

September 2025

Naslov / *Title*
Govorjeni jezik med raziskovanjem in tehnologijo
Spoken Language Between Research and Technology

Podnaslov / *Subtitle*
Zbornik povzetkov
Book of Abstracts

Urednika / *Editors*
Darinka Verdonik
(Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko)

Nikola Ljubešić
(Inštitut Jožef Stefan)

Lektoriranje / *Language editing*
Darinka Verdonik
(Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko)

Tehnični urednik / *Technical editor*
Jan Perša
(Univerza v Mariboru, Univerzitetna založba)

Oblikovanje ovitka / *Cover designers*
Jan Perša
(Univerza v Mariboru, Univerzitetna založba)

Grafika na ovitku / *Cover graphics*
Logotip konference MEZZANINE, Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, 2025

Konferenca / *Conference*
Zaključna konferenca MEZZANINE: Govorjeni jezik med raziskovanjem in tehnologijo

Datum in kraj / *Date & location*
18. september 2025, Ljubljana, Slovenija

Programski odbor / *Program committee*
Darinka Verdonik (Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko) in Nikola Ljubešić (Inštitut Jožef Stefan)

Organizacijski odbor / *Organizing committee*
Špela Antloga (Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko), Sara Kosi (Univerza v Ljubljani, Fakulteta za računalništvo in informatiko), Nejc Robida (Univerza v Ljubljani, Filozofska fakulteta) in Jaka Čibej (Univerza v Ljubljani, Filozofska fakulteta.

Založnik / *Published by*
Univerza v Mariboru
Univerzitetna založba
Slomškov trg 15, 2000 Maribor, Slovenija
https://press.um.si, zalozba@um.si

Izdajatelj / *Issued by*
Univerza v Mariboru
Fakulteta za elektrotehniko, računalništvo in informatiko
Koroška cesta 46, 2000 Maribor, Slovenija
https://feri.um.si, feri@um.si

Izdaja
Edition Prva izdaja

Vrsta publikacije
Publication type E-knjiga

Dostopno na
Available at https://press.um.si/index.php/ump/catalog/book/1052

Izdano
Published Maribor, Slovenija, september 2025

CIP - Kataložni zapis o publikaciji
Univerzitetna knjižnica Maribor

81'373 : 004.9(0.034.2)

ZAKLJUČNA konferenca Mezzanine (2025 ; Maribor)
Govorjeni jezik med raziskovanjem in tehnologijo [Elektronski vir] : zbornik povzetkov / uredila Darinka Verdonik in Nikola Ljubešić. - 1. izd. - E-publikacija. - Maribor : Univerza v Mariboru, Univerzitetna založba, 2025

Način dostopa (URL): https://press.um.si/index.php/ump/catalog/book/1052

ISBN 978-961-299-050-3 (PDF)
doi: 10.18690/um.feri.9.2025
COBISS.SI-ID 248221699

ISBN 978-961-299-050-3 (pdf)

DOI https://doi.org/10.18690/um.feri.9.2025

Cena
Price Brezplačni izvod

Odgovorna oseba založnika
For publisher Prof. dr. Zdravko Kačič,
rektor Univerze v Mariboru

Citiranje
Attribution Verdonik D., Ljubešić, N. (ur.). (2025). *Govorjeni jezik med raziskovanjem in tehnologijo: zbornik povzetkov*. Univerza v Mariboru, Univerzitetna založba. doi: 10.18690/um.feri.9.2025

Univerzitetna založba
Univerze v Mariboru

Kazalo

GOVORJENI JEZIK MED RAZISKOVANJEM IN TEHNOLOGIJO:
ZBORNIK POVZETKOV
D. Verdonik, N. Ljubešić (ur.)

Univerzitetna založba
Univerze v Mariboru

Croatian Child Language Corpora in Childes: Public Resources for Research on Early Language Development

GORDANA HRŽICA
University of Zagreb, Faculty of Education and Rehabilitation Research, Department of Speech and Language Pathology, Zagreb, Croatia
gordana.hrzica@erf.unizg.hr

We present publicly available corpora of Croatian child language in the CHILDES database within the TalkBank infrastructure. The datasets include spontaneous and elicited speech of children aged 1;8–10;0, transcribed in CHAT and enriched with manual and automatic linguistic annotation. They provide valuable resources for research on language acquisition, narrative development, and language impairment, and are freely accessible through TalkBank.

1 Croatian child language corpora

We present a set of publicly available corpora of Croatian child language hosted in the CHILDES database within the TalkBank infrastructure. The corpora cover both spontaneous and elicited speech data from typically developing children aged between 1;8 and 10;0. The datasets include the longitudinal Kovačević corpus, the cross-sectional Croatian Corpus of Preschool Child Language (CCPCL), narrative retellings in the Croatian MAIN corpus, and the TKH Frog Story corpus. The Kovačević corpus consists of longitudinal recordings of three monolingual children

from the onset of speech to 3;4, richly annotated for morphosyntactic development. The CCPCL corpus (Croatian Corpus of Preschool Child Language) includes cross-sectional data from over 90 children aged 2;6 to 6;0, elicited through standardized conversational settings. The Croatian MAIN corpus features narrative data collected using the Multilingual Assessment Instrument for Narratives (MAIN; Gagarina et al. 2019; Hržica & Kuvač Kraljević 2019), with materials in both telling and retelling tasks. Finally, the TKH Frog Story corpus provides retellings of the wordless picture book Frog, Where Are You?, offering insights into discourse-level abilities in preschool and early school-aged children. All corpora are encoded in CHAT format and transcribed according to the CHILDES conventions.

2 Annotation and applications

The corpora have been enriched with multiple layers of annotation. Manual coding includes phonological, morphosyntactic, and semantic error tagging, as well as selected types of disfluencies. Automatic linguistic annotation has been performed using the CLAN tools, including morphological tagging via the MOR program. The corpora support a wide range of applications in the study of language acquisition, narrative development, and language impairment. All datasets are freely available through TalkBank:

https://talkbank.org/childes/access/Slavic/Croatian/Kovacevic.html
https://talkbank.org/childes/access/Slavic/Croatian/CCPCL.html
https://talkbank.org/childes/access/Slavic/Croatian/MAIN.html
https://talkbank.org/childes/access/Frogs/Croatian-TKH.html

Keywords: Croatian, child language, corpus linguistics, error tagging, narrative development, morphosyntactic annotation

Hrvaški korpusi otroškega jezika v zbirki Childes: javni viri za raziskave zgodnjega jezikovnega razvoja

Predstavljamo javno dostopne korpuse hrvaškega otroškega jezika, ki so vključeni v podatkovno zbirko CHILDES znotraj infrastrukture TalkBank. Korpusi zajemajo tako spontani kot vodeni govor otrok med 1. in 10. letom. Korpusi so označeni na

več plasteh. Ročno označevanje je zajemalo fonološko, oblikoslovno in semantično označevanje napak ter izbrane tipe netekočnosti. Avtomatsko jezikoslovno označevanje je bilo izvedeno z orodji CLAN, med drugim s programom MOR za morfološko razčlenjevanje. Korpusi podpirajo raziskave usvajanja jezika, razvoja kompetenc za pripovedovanje zgodb in raziskave motenj govora.

Ključne besede: hrvaščina, otroški jezik, korpusno jezikoslovje, označevanje napak, razvoj kompetenc za pripovedovanje zgodb, oblikoslovno označevanje

GOVORJENI JEZIK MED RAZISKOVANJEM IN TEHNOLOGIJO:
ZBORNIK POVZETKOV
D. Verdonik, N. Ljubešić (ur.)

Univerzitetna založba
Univerze v Mariboru

The Parlaspeech V3 Collection of Spoken Parliamentary Corpora From the Croatian, Czech, Polish and Serbian Parliament

NIKOLA LJUBEŠIĆ, PETER RUPNIK, IVAN PORUPSKI,
TAJA KUZMAN PUNGERŠEK
Jožef Stefan Institute, Ljubljana, Slovenia
nikola.ljubesic@ijs.si, peter.rupnik@ijs.si, ivan.porupski@ijs.si, taja.kuzman@ijs.si

ParlaSpeech is a collection of spoken parliamentary corpora, currently spanning the Croatian, Czech, Polish and Serbian parliaments, built by automatically aligning speech recordings to the corresponding ParlaMint transcripts, ensuring also the availability of the rich ParlaMint metadata. The corpus collection has been automatically enriched on multiple linguistic and paralinguistic levels, using both text and speech as the source for our automatic annotations.

1 Enrichment levels

The first enrichment level, ParlaSpeech-Ling, is performed on the textual modality in the form of linguistic annotation that follows the Universal Dependencies format. The second level of enrichment, ParlaSpeech-Pause, is performed on the spoken modality by identifying filled pauses with a fine-tuned speech transformer model. The following two layers are applied at this point only to the Croatian and the Serbian corpora. The third layer, ParlaSpeech-Align, additionally aligns the spoken and textual modality with highly accurate grapheme-level and word-level alignments

based on an acoustic-model-based forced aligner. The final layer, ParlaSpeech-Stress, adds the primary stress information to each multi-syllabic word via predicting primary stress on the spoken modality with a fine-tuned speech transformer model. The spoken modality prediction is post-processed to be aligned to a syllable nucleus on the textual modality for simpler downstream use of primary stress information.

2 Data formats and availability

All corpora are available as JSONL files with all available information encoded for computational use, as searchable corpora on the CLARIN.SI concordancers, and as TextGrid files to facilitate use by phoneticians and other speech specialists. Documentation on the encoding of the corpus in the JSONL format can be accessed here: https://clarinsi.github.io/parlaspeech/. A short tutorial on exploiting the Croatian corpus through the CLARIN.SI concordancer is available here: https://clarinsi.github.io/parlaspeech/concordancer/concordancer-guide.html.

Keywords: spoken parliamentary corpora, linguistic annotation, speech and text alignment, primary stress detection, filled pause detection

Zbirka govornih parlamentarnih korpusov parlaspeech V3 iz hrvaškega, češkega, poljskega in srbskega parlamenta

ParlaSpeech je zbirka govornih parlamentarnih korpusov, ki trenutno zajema govore iz hrvaškega, češkega, poljskega in srbskega parlamenta. Razvoj zbirke je zajemal avtomatsko poravnavo zvočnih posnetkov govorov s transkripcijami iz korpusov ParlaMint, iz katerih smo pridobili tudi bogate metapodatke. Zbirko korpusov smo nato obogatili z dodatnimi informacijami na več jezikovnih in parajezikovnih ravneh, pri čemer je avtomatsko označevanje temeljilo tako na besedilu kot na govoru.

Ključne besede: govorni parlamentarni korpusi, jezikoslovno označevanje, poravnava govora in besedila, zaznavanje glavnega naglasa, zaznavanje zapolnjenih premorov

GOVORJENI JEZIK MED RAZISKOVANJEM IN TEHNOLOGIJO:
ZBORNIK POVZETKOV
D. Verdonik, N. Ljubešić (ur.)

Univerzitetna založba
Univerze v Mariboru

Vloga občanske znanosti pri množičnem zbiranju govornih virov v slovenščini

ANDREJA BIZJAK
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor, Slovenija
andreja.bizjak@um.si

V prispevku obravnavamo inovativne pristope k pridobivanju spontanih govornih virov za slovenščino, kot so občanska znanost, množičenje, GWAP in Collect4NLP. Na podlagi analize treh evropskih pobud (CorCenCC, Anneta kõnet, Lahjoita puhetta) izpostavljamo pomen motivacije govorcev, pravne in etične varnosti ter vloge nacionalnih promocijskih kampanj, ter podajamo priporočila za zasnovo spletne platforme za pridobivanje govornih virov na daljavo.

1 Pregled stanja

Zaradi pomanjkanja spontano tvorjenih govornih virov za slovenščino smo pregledali inovativne pristope k njihovemu pridobivanju, kot so množičenje (Eskénazi in sod., 2013), občanska znanost (Mlinar, 2021), Games-With-A-Purpose in Collect4NLP (Lyding in sod., 2022). Izpostavili smo različne dejavnike, med njimi pravne in etične, ki so ključni za zasnovo trajnostnega in razširljivega sistema za pridobivanje govornih posnetkov. Med glavnimi izzivi ostaja pridobivanje zaupanja govorcev zaradi dvomov, povezanih s snemanjem in deljenjem govora ter zaščite osebnih podatkov (Rutten in sod., 2017). Nadaljnji izziv je doseči pristnost govora zaradi Labovega paradoksa opazovalca (Lindén in sod., 2022), ki ga lahko nekoliko

omilimo z vključevanjem daljših posnetkov nepripravljenega govora več govorcev v zasebnem okolju.

Na podlagi analize uspešnih pobud za izgradnjo korpusov manj razširjenih evropskih jezikov smo ugotovili, da ostaja ključen dejavnik uspeha ustrezno motiviranje govorcev. Analizirali smo primere pridobivanja posnetkov za korpus CorCenCC za valižanščino, kjer so zbrali več kot 11 milijonov besed, od tega 2,8 milijona v govorjeni obliki (Neale in sod., 2017; Knight in sod., 2020; 2021), Anneta kõnet za estonščino z več kot 100 urami govora, zbranih v osmih mesecih, in Lahjoita puhetta za finščino z več kot 4000 urami v enem letu (Lindén in sod., 2022). K uporabnikom usmerjene aplikacije so prostovoljcem omogočile bolj osebno, igrivo izkušnjo ter nadzor nad lastnimi posnetki. Namen vseh treh pobud sta bila podpora razvoju jezikovnih tehnologij in jezikovno ohranjanje, njihovo uspešnost pa so bistveno okrepile nacionalno podprte promocijske kampanje z duhovitimi in strateško načrtovanimi nastopi v medijih ter digitalnem in lokalnem okolju.

2 Priporočila

Odločitev za sodelovanje v projektih občanske znanosti je v veliki meri odvisna od dejavnikov notranje motivacije (Nov in sod., 2014), zato bi jih bilo v prihodnje smiselno sistematično kategorizirati in implementirati. Poleg teh je v analizo treba vključiti tudi demografske značilnosti in vrednote govorcev, ki pomembno vplivajo na njihovo pripravljenost za sodelovanje (Levontin idr., 2022). Priporočila za načrtovanje spletne platforme za zbiranje govornih virov vključujejo pripravo razumljivih in nedvoumnih navodil za snemanje ter oblikovanje nalog, ki so enostavne, a hkrati dovolj zanimive, da ne odvrnejo izkušenih uporabnikov. Ključno je vzpostaviti zaupanje udeležencev – zlasti glede anonimizacije in varstva podatkov – ter zagotoviti transparentnost glede namena zbiranja in uporabe posnetkov. Potrebna je večnivojska validacija kakovosti (pred, med in po oddaji posnetkov) ter vzpostavitev sistema usposabljanja in predtestiranja govorcev. Platforma naj bo oblikovana kot atraktivna, občanskemu raziskovalcu prijazna rešitev, ki mu omogoča vključevanje v različne faze raziskovalnega procesa in seznanitev z diseminacijo rezultatov. Uspešnost projekta se poveča, če je v lokalnem okolju prepoznan kot pobuda v javnem interesu, kar zahteva strateško komuniciranje in ciljno usmerjeno medijsko kampanjo, brez katere množične udeležbe v kratkem času ni mogoče doseči.

Ključne besede: pridobivanje govornih virov, občanska znanost, množičenje, spontani govor

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta ARIS Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik (J7-4642).

Literatura

Eskenazi, M., Levow, G. A., Meng, H., Parent, G., & Suendermann, D. (2013). *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment.* John Wiley & Sons.

Knight, D., Loizides, F., Neale, S., Anthony, L., & Spasić, I. (2021). *Developing computational infrastructure for the CorCenCC corpus: the national corpus of contemporary Welsh. Language Resources and Evaluation, 55*, 789–816. Pridobljeno 15. 5. 2025 s
https://link.springer.com/content/pdf/10.1007/s10579-020-09501-9.pdf

Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I., & Thomas, E. M. (2020). *The national corpus of contemporary Welsh: Project report| Y corpws cenedlaethol Cymraeg cyfoes: adroddiad y prosiect.* Pridobljeno 15. 5. 2025 s https://arxiv.org/abs/2010.05542

Levontin, L., Gilad, Z., Shuster, B., Chako, S., Land-Zandstra, A., Lavie-Alon, N., & Shwartz, A. (2022). Standardizing the assessment of citizen scientists' motivations: A motivational goal-based approach. *Citizen Science: Theory and Practice*, *7*(1). Pridobljeno 16. maja 2025 s Standardizing the Assessment of Citizen Scientists' Motivations: A Motivational Goal-Based Approach | Citizen Science: Theory and Practice

Lindén, K., Jauhiainen, T., Lennes, M., Kurimo, M., Rossi, A., Kurki, T., & Pitkänen, O. (2022). Donate Speech: *Collecting and Sharing a Large-Scale Speech Database for Social Sciences, Humanities and Artificial Intelligence Research and Innovation.* V CLARIN: the infrastructure for language resources (Digital Linguistics; Vol. 1). De Gruyter. doi: 10.1515/9783110767377-019

Lyding, V., Nicolas, L., & König, A. (2022). *About the applicability of combining implicit crowdsourcing and language learning for the collection of NLP datasets. V Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results within LREC 2022* (str. 46–57). Pridobljeno 15. maja 2025 s
https://aclanthology.org/2022.nidcp-1.8.pdf

Mlinar, Z. (2021). Kaj nam prinašata koncept in gibanje občanska znanost/Citizen Science? Uveljavljanje raziskovanja kot sestavine vsakdanjega življenja. *Casopis za Kritko Znanosti, Domisljijo in Novo Antropologijo (Journal for the Critique of Science, Imagination & New Anthropology), 49*(282).

Neale, S., Spasic, I., Needs, J., Watkins, G., Morris, S., Fitzpatrick, T., ... & Knight, D. (2017). *The CorCenCC crowdsourcing app: A bespoke tool for the user-driven creation of the national corpus of contemporary Welsh. V Corpus Linguistics Conference, Birmingham.* Pridobljeno 15. 5. 2025 s
https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper273.pdf

Nov, O., Arazy, O., & Anderson, D. (2014*). Scientists@Home: What Drives the Quantity and Quality of Online Citizen Science Participation?* PLoS ONE, 9(4), e90375. Pridobljeno 15. 5. 2025 s
https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0090375&type=printable

Rutten, M., Minkman, E., & van der Sanden, M. (2017). *How to get and keep citizens involved in mobile crowd sensing for water management? A review of key success factors and motivational aspects. Wiley Interdisciplinary Reviews: Water, 4*(4), e1218.
https://wires.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/wat2.1218

The Role of Citizen Science in Crowdsourced Collection of Speech Resources in Slovenian

This paper examines innovative approaches to collecting spontaneous speech resources for Slovenian, including citizen science, crowdsourcing, GWAP, and Collect4NLP. Based on an analysis of three European initiatives (CorCenCC, Anneta kõnet, Lahjoita puhetta), we highlight the importance of speaker motivation, legal and ethical safeguards, and the role of national promotional campaigns, and offer recommendations for the design of an online platform for remote speech data collection.

Keywords: speech resources acquisition, citizen science, crowdsourcing, spontaneous speech

GOVORJENI JEZIK MED RAZISKOVANJEM IN TEHNOLOGIJO:
ZBORNIK POVZETKOV
D. Verdonik, N. Ljubešić (ur.)

Univerzitetna založba
Univerze v Mariboru

Podporna orodja za obdelavo govornih posnetkov v jezikoslovnem raziskovanju

JANEZ KRIŽAJ, SIMON DOBRIŠEK
Univerza v Ljubljani, Fakulteta za elektrotehniko, Ljubljana, Slovenija
janez.krizaj@fe.uni-lj.si, simon.dobrisek@fe.uni-lj.si

Članek predstavlja pet odprtokodnih orodij za obdelavo govora, razvitih za podporo raziskavam govorjene slovenščine. Orodja pokrivajo različne vidike dela z govornimi jezikovnimi podatki in se medsebojno dopolnjujejo. Vsa orodja so javno dostopna, dokumentirana in namenjena široki skupnosti raziskovalcev govorjenega jezika ter razvijalcev jezikovnih tehnologij.

1 Uvod

Raziskovanje govorjenega jezika zahteva specializirana orodja za obdelavo zvočnih posnetkov. Na Fakulteti za elektrotehniko UL smo razvili pet odprtokodnih in prosto dostopnih orodij, ki pokrivajo vse ključne faze obdelave govornih podatkov

2 Vsiljena poravnava govora

Orodje (https://github.com/jan3zk/forced_alignment) temelji na Montreal Forced Aligner (MFA) (McAuliffe, 2017) in omogoča poravnavo govora s transkripcijo. Razširjeno je z jezikoslovnimi sloji, kot so zlogi, prozodija in menjave govorcev. Akustične meritve vključujejo trajanje, formante, jakost ipd. Prilagojeno je za slovenščino s slovarjem izgovorjav Optilex (Žganec Gros, 2022) in akustičnim

modelom, naučenim na zbirki ARTUR (Verdonik, 2023). Poravnava na korpusu Gos 2 (Verdonik, 2024) dosega točnost >90 % znotraj 100 ms (Križaj, 2024).

3 Anonimizacija zvočnih posnetkov

Orodje (https://github.com/jan3zk/audio_anonymizer) uporablja kombinacijo poravnave z MFA in prepoznavanja imenskih entitet prek knjižnice spaCy (Honnibal, 2020). Občutljive informacije, kot so imena oseb in krajev, se v posnetku samodejno nadomestijo z nevtralnim zvokom (1 kHz pisk). Možna je tudi ročna določitev ključnih besed. Orodje omogoča etično rabo posnetkov govora v raziskavah, skladno z zahtevami GDPR.

4 Validacija govornih posnetkov

Orodje (https://github.com/jan3zk/audio_validation) omogoča tehnično preverjanje (format, glasnost, tišina) in (pol)samodejno preverjanje ujemanja govora in transkripcije s samodejnim prevajalnikom. Uporabnik ima vizualni pregled valovne oblike in razlik v transkripciji. Orodje je bilo uspešno uporabljeno pri validaciji posnetkov za korpus ARTUR (Verdonik, 2023), kjer je avtomatsko zavrnilo približno 15 % posnetkov in pospešilo postopek gradnje korpusa (Križaj, 2022b).

5 Akustična normalizacija

Orodje (https://github.com/jan3zk/akusticna_normalizacija) vključuje tri metode: spektralno filtriranje NRSG (Sainburg, 2024), nevronski model DFL (Germain, 2018) in pristop z nasprotniškimi mrežami SEGAN (Pascual, 2017). DFL in SEGAN omogočata učinkovito odstranjevanje šuma brez večjih izgub prozodije. Merjeni kazalniki kažejo občutne izboljšave kakovosti posnetkov.

6 Grafemsko-fonemski pretvornik

Orodje (https://github.com/jan3zk/rsdo_gfp_v2) podpira postopke Sequitur (Bisani, 2008), Phonetisaurus (Novak, 2016) in Deep Phonemizer. Modeli so naučeni na leksikonu Gigafidaleks (Krek, 2020) in testirani na Sofesleks (Dobrišek, 2017). Deep Phonemizer (Yolchuyeva, 2019) izstopa po natančnosti (Križaj, 2022a). Pretvornik omogoča enostavno fonetizacijo nestandardnih zapisov.

Ključne besede: govorni posnetki, vsiljena poravnava, anonimizacija, validacija, akustična normalizacija

Zahvala

Razvoj predstavljenih orodij je bil podprt s strani Javne agencije za raziskovalno dejavnost Republike Slovenije v okviru raziskovalnega projekta Mezzanine (teMeljnE raZiskave Za rAzvoj govorNih vIrov in tehNologij za slovEnščino), šifra projekta: J7-4642.

Literatura

Bisani, M. in Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5), 434-451.

Dobrišek, S., Žganec Gros, J., Žibert, J., Mihelič, F. in Pavešić, N. (2017). Speech Database of Spoken Flight Information Enquiries SOFES 1.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1125

Germain, F.G., Chen, Q. in Koltun, V. (2018). Speech Denoising with Deep Feature Losses. *Proceedings of Interspeech*.

Honnibal, M., Montani, I., Van Landeghem, S. in Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., … Dobrovoljc, K. (2020, May). Gigafida 2.0: The Reference Corpus of Written Standard Slovene. V *Proceedings of the Twelfth Language Resources and Evaluation Conference* (str. 3340–3345). https://aclanthology.org/2020.lrec-1.409/

Križaj, J., Dobrišek, S., Mihelič, A. in Žganec-Gros, J. (2022a). Uporaba postopkov strojnega učenja pri samodejni slovenski grafemsko-fonemski pretvorbi. *Jezikovne tehnologije in digitalna humanistika*, 248-251.

Križaj, J., Žganec Gros, J. in Dobrišek, S (2022b). Validation of Speech Data for Training Automatic Speech Recognition Systems. *30th European Signal Processing Conference (EUSIPCO)*, str. 1165-1169, Beograd, Srbija, doi: 10.23919/EUSIPCO55093.2022.9909734.

Križaj, J., Žganec Gros, J. in Dobrišek, S. (2024). Utilizing Forced Alignment for Phonetic Analysis of Slovene Speech. *Proceedings of the Language Technologies and Digital Humanities conference*. https://www.sdjt.si/wp/wp-content/uploads/2024/09/JT-DH_2024_Krizaj_Gros_Dobrisek.pdf

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. in Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of Interspeech*, str. 498–502. https://doi.org/10.21437/Interspeech.2017-1386

Novak, J. R., Minematsu, N. in Hirose, K. (2016). Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering*, 22(6), 907-938.

Pascual, S., Bonafonte, A. in Serrà, J. (2017). SEGAN: Speech Enhancement Generative Adversarial Network. In F. Lacerda (Ed.), *18th Annual Conference of the International Speech Communication Association, Interspeech 2017*, Stockholm, Sweden, August 20-24, 2017 (str. 3642–3646). doi:10.21437/INTERSPEECH.2017-1428

Sainburg, T. in Zorea, A. (2024). Noisereduce: Domain General Noise Reduction for Time Series Signals. 10.48550/arXiv.2412.17851.

Verdonik, D., Bizjak, A. in Dobrišek, S. (2023). Description of the Artur speech database in the framework of the RSDO project. Elaborat, predštudija, študija. Univerza v Mariboru. https://dk.um.si/IzpisGradiva.php?lang=slv&id=85196

Verdonik D., Dobrovoljc K., Erjavec T. in Ljubešić N. (2024). Gos 2: A New Reference Corpus of Spoken Slovenian. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, str. 7825–7830, Torino, Italia. ELRA and ICCL.

Yolchuyeva, S., Németh, G. in Gyires-Tóth, B. (2019). Transformer Based Grapheme-to-Phoneme Conversion. 2095-2099. Doi: 10.21437/Interspeech.2019-1954.

Žganec Gros, J., Mirtič, T., Romih, M. in Ahačič, K. (2022). Slovar izgovarjav OptiLEX (1. e-izd.). Založba ZRC. https://doi.org/10.3986/9789610506720

Support Tools for Speech Processing in Linguistic Research

This paper presents five open-source tools for speech processing, developed to support research on spoken Slovenian language. The tools cover various aspects of working with speech language data and complement each other. All tools are publicly available, documented, and aimed at the broad community of spoken language researchers and language technology developers.

Keywords: speech processing, forced alignment, anonymization, validation, acoustic normalization

GOVORJENI JEZIK MED RAZISKOVANJEM IN TEHNOLOGIJO:
ZBORNIK POVZETKOV
D. Verdonik, N. Ljubešić (ur.)

Univerzitetna založba
Univerze v Mariboru

Uporabnost tehnik prenosa znanja pri razvoju modelov za prepoznavo jezika in govorca

MARKO BAJEC, IZTOK LEBAR BAJEC
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Ljubljana, Slovenija
marko.bajec@fri.uni-lj.si, ilb@fri.uni-lj.si

V prispevku kratko opišemo ključne ugotovitve raziskave o pomenu in vplivu tehnik prenosa znanja pri učenju modelov prepoznave govora ter segmentacije na govorce, in sicer za jezike, ki imajo malo jezikovnih virov.

1 Raziskava

Eden ključnih izzivov pri razvoju modelov prepoznave slovenskega govora je t.i. majhnost jezika. Slovenija ima z dvema milijonoma prebivalcev in relativno majhno razseljenostjo malo govorcev in piscev, ki bi ustvarjali tekstualne in zvočne zapise. Posledično je precej težje zbrati učne vire za učenje modelov, povezanih z govorom. V projektu RSDO[1] smo zbrali 1000 ur govorjene slovenščine (ter za 840 ur pripravili natančne prepise), kar ni zanemarljiva množica, vendar se izkaže v primerjavi z velikimi jeziki, kjer je na voljo po več sto tisoč učnih ur, izjemno majhna.

V projektu MEZZANINE[2] smo eno aktivnost namenili preučevanju tehnik prenosa znanja, saj smo želeli preveriti, ali si lahko kakorkoli pomagamo z obsežnimi učnimi viri in kakovostnimi modeli, ki so na voljo za nekatere druge jezike. Konkretno nas

[1] https://rsdo.slovenscina.eu/
[2] https://mezzanine.um.si/

je zanimalo: a) ali obstaja možnost prenosa znanja iz večjih modelov oz. večjih jezikov na slovenski jezik, b) kako vpliva velikost učne množice na robustnost modela ter splošno natančnost, ter c) ali za kakovostno diarizacijo govorcev (za slovenski govor) zadošča dober splošen večjezični model ali je potrebna slovenska učna množica.

2 Ključne ugotovitve

V tem razdelku so kratko opisane ključne ugotovitve raziskave. Podrobneje jih bomo predstavili na delavnici.

- Vpliv velikosti učne množice je nelinearen: poskušali smo z množicami od 50 ur do približno 15.000 prečiščenih ur. Uporabili smo oba pristopa, dvostopenjski, kot je KALDI (Povey et al, 2011), in e2e modele, kot jih podpirata NeMo[3] in Whisper[4]. Testirali smo na enostavnejših in zahtevnejših akustičnih okoliščinah. Že pri 1000 urah se e2e modeli večinoma izkažejo boljši za splošno prepoznavo. S povečevanjem učne množice WER (angl. Word Error Rate) pričakovano pada, vendar padec ni linearen, temveč upada. Od 10.000 naprej pri enostavnejših akustičnih okoliščinah bistvenih izboljšav WER nismo več zaznali, medtem ko se je za zahtevnejše okoliščine WER še vedno zmanjševal, saj je model postajal robustnejši.
- Prenos znanja iz drugih jezikov: s prenosom znanja iz drugih jezikov pridobimo zgolj to, da je za isti rezultat potrebnih manj epoh, ker model hitreje konvergira. Žal nimamo dostopa do kakovostnih slovanskih modelov (npr. hrvaščina, srbščina, češčina ...), zato prenosa znanja iz jezikov iste jezikovne družine nismo uspeli preskusiti. Predvidevamo, da bi imel prenos znanja za jezike iz iste družine večji vpliv, kot smo ga zaznali v naših raziskavah.
- Prenos znanja z učenjem večjezičnih modelov: ena od možnosti prenosa znanja je tudi učenje večjezičnih modelov, kjer si jeziki delijo vse nivoje nevronske mreže z izjemo zadnjega, ki dekodira končne signale v žetone (angl. tokens) združenega ali sestavljenega tokenizatorja (angl. tokenizer). V naših poskusih se je izkazalo, da pri skupnem učenju dejansko pride do prenosa znanja (bolje rečeno skupnega učenja), vendar morajo biti učne množice vsaj približno enako

[3] https://github.com/NVIDIA/NeMo
[4] https://github.com/openai/whisper

velike. V nasprotnem primeru močneje zastopani jeziki prevagajo manj zastopane, kar se pozna pri inferenci.

- Specializacija za novo domeno: pri starih pristopih, kot je KALDI, za specializacijo na novo domeno zadošča izdelava novega jezikovnega modela, ki je prilagojen novi domeni. Pri e2e modelih je vpliv jezikovnega modela, ne glede na parametre, manjši in posledično je potrebno specializirati akustični model. Poskuse smo izvedli na medicinski domeni, ki ima zelo specifično besedišče. Uporabili smo od 200 do 3000 dodatnih ur. Po 50 epohah se je model naučil zelo dobro (pri naši validacijski množici je WER znašal le pičlih 0,006), pri čemer »pozabljanja« nismo zaznali.
- Segmentacija na govorce: preskusili smo vse možnosti, ki so podprte v NeMo (Park et al, 2022) in Pyannote ogrodjih (Khoma, 2023). Pripravili smo lastno učno množico in izvedli specializacijo za MSDD (Park, 2022) in nevronski model. Izkazalo se je, da z dodatnim učenjem ne pridobimo veliko, pravzaprav skoraj nič. Pyannote 3.1 je že v osnovi relativno natančen, glede na druge modele, ki so na voljo, pri uporabi MSDD pa smo naleteli na težave z daljšimi posnetki (> 30 minut) in/ali več govorci (n>8). Pri slednjih kakovost segmentacije zelo pade, zato smo MSDD kot tak opustili pri nadaljnjem testiranju.

Ključne besede: prepoznava govora, segmentacija na govorce, prenos znanja

Literatura

Khoma, V., Khoma, Y., Brydinskyi, V., & Konovalov, A. (2023). Development of Supervised Speaker Diarization System Based on the PyAnnote Audio Processing Library. *Sensors*, *23*(4). https://doi.org/10.3390/s23042082

Park, T. J., Koluguri, N. R., Jia, F., Balam, J., & Ginsburg, B. (2022). NeMo Open Source Speaker Diarization System. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2022-September*.

Park, T. J., Koluguri, N. R., Balam, J., & Ginsburg, B. (2022). Multi-scale Speaker Diarization with Dynamic Scale Weighting. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2022-September*. https://doi.org/10.21437/Interspeech.2022-991

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. IEEE Signal Processing Society.

Winata, G. I., Cahyawijaya, S., Lin, Z., Liu, Z., Xu, P., & Fung, P. (2020). Meta-transfer learning for code-switched speech recognition. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.348

Zhou, R., Koshikawa, T., Ito A., Nose, T. and Chen, C. P. Multilingual Meta-Transfer Learning for Low-Resource Speech Recognition, in IEEE Access, vol. 12, pp. 158493-158504, 2024, doi: 10.1109/ACCESS.2024.3486711.

Efficiency of Knowledge Transfer Techniques in the Development of Speech and Speaker Recognition Models

In the paper, we briefly describe the key findings of a study on the significance and impact of knowledge transfer techniques in training models for speech recognition and speaker segmentation, particularly for low-resource languages.

Keywords: speech recognition, speaker diarization, knowledge transfer

GOVORJENI JEZIK MED RAZISKOVANJEM IN TEHNOLOGIJO:
ZBORNIK POVZETKOV
D. Verdonik, N. Ljubešić (ur.)

Univerzitetna založba
Univerze v Mariboru

Besedilo kot ogledalo narečne raznolikosti: gradivo, zapis, uporaba in izzivi

KLARA ŠUMENJAK,[1] JOŽICA ŠKOFIC[2]

[1] Univerza na Primorskem, Fakulteta za humanistične študije, Koper, Slovenija
klara.sumenjak@fhs.upr.si
[2] Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti, Inštitut za slovenski jezik Frana Ramovša, Ljubljana, Slovenija
jozica.skofic@zrc-sazu.si

V prispevku je obravnavano vprašanje ustreznosti zbiranja narečnih podatkov s pomočjo vnaprej pripravljenega (knjižnega) besedila in njihove (jezikoslovne) uporabnosti.

1 Narečjeslovje

Dialektologija oz. narečjeslovje je jezikoslovna veda, ki preučuje narečja tako na sinhroni kot diahroni ravni. Veda uporablja različne raziskovalne metode na različnih stopnjah raziskave, dogovorjeni znanstveni zapis, dokumentacijo z natančnimi metapodatki in jezikoslovno analizo, ki od raziskovalca zahteva poznavanje vseh ravnin jezikovnega sistema posameznih krajevnih govorov.

2 Narečno besedilo, nastalo po glasoslovno premišljeni knjižni predlogi

V prispevku je obravnavano vprašanje ustreznosti zbiranja narečnih podatkov s pomočjo vnaprej pripravljenega (knjižnega) besedila in njihove (jezikoslovne) uporabnosti. Besedilo, zapisano v slovenskem knjižnem jeziku, so informanti iz

različnih slovenskih krajev pretvorili v svoj domači, neformalni jezikovni kod (narečje, pogovorni jezik, mestno govorico …), ga prebrali in posneli ter s tem omogočili njihovo nadaljnjo jezikoslovno uporabo – fonetični zapis in jezikoslovno analizo. V besedilu so bile uporabljene besede iz gramatičnega dela vprašalnice za *Slovenski lingvistični atlas,* s katerimi je mogoče preverjati razvoj izbranih praslovanskih fonemov (npr. nosnikov, dolgega jata, dolgega cirkumflektiranega *o* …) v posnetih govorih. V razpravi je predstavljeno, kako je tovrstno gradivo uporabno za analizo narečnih glasoslovnih pojavov in njihovo (mednarečno) primerjavo. Izkušnje kažejo, da ima zbrano gradivo poleg raziskovalne predvsem didaktično vrednost, saj omogoča neposreden vpogled v fonetične posebnosti slovenskih narečij/izbranih krajevnih govorov in predstavlja dragocen pripomoček pri poučevanju narečne raznolikosti.

Med izzive raziskave med drugim uvrščamo vprašanja standardizacije zapisa in oblikovanja optimalnega nabora fonemov, ki omogoča učinkovito prepoznavanje narečne raznolikosti slovenskega jezikovnega prostora. Doslej zbranih približno 40 zvočnih posnetkov iz vseh narečnih skupin že izkazuje precejšnjo narečno raznolikost tako na glasoslovni kot tudi na leksikalni ravnini.

Ključne besede: dialektologija, narečna besedila, narečna transkripcija, narečno glasoslovje, narečno besedje

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta ARIS Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik (J7-4642).

Literatura

Benedik, F. (1999). *Vodnik po zbirki narečnega gradiva za Slovenski lingvistični atlas (SLA)*. Ljubljana: ZRC SAZU.

Rigler, J. (1963). Pregled osnovnih razvojnih etap v slovenskem vokalizmu. *Slavistična revija* 14/1–4. 25–78.

Text as a Mirror of Dialect Diversity: Material, Recording, Use and Challenges

The paper discusses the issue of the appropriateness of collecting dialect data using a pre-prepared (Slovene standard) text, and their (linguistic) utility.

Keywords: dialectology, dialect texts, dialect transcription, dialect phonology, dialect vocabulary

GOVORJENI JEZIK MED RAZISKOVANJEM IN TEHNOLOGIJO:
ZBORNIK POVZETKOV
D. Verdonik, N. Ljubešić (ur.)

Univerzitetna založba
Univerze v Mariboru

Izzivi pri standardizaciji narečne transkripcije samoglasnikov v prekmurskem in prleškem narečju

MELITA ZEMLJAK JONTES, MIHAELA KOLETNIK
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
melita.zemljak@um.si, mihaela.koletnik@um.si

Prispevek prikazuje izzive pri preverjanju zanesljivosti veljavne slovenske narečne transkripcije v panonskih prekmurskem in prleškem narečju. Po gradivu za SLA je pregledana in ovrednotena narečna fonetična transkripcija odrazov za issln. * ə̄ in issln. *ə̀- v prekmurskem narečju ter za issln. * ẽ̄ in za issln. *ē̄ v prleškem narečju zaradi variantnosti v zapisu odrazov ter kakovosti in trajanju glasov. Z izsledki instrumentalne analize narečnega gradiva se skuša pojasniti vzroke za razlike v kvaliteti in/ali kvantiteti glasov oz. vsaj utemeljiti njihov obstoj v obravnavanem narečnem govoru.

1 Cilji raziskave v okviru projektnega[1] delovnega sklopa Narečna variabilnost

Eden izmed ključnih ciljev sodobnih raziskav spontane govorjene rabe jezika je pregled stanja in opredelitev potreb po govornih podatkih ter pripadajoči raziskovalni infrastrukturi, pri čemer poseben raziskovalni izziv predstavlja tudi socialnozvrstna členjenost (slovenskega) jezika.

[1] Prispevek je nastal v okviru raziskovalnega projekta ARIS Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik (J7-4642), ki ga sofinancira Agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS) iz državnega proračuna.

V okviru projektnega delovnega sklopa Narečna variabilnost smo se predstavniki slovenskih dialektoloških središč ukvarjali z zanesljivostjo veljavne slovenske narečne transkripcije, določitvijo prostorske razširjenosti nestandardnih fonemov, izdelavo prostorskega modela za pripravo diasistemskih narečno-knjižnih kontrastivnih tabel fonemov in z opredelitvijo optimalnega nabora slovenskih fonemov, uravnoteženega med standardnimi in narečnimi različicami fonemov.

2 Pregled in ovrednotenje fonetičnih transkripcij gradiva za *SLA* za izbrane problematične izvornoslovenske samoglasnike v prekmurskem in prleškem narečju panonske narečne skupine

Na Filozofski fakulteti Univerze v Mariboru smo se osredinili na (pilotno) obravnavo panonske narečne skupine, natančneje prekmurskega in prleškega narečja, s ciljem preveriti primernost v projektu predvidenega postopka standardizacije slovenske narečne transkripcije samoglasnikov. Pregledali in ovrednotili smo fonetično transkripcijo gradiva za SLA v (1) prekmurskem narečju, in sicer za issln. **ə̄* zaradi eno- oz. dvoglasniškosti odraza zanj (*ẹ̄ː* – *eːi̯*/*ẹ̄ːi̯*) in različne fonetične vrednosti dvoglasnika (*eːi̯* – *ẹ̄ːi̯*) ter za issln.* *ə̀-* in **-ə̀* – zaradi nejasne kakovosti odraza zanju (široki *e* – zelo široki *a*) in problematike dvoglasniškosti (*ie*, *e̯a*, *i̯a*); (2) v prleškem narečju, in sicer za issln. **ẹ̄* zaradi štirih kakovostnih različic e-jev (*ẹ̄ː*, *e̝ː*, *e̯ːˇ* in *eːˇ*) in izkazane kolikostne opozicije (*ẹ̄ː* : *ẹ̄*) ter za issln. **ẹ̀* zaradi treh kakovostnih različic e-jev (*eːˇ*, *e̯ːˇ*, *ẹ̄ːˇ*). Gradivo za SLA pri evidentiranem besedju za oba glasova sicer izkazuje prleško enoglasniško naravo samoglasnikov.

Na izbranih primerih je bila opravljena instrumentalno-slušna analiza dostopnih zvočnih posnetkov (arhivsko zvočno gradivo, kolikor je ohranjeno v zvočnem arhivu Dialektološke sekcije ZRC SAZU in v zvočnem arhivu Filozofske fakultete Univerze v Mariboru), s čimer se je tako skušalo pojasniti vzroke za razlike v kvaliteti in/ali kvantiteti glasov oz. vsaj utemeljiti njihov obstoj v obravnavanih narečnih govorih (Koletnik, Zemljak Jontes, 2024; Koletnik, Zemljak Jontes, 2025).

Ključne besede: dialektologija, prekmursko in prleško narečje, gradivo za SLA, fonetični zapis, eksperimentalnofonetična analiza

Literatura

Koletnik, M., Zemljak Jontes, M. (2024). Standardizacija prekmurske transkripcije samoglasnikov: študija primera. V M. Krajnc Ivič (Ur.), *Stanje in perspektive uporabe govornih virov v raziskavah govora* (str. 121–150). Univerzitetna založba. https://doi.org/10.18690/um.ff.4.2024.7

Logar, T. (1966). Prispevek k dialektologiji Goričkega. V F. Zadravec (ur.), *Panonski zbornik* (str. 29–30). Pomurska založba.

Logar, T. (1996). *Dialektološke in jezikovnozgodovinske razprave.* ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša.

Ramovš, F. (1935). *Historična gramatika slovenskega jezika VII, Dialekti.* Učiteljska tiskarna.

Ramovš, F. (1936). *Kratka zgodovina slovenskega jezika.* Akademska založba.

Rigler, J. (1986). *Razprave o slovenskem jeziku.* Slovenska matica.

Zemljak Jontes, M., Koletnik, M. (2025). Instrumentalno-slušna določitev kvalitete odrazov za **ẽ̄* in **ē* v prleškem narečju. V J. Škofic, D. Zuljan Kumar in K. Kenda-Jež (Ur.), *1. DiaClas – Konferenca Narečna klasifikacija in 5. Slovenski dialektološki posvet: povzetki prispevkov* (str. 107–108). Slovenska akademija znanosti in umetnosti.

Zorko, Z. (2009). *Narečjeslovne razprave o koroških, štajerskih in panonskih govorih.* (Zora, 64). Filozofska fakulteta, Mednarodna založba Oddelka za slovanske jezike in književnosti.

Challenges in Standardising the Dialectal Transcripts of Vowels in the Prekmurje and Prlekija Dialects

The paper presents challenges in verifying the reliability of the valid Slovene dialect transcription in the Pannonian Prekmurje and Prlekija dialect. Based on the dialectal material for SLA, the dialect phonetic transcription of reflections for issln. **ə̄* and the issln. **ə̀-* in the Prekmurje dialect and issln. **ẽ̄* and issln. **ē* in the Prlekija dialect is reviewed and evaluated due to variations in the recording of reflections and the quality and duration of sounds. The results of the instrumental analysis of the dialectal material are used to explain the reasons for differences in the quality and/or quantity of sounds, or at least to justify their existence in the dialect speech under consideration.

Keywords: dialectology, Prekmurje and Prlekija dialect, material for SLA, phonetic notation, experimental phonetic analysis

GOVORJENI JEZIK MED RAZISKOVANJEM IN TEHNOLOGIJO:
ZBORNIK POVZETKOV
D. Verdonik, N. Ljubešić (ur.)

Univerzitetna založba
Univerze v Mariboru

Strojni razrez in fonetične meritve kot temelj zapisa izgovora v DSBS

NEJC ROBIDA
Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija
nejc.robida@ff.uni-lj.si

V Digitalni slovarski bazi za slovenščino (DSBS) izgovor zapisujemo v pisavah IPA in SAMPA. V okviru projekta Mezzanine smo pregledali normativne priročnike in pripravili izhodiščno gradivo za kvantitativno fonetično analizo posnetkov govora iz korpusov Gos 2.1 in Artur. Tudi na podlagi te analize bomo lahko oblikovali standardni nabor glasov za DSBS, ga podprli z referenčnim gradivom in pripravili jasna navodila za zapis izgovora.

1 Zapis izgovora v DSBS

Zapisovanje govora je zapleteno vprašanje, ki si ga redno zastavljajo jezikoslovci različnih disciplin. Raziskave govorjene slovenščine so bile doslej sporadične in omejene predvsem na discipline, kot sta fonetika s fonologijo in dialektologija, ki sta vedi, tipično usmerjeni k preučevanju govorjenega jezika, medtem ko je bila pred projektom Mezzanine na drugih znanstvenih področjih govorjena slovenščina le izjemoma predmet preučevanja (na primer Verdonik, 2007; Krajnc Ivič, 2009; Smolej, 2012).

V Digitalni slovarski bazi za slovenščino, ki vsebuje podatke o pomenskih, stilnih, skladenjskih, slovničnih, kolokacijskih, frazeoloških in drugih lastnostih dela besedišča sodobne slovenščine (Gantar, 2015), za zapis izgovora uporabljamo fonetični pisavi IPA in SAMPA, trenutni nabor glasov pa je bil osnovan na

Horjakovem (2016) predlogu fonetične transkripcije za slovenščino. V okviru projekta Mezzanine smo pridobili in pripravili podatke, na podlagi katerih se bomo lahko opredelili do različnih fonetičnih vprašanj, na katera jezikoslovje še ni podalo dokončnega odgovora (izgovor sklopa *ij*, izgovor fonema /v/, vprašanje fonetičnih dvoglasnikov, obravnava naglašenih samoglasnikov brez kolikostne razlike itn.).

2 Izbor gradiva in strojni razrez posnetkov

Fonetične analize govora vzamejo zelo veliko časa, vsaj tako je veljalo do zdaj. Raziskovalec je moral ročno segmentirati posnetke na posamezne glasove in s pomočjo računalniških programov tudi analizirati spektrogram posameznega glasu. Novejši modeli, kot je Montreal Forced Aligner (MFA; McAuliffe idr., 2017), nam omogočajo, da posnetke govora na posamezne glasove s pomočjo slovarja izgovarjav in akustičnega modela razrežemo strojno. Na takšnem gradivu lahko potem opravimo tudi strojne fonetične meritve. To zelo skrajša čas zbiranja podatkov in nam omogoči širok nabor posnetkov, govorcev in govorjenih besedil.[1]

V okviru projekta smo za razrez izbrali 131 govorcev in govork iz korpusov Gos 2.1 (Verdonik, 2023) in Artur 1.0 (Verdonik, 2023), ki so govorili pretežno standardni jezik, pri izboru pa smo upoštevali tudi, da smo imeli vsaj enega ali več govorcev iz vsake statistične regije, da bomo lahko ob zaključkih analize izključili vpliv narečne osnove na standardni izgovor govorca oz. govorke. Za poskusne razreze smo uporabili različne nabore glasov. Prvi je bil najpodrobnejši in je bil osnovan na *Slovenski slovnici* (2004) in obsega 61 grafemov v pisavi SAMPA (npr. [w], [W], [U], [d_n], [d_l], [n'], [N] itd.).[2] Drugi nabor, ki smo ga uporabili pri razrezu, vsebuje tudi dvoglasnike (npr. [aI] in [OU]), to nam bo v bodoče omogočilo tudi natančnejše meritve t. i. fonetičnih dvoglasnikov. Tretji nabor pa je bil poenostavljen, gre predvsem za fonemski zapis, obsega 44 grafemov in uporablja na primer enotni grafem [w] za vse tri dvoustnične variante fonema /v/.

3 Slovar izgovarjav in orodje MFA

Za uspešno vsiljeno poravnavo s pomočjo orodja MFA potrebujemo tudi slovar izgovarjav. Odločili smo se, da bomo uporabili slovar OptiLEX (Žganec Gros idr., 2022). Ta vsebuje 698.375 naglašenih oblik. Slovarju pa smo dodali še izgovore 835

[1] Več o tem v Robida idr. (2024).
[2] V članku uporabljamo za zapis izgovora fonetično abecedo SAMPA (Zemljak Jontes idr., 2002).

nestandardnih oblik, ki se najpogosteje pojavljajo kot del pogovornega zapisa v korpusu Gos 2.1. Tako bomo lahko v bodoče ustrezno analizirali tudi izgovor nestandardnega besedišča, kot so na primer besede *dej*, *toti* in *jəz*, ki ga v okviru projekta Mezzanine tudi vključujemo v DSBS. Po prvih preizkusih smo ugotovili, da moramo slovar OptiLEX poenotiti, dopolniti in tudi popraviti. Večja težava so bili predvsem pomešan zapis dvoglasniške variante in ustnično-ustničnih šumnih zvenečih in nezvenečih /v/ (beseda *cvreti* je imela zapisan izgovor [tsur"e:ti]), neenoten zapis izgovora drsnika *j* v besedah, kot sta na primer *radio* in *aksialen*, nestandardni zapis polglasnika v mestniku ednine pri moškem spolu (npr. [n"EizgOvOrj"E:n@m]) ipd.

V slovarju, ki vsebuje različne variante izgovora posamezne besede, na primer izgovor sklopa *ij,* smo določili dva možna izgovora, in sicer kot [param"e:tsij] in [param"e:tsi], tako bomo lahko preverili, ali model MFA razliko med izgovorom sklopa *ij* zazanava kot en glas ali dva različna glasova. Model na primer pri razrezu s prvim najobširnejšim naborom ni zaznaval favkalnih in obstranskih variant fonemov /t/ in /d/, tudi zaradi tega bi veljalo razmisliti o fonemskem zapisu ali zapisu z manj variantami posameznih fonemov, a moramo prej na našem gradivu opraviti še natančnejšo fonetično raziskavo glasov, kot so [n'], [l'], [I], [F], [w], [W] in [U].

4 Standardni nabor glasov v DSBS

Ker ima izbor glasov in grafemov, ki jih opisujejo, pomembne posledice za konsistentnost opisa govorjene slovenščine, zahteva premišljen in strokovno utemeljen pristop. A sam nabor glasov pri zapisovanju izgovora ni edini problem. V okviru projekta smo se morali posvetiti tudi vprašanjem zapisa izgovora dvojnih soglasnikov (npr. *oddati*), asimilacij (*izhod*) in posebnosti, kot sta na primer besedi *uvleči* in *prečul*.

DSBS bo javno dostopen, zato morajo biti predlagani nabor standardnih glasov in navodila zapisa izgovora dovolj preprosti za vse uporabnike, hkrati pa morajo ohranjati fleksibilnost za vključevanje novih fonetičnih in jezikovnotehnoloških spoznanj. Ključna bo tudi usklajenost z novim *Pravopisom 8.0*, a bomo, če bo koncept DSBS to zahteval, na podlagi analize našega gradiva sprejeli tudi drugačno rešitev.

Ključne besede: Digitalna slovarska baza za slovenščino, vsiljena poravnava, fonetična analiza

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta ARIS Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik (J7-4642).
Za učenje modelov za vsiljeno poravnavo in strojni razrez posnetkov se iskreno zahvaljujem dr. Janezu Križaju in dr. Simonu Dobrišku s Fakultete za elektrotehniko Univerze v Ljubljani.

Literatura

Gantar, P. (2015). *Leksikografski opis slovenščine v digitalnem okolju*. Znanstvena založba Filozofske fakultete. http://www.ff.uni-lj.si/Portals/0/Dokumenti/ZnanstvenaZalozba/e-knjige/Leksikografski.pdf.

Horjak, L. (2016). *Problematika slovenske fonetične transkripcije: predlog nove različice mednarodne fonetične transkripcije za slovenščino: diplomsko delo*. [L. Horjak].

Krajnc Ivič, M. (2009). *Razgovor kot vrsta komunikacijskega stika*. Filozofska fakulteta, Mednarodna založba Oddelka za slovanske jezike in književnosti.

Robida, N., Čibej, J., in Krek, S. (2024). Strojni razrez posnetkov iz korpusa govorjene slovenščine GOS 2.1 in fonetične meritve. V *Predpis in norma v jeziku* (str. 267–274). Založba Univerze. https://centerslo.si/wp-content/uploads/2024/11/Robida-et-al._Obdobja-43.pdf.

Smolej, M. (2012). *Besedilne vrste v spontanem govoru*. Znanstvena založba Filozofske fakultete.

Toporišič, J. (2004). *Slovenska slovnica*. Obzorja.

Verdonik, D. (2007). *Jezikovni elementi spontanosti v pogovoru: diskurzni označevalci in popravljanja*. Slavistično društvo.

Zemljak Jontes, M., Kačič, Z., Dobrišek, S., Žganec Gros, J., in Weiss, P. (2002). Računalniški simbolni fonetični zapis slovenskega govora. *Slavistična revija*, 50(2), 159–169. https://srl.si/ojs/srl/article/view/COBISS_ID-19009634/PDF-2002-2-1-1.

Žganec Gros, J., Mirtič, T., Romih, M., in Ahačič, K. (2022). *Slovar izgovarjav OptiLEX*. Založba ZRC. https://doi.org/10.3986/9789610506720.

Forced Alignment and Phonetic Measurements as the Basis for Speech Transcription in DDDS

In the Digital Dictionary Database of Slovene (DDDS), pronunciation is recorded using IPA and SAMPA notation. As part of the Mezzanine project, we reviewed normative reference sources and prepared an initial dataset for quantitative phonetic analysis of speech recordings from the Gos 2.1 and Artur corpora. This analysis will also serve as a basis for establishing a standard phoneme inventory in the DDDS, supporting it with reference materials, and preparing clear guidelines for phonetic transcription.

Keywords: Digital Dictionary Database of Slovene, forced alignment, phonetic analysis

GOVORJENI JEZIK MED RAZISKOVANJEM IN TEHNOLOGIJO:
ZBORNIK POVZETKOV
D. Verdonik, N. Ljubešić (ur.)

Univerzitetna založba
Univerze v Mariboru

Tipično govorjeni leksemi in digitalna slovarska baza za slovenščino

JAKA ČIBEJ[1,2]
[1] Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija
jaka.cibej@ff.uni-lj.si
[2] Univerza v Ljubljani, Center za jezikovne vire in tehnologije, Ljubljana, Slovenija
jaka.cibej@ff.uni-lj.si

V prispevku opisujemo postopek luščenja in pregleda kandidatov za lekseme za Digitalno slovarsko bazo za slovenščino. Kandidate smo izluščili iz korpusa GOS, podatkovne zbirke Artur in transkripcij Online Notes. Opišemo postopek pregleda kandidatov in strojnega generiranja njihovih pregibnih ter naglašenih oblik in izgovorov s pomočjo Pregibalnika, orodja za strojno širjenje leksikona.

1 Digitalna slovarska baza za slovenščino: trenutno stanje

Digitalna slovarska baza za slovenščino (DSBS; Kosem in sod., 2021) je zasnovana kot osrednja podatkovna baza s strojno berljivimi podatki o slovenščini. V prispevku se osredotočamo na njen oblikoslovni del – *Slovenski oblikoslovni leksikon Sloleks* (Čibej in sod., 2022) –, ki trenutno vsebuje približno 380.000 leksemov ter njihovih pregibnih in naglašenih oblik skupaj z izgovori v mednarodni fonetični abecedi IPA ter njenem ekvivalentu X-SAMPA. Najnovejša širitev DSBS v okviru projekta RSDO je gradivo črpala iz korpusa pisne standardne slovenščine *Gigafida 2.0* (Krek in sod., 2021), pred tem pa razen posameznih ročno dodanih potencialno

nestandardnih leksemov (npr. *žleht*, *pošlihtan*, *zaštekati*) DSBS še nikoli ni bil sistematično razširjen s tipično govorjenim in nestandardnim besediščem.

2 Širjenje DSBS z novimi leksemi

Širjenja DSBS smo se lotili v okviru projekta MEZZANINE, zato v prispevku predstavljamo postopek širitve baze z besediščem, ki smo ga strojno izluščili iz več jezikovnih virov, ki vsebujejo transkripcije govorjene slovenščine – npr. korpus govorjene slovenščine *GOS* (Verdonik in sod., 2023a), zbirka za razpoznavo govora Artur 1.0 (Verdonik in sod., 2023b) ter transkripcije, ki so bile uporabljene za namene razvoja sistema za sprotno strojno prevajanje slovenskih predavanj Online Notes (Bajec in sod., 2023). V prispevku opišemo postopek luščenja in pregleda kandidatov za lekseme: strojno označenim leksemom smo ročno popravili oblikoskladenjske oznake in leme ter jim pripisali še nekatere druge značilnosti, ki bodo uporabne pri njihovi nadaljnji leksikografski obravnavi v DSBS. Za vsak leksem smo nato strojno generirali še njegove pregibne oblike, naglašene oblike in izgovore s pomočjo *Pregibalnika*,[1] orodja za strojno širjenje oblikoslovnega leksikona. Opišemo nekatere izzive, s katerimi smo se soočali pri vključevanju tipično govorjenega besedišča v podatkovno bazo, ter rešitve, ki smo jih pri tem implementirali. Postopek bo mogoče ponoviti tudi pri luščenju besedišča iz prihodnjih različic korpusov govorjene slovenščine.

Ključne besede: tipično govorjeni leksemi, govorjena slovenščina, Digitalna slovarska baza za slovenščino

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta *Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik* (MEZZANINE, J7-4642) in raziskovalnega programa *Jezikovni viri in tehnologije za slovenski jezik* (P6-0411), ki ju financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS).

Literatura

Bajec, M., Lebar Bajec, I., Šoltes, T., Cvek, J., Čibej, J., Gantar, K., Sever, S., & Krek, S. (2023). *Online Notes – a real-time speech recognition and machine translation system for Slovene university lectures*. (str. 7–10). https://is.ijs.si/wp-content/uploads/2023/11/IS2023_Volume-H.pdf

[1] *Pregibalnik* je na voljo v obliki odprto dostopne kode (https://github.com/clarinsi/SloInflector) ali preko aplikacijskega programskega vmesnika (https://orodja.cjvt.si/pregibalnik/docs).

Čibej, J., Gantar, K., Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Krsnik, L., Robnik-Šikonja, M. (2022), *Morphological lexicon Sloleks 3.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1745.

Kosem, I., Krek, S., Gantar, P. (2021) Semantic data should no longer exist in isolation: the digital dictionary database of Slovenian. V: Gavriilidou, Z., Mitits, L., Kiosses, S. (ur.): *Proceedings of the XIX EURALEX International Congress: Lexicography for Inclusion* (str. 81–83). Komotini: SynMorPhoSe Lab, Democritus University of Thrace. https://elex.is/wp-content/uploads/2021/09/Semantic-Data-should-no-longer-exist-in-isolation-the-Digital-Dictionary-Database-of-Slovenian_Kosem-Krek-Gantar_EURALEX2020.pdf

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I., Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. V: Calzolari, N. (ur.): *LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11-16, 2020, Marseille, France*. Paris: ELRA - European Language Resources Association. 2020, (str. 3340–3345). http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf

Verdonik, D., Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., Erjavec, T., Potočnik, T., Sepesy Maučec, M., Majhenič, S., Žgank, A., Bizjak, A., Gril, L., Dobrišek, S., Križaj, J., Bajec, M., Lebar Bajec, I., Jelovšek, T., Trojar, M., Bernjak, M., Dretnik, N., Strle, G., Dobrovoljc, K., Ljubešić, N., Rupnik, P. (2023a). *Spoken corpus Gos 2.1 (transcriptions)*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1863.

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., Zelenik, M., Ivanovska, M., Grm, K., Longyka, J., Mihelič, A., Vesnicer, B., Dretnik, N. (2023b). *ASR database ARTUR 1.0 (transcriptions)*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1772.

Typically Spoken Lexemes and the *Digital Dictionary Database of Slovene*

We present the results of extracting and analyzing lexeme candidates for the *Digital Dictionary Database of Slovene*. The candidates were extracted from the *GOS Corpus of Spoken Slovene*, the *Artur* database, and transcriptions used in the *Online Notes* project. We describe the analysis of the candidates and the automatic generation of their inflected and accentuated forms as well as their pronunciations using *Pregibalnik*, a custom tool for Slovene lexicon expansion.

Keywords: typically spoken lexemes, spoken Slovene, Digital Dictonary Database of Slovene

GOVORJENI JEZIK MED RAZISKOVANJEM IN TEHNOLOGIJO:
ZBORNIK POVZETKOV
D. Verdonik, N. Ljubešić (ur.)

Univerzitetna založba
Univerze v Mariboru

Govorjena slovenščina: k teoretski klasifikaciji žanrov vsakdanje komunikacije

MIRA KRAJNC IVIČ
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
mira.krajnc@um.si

Prispevek obravnava izzive sistematizacije govorjenih konverzacijskih besedil, ki podobno kot pisni žanri izkazujejo skupne jezikovne in/ali nejezikovne prvine. Njihova sistematizacija je otežena zaradi spontanosti, funkcijske odprtosti in terminološke nedoločnosti. Kljub temu se kot potencialna merila kažejo komunikacijsko področje, stopnja institucionalnosti, funkcija oziroma govorečev cilj ter tematika.

1 Izzivi klasifikacije konverzacijskih vrst

Pri raziskovanju besedilnih vrst (žanrov) je ključno tudi vprašanje njihove klasifikacije in tipologizacije. Adamzik (2018, str. 5) ugotavlja, da številni pristopi niso sistematični, saj operirajo z različnimi kategorijami in merili, npr. telefonski pogovor ali podkast ne poimenujeta žanra, temveč besedilni/konverzacijski tip. Ta spoznanja so prenosljiva na raziskovanje in klasifikacijo konverzacijskih vrst. Ključna vprašanja so: kaj je vsakdanja komunikacija, na katerih komunikacijskih področjih poteka, kaj je značilno za konverzacijsko vrsto in kako jih razvrstiti.

2 Potencialna merila razvrščanja konverzacijskih vrst

Vsakdanja komunikacija je zlasti govorjena, dialoška oz. interaktivna, spontana in ima v primerjevi s pisnimi besedili bistveno manjšo veljavnost trajanja (Adamzik, 2018, str. 17–18). Zajema konverzacijske prakse (Heinemann, 2000): a) institucionalno regulirane, b) razregulirane, c) socialno indiferentne ali d) nerelevantne. Gre za tipične jezikovne in nejezikovne dejavnosti na različnih komunikacijskih področjih, ki so s temi dejavnostmi oziroma konverzacijskimi vrstami vzratno konstituirana. To pomeni, da prvine konverzacijske vrste določa komunikacijsko področje in obratno. Funkcije vsakdanje konverzacije so vezane na področje zadovoljevanja osnovnih življenjskih potreb, na področje družbeno-kulturnih okoliščin in na delo. Funkcijsko so vsakdanje konverzacijske vrste odprte, saj udeleženci lahko zasledujejo več različnih ciljev (Heinemann 2000, str. 609), pri čemer lahko imata udeleženca istega dogodka različno dojemanje funkcije. To dejstvo poleg a) močne vpetosti v neposredni kontekst, b) nejasnih in nenatančnih laičnih poimenovanj konverzacijskih vrst, c) pogostih implikacij in poleg č) težke določljivosti za večino sprejemljivega, a individualnega vsakdana, omogoča le grobo klasifikacijo konverzacijskih vrst (Heinemann 2000, str. 605, 609).

Pregled v korpus ROG vključenih konverzacijskih skupin kaže omejen doseg pri zbiranju gradiva. Zbrani so večinoma javni, v manjši meri pa tudi zasebni, neformalni pogovori. Priporočena je vpeljava komunikacijskih področij, razlikovanje med konverzacijskimi skupinami, ki imajo skupne le jezikovne prvine, t. i. različnimi konverzacijskimi tipi, in razlikovanje med konverzacijskimi vrstami s skupnimi jezikovnimi in nejezikovnimi prvinami. Smiselna je tudi določitev stopnje institucionalnosti, funkcije oziroma govorčevega cilja in tematike.

Ključne besede: govorjeni jezik, komunikacijsko področje, konverzacijska vrsta, klasifikacija in tipologija konverzacijskih vrst

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta ARIS Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik (J7-4642).

Literatura

Adamzik, K. (2019). Textsorten und ihre Beschreibung. V N. Janich (Ur.) *Textlinguistik. 15 Einführungen und eine Diskussion.* Tübingen: Narr. Str. 135–168. (Narr-Studienbücher). Pridobljeno 18. 6. 2025, https://archive-ouverte.unige.ch/unige:113921

Bizjak, A. (2024). Korpusne oznake za opis konteksta govornih dogodkov v slovenskih govornih korpusih. *Slovenščina 2.0*, 12(1), 54–94.

Brinker, K., Antos, G., Heinemann, W., Sager, S. F (2000). Vorword. V K. Brinker, G. Antos, W. Heinemann, S. F. Sager (Ur.), *Text- und Gesprächslinguistik: ein internationales Handbuch zeitgenössischer Forschung; Linguistics of text and conversation: an international handbook of contemporary research; Linguistics of text and conversation* (str. XVII–XXIII). W. de Gruyter.

Firbas, J. (1992). Functional Sentence Perspective in Written and Spoken Communication. Cambridge University Press.

Gansel, Ch., in Jürgens, F. (2007). *Textlinguistik und Textgrammatik (eine Einführung).* Vandenhoeck & Ruprecht.

Heinemann, M. (2000): Textsorten des Alltags. V K. Brinker, G. Antos, W. Heinemann, S. F. Sager (Ur.), *Text- und Gesprächslinguistik: ein internationales Handbuch zeitgenössischer Forschung; Linguistics of text and conversation: an international handbook of contemporary research; Linguistics of text and conversation* (str. 604–614). W. de Gruyter.

Krajnc Ivič, M. (2020). Obravnava besedil. *Slavistična revija*, 68(1), 55–71.

Mikolič, V. (2007). Modifikacija podstave in argumentacijska struktura besedilnih vrst. *Slavistična revija*, 55(1-2), 345–355.

Verdonik, D., Ljubešić, N., Rupnik, P., Dobrovoljc, K., Čibej, J. (2024). Izbor in urejanje gradiv za učni korpus govorjene slovenščine – ROG. Pridobljeno 12. 2. 2025, https://www.sdjt.si/wp/wp-content/uploads/2024/09/JT-DH-2024_Verdonik_Ljubesic_Rupnik_Dobrovoljc_Cibej.pdf

Skubic, A. (2005). *Obrazi jezika.* Študentska založba.

Starc, S. (2020). Vrednotenje v reklami in antireklami: primer alkoholnih in tobačnih izdelkov. V J. Vogel (Ur.), *Slovenščina – diskurzi, zvrsti in jeziki med identiteto in funkcijo.* Obdobja 39. (str. 67–78). Znanstvena založba Filozofske fakultete.

Theorizing Spoken Slovene: A Genre-Based Classification of Everyday Conversation

The paper addresses the challenges of systematising spoken conversational texts, which, similar to written genres, display shared linguistic and/or non-linguistic features. Their classification, however, is complicated by spontaneity, functional openness, and terminological indeterminacy. Nevertheless, communication domain, degree of institutionalisation, function or speaker's intention, and topic emerge as potential criteria for their classification.

Keywords: spoken language, communicative domain, conversational genre, classification and typologization of conversational genres

GOVORJENI JEZIK MED RAZISKOVANJEM IN TEHNOLOGIJO:
ZBORNIK POVZETKOV
D. Verdonik, N. Ljubešić (ur.)

Univerzitetna založba
Univerze v Mariboru

Raziskave (ne)tekočnosti v projektu Mezzanine

DARINKA VERDONIK
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor, Slovenija
darinka.verdonik@um.si

Prosti govor je kognitivno zahteven proces, v katerem se odražajo številne (ne)tekočnosti, kot so premori, ponavljanja in samopopravljanja. V učnem korpusu ROG-Artur smo razvili večnivojsko shemo za njihovo označevanje ter analizirali razlike med javnim in zasebnim govorom.

1 (NE)tekočnosti

Prosti govor je kompleksen kognitivni proces, ki vključuje hkrati načrtovanje, ubesedovanje in artikulacijo. Rezultat ni linearno besedilo, členjeno po vnaprej postavljenih skladenjskih pravilih, ampak besedilo, skozi katero se odslikavajo govorčevi procesi tvorjenja in njegovo prilagajanje bodisi sogovorniku bodisi nepredvidljivim okoliščinam govorjenja. Besedilo tako vsebuje opuščene enote govora, popravljene enote in različne signale upočasnjenega tempa tvorjenja govora, kot so ponavljanja, premori, podaljševanja. Za vse te pojave se je uveljavilo skupno poimenovanje disfluencies (Kosmala, 2024), kar prevedemo kot netekočnosti. Poimenovanje ni najbolj posrečeno (Taylor, 1997; Clark, 2002; Crible, 2018; Kosmala, 2024), saj je »tekoč govor« ideal, ki v jezikovni rabi niti ne obstaja (razen morda, če je bran ali naučen vnaprej) niti ne bi bil najbolj učinkovita oblika govorjenja, navedeni pojavi pa govor ne samo razmejujejo, ampak tudi povezujejo, zato predlagamo uporabo predpone ne v oklepajih: (ne)tekočnosti.

2 Označevanje v učnem korpusu ROG-Artur

(Ne)tekočnosti v slovenskem govoru smo raziskovali v korpusu ROG-Artur (Verdonik et al., 2024). Ta vsebuje 57 posnetkov 72 različnih govorcev v skupnem obsegu nekaj več kot 5 ur govora. Približno 40 % govora je iz zasebnih pogovorov ali pripovedi, približno 50% iz različnih medijskih vsebin in 10% iz sej slovenskega državnega zbora. Na podlagi natančne analize tujih shem za označevanje (ne)tekočnosti smo izdelali prilagojeno shemo (Verdonik, 2024), ki omogoča označevanje teh pojavov na več ravneh: (1) glasovne netekočnosti (tihi in zapolnjeni premori, podaljševanja, blokade in nejezikovni zvoki), (2) besedne netekočnosti (ponavljanja, različne vrste samopopravljanj, nepopravljene netekočnosti in opuščene strukture), (3) komentarji netekočnosti, (4) struktura netekočnosti za tiste tipe, v katerih lahko ločimo popravljeno in popravek (in opcijsko še prehod).

3 (Ne)tekočnosti v javnem in zasebnem govoru

Na podalgi označenega gradiva smo raziskovali, ali obstajajo razlike med javnimi in nejavnimi govornimi situacijami, ter ugotavljali, da so razlike izrazite: nasploh so netekočnosti pogostejše v nejavnem govoru, a to ne velja za vse različne tipe, izjema so namreč zapolnjeni premori, nepopravljene izgovorjave in blokade, ki se pojavljajo pogosteje v javnem govoru. Razlike smo pojasnjevali na podlagi kontekstnih faktorjev (van Dijk, 1997): govorci v javnem govoru reducirajo netekočnosti zaradi visoke pomembnosti govora, formalnih pričakovanj, delne vnaprejšnje priprave, časovnih omejitev in dobrih govorniških veščin, medtem ko večja pogostost zapolnjenih premorov, nepopravljenih izgovorjav in blokad v javnem govoru odraža daljše vloge govorcev, časovne omejitve in čustvene obremenitve.

Ključne besede: tekočnost, diskurz, kontekst

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta ARIS Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik (J7-4642).

Literatura

Crible, L. (2018). *Discourse Markers and (Dis)fluency: Forms and functions across languages and registers*. John Benjamins Publishing Company. https://doi.org/10.1075/pbns.286.

Kosmala, L. (2024). *Beyond Disfluency: The interplay of speech, gesture, and interaction*. John Benjamins. https://doi.org/10.1075/ais.11?locatt=mode:legacy.

Taylor, T. J. (1997). *Theorizing language: analyses, normativity, rhetoric, history*. Pergamon.

van Dijk, T. A. (1997). Cognitive Context Models and Discourse. V M. I. Stamenov (Ur.), *Advances in Consciousness Research* (str. 189-226). John Benjamins Publishing Company. https://doi.org/10.1075/aicr.12.09dij?locatt=mode:legacy.

Verdonik, D. (2024). *Označevanje netekočnosti v govoru: primer označevanja z uporabo orodja Exmaralda*. Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko. https://dk.um.si/IzpisGradiva.php?id=87952.

Verdonik, D., Ljubešić, N., Rupnik, P., Dobrovoljc, K., in Čibej, J. (2024). Izbor in urejanje gradiv za učni korpus govorjene slovenščine ROG. V Š. Arhar Holdt, T. Erjavec (Ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (JT&DH 2024) (str. 469-484). Inštitut za novejšo zgodovino. https://doi.org/10.5281/zenodo.13936426

Research on (Dis)Fluency in the Mezzanine Project

Spontaneous speech is a cognitively demanding process that manifests various (dis)fluencies, such as pauses, repetitions, and self-repairs. Using the training corpus ROG-Artur, we developed a multi-level annotation scheme for annotating these phenomena and analyzed the differences in (dis)fluency use between public and private speech.

Keywords: fluency, discourse, context

GOVORJENI JEZIK MED RAZISKOVANJEM IN TEHNOLOGIJO:
ZBORNIK POVZETKOV
D. Verdonik, N. Ljubešić (ur.)

Univerzitetna založba
Univerze v Mariboru

Building a Filled Pause Detector for Slovenian and Evaluating Its Applicability to Various Slavic Languages

PETER RUPNIK, IVAN PORUPSKI, NIKOLA LJUBEŠIĆ
Jožef Stefan Institute, Ljubljana, Slovenia
peter.medle.rupnik@ijs.si, ivan.porupski@ijs.si, nikola.ljubesic@ijs.si

In spontaneous speech, filled pauses (such as "erm", or "umm") are very frequent, but during transcribing they are mostly skipped over. Consequently, speech disfluency research involves large amounts of manual annotation, which is time consuming and potentially error prone. We leveraged the Slovenian ROG dataset to fine-tune a transformer model that would identify filled pauses in speech in Slovenian, but potentially also other related Slavic languages. We discovered the resulting model to be highly accurate for Slovenian, and expanded our analysis to Croatian, Serbian, Czech, and Polish speech as well. We found the model generalises well, but with some performance drop in less closely related languages. In addition, comparing human- and machine-annotated filled pauses, we found that on some metrics (namely recall and F1 score) the model might actually be outperforming human annotators, but cannot beat them on precision. The human annotated test sets in four South- and West-Slavic languages as well as the fine-tuned model are freely available for applications in disfluency research and other speech related fields.

1 Data

The study relies on in-language and cross-lingual datasets to train and evaluate a speech transformer model for identifying filled pauses from the speech signal. The primary training and in-language test data are sourced from the ROG dataset (Verdonik et al., 2024), which contains Slovenian speech annotated with various disfluencies, including filled pauses. For training, the data were segmented into overlapping 30-second audio chunks to maximize usable content. The training set included 1,314 filled pauses, while the test set featured 558.

To assess the model's cross-lingual generalization, we constructed datasets in Croatian, Serbian, Czech, and Polish using the ParlaSpeech corpus (Ljubešić et al., 2024). For each language, 400 instances were selected (speech segments of 6–20 seconds), ensuring gender balance and an equal split between segments likely containing filled pauses and those without, based on predictions from the Slovenian model.

Manual annotations were conducted with minimalist guidelines: annotators marked "schwa"-like sounds indicative of filled pauses, carefully excluding partial words resulting from ASR-based segmentation. Croatian and Serbian datasets had a second annotator for reliability estimation, revealing strong inter-annotator agreement (F1 $\approx$ 0.89–0.93, Krippendorff's $\alpha \approx$ 0.79–0.81), setting a high-quality standard for further evaluation. Overall, the preparation ensured that both training and evaluation data are robust, diverse, and well-annotated, establishing a solid foundation for cross-lingual model assessment.

2 Experiments

The model used in the study is a Wav2Vec2Bert transformer configured for Audio Frame Classification (Barrault et al., 2023). The task is framed as a binary classification at the 20ms frame level, where each frame is labeled for the presence or absence of a filled pause. Labels for the training data were constructed accordingly.

Hyperparameter tuning was conducted on provisional splits of the training data, evaluating learning rates (3×10^{-5}, 1×10^{-6}, 8×10^{-6}), number of epochs (10, 20), and gradient accumulation steps (1, 4). The final configuration used a learning rate of 3×10^{-5}, 20 training epochs, and gradient accumulation steps of 4.

For evaluation, the frame-based binary outputs of the model were transformed into event-based spans, identifying the start and end times of predicted filled pauses. This span-based evaluation better matches human annotation practices and is more informative for real-world applications. Evaluation metrics included precision, recall, and F1 score based on overlaps between predicted and annotated spans.

Additionally, a post-processing step was applied to address known segmentation errors from the ParlaSpeech corpus. This included removing predictions at the start or end of segments (likely to be incomplete words) and discarding very short filled pauses (<80ms), which are generally imperceptible.

This methodology allowed us to evaluate the model's performance both within the training language (Slovenian) and in cross-lingual contexts, testing its robustness and generalization across related Slavic languages. The setup is noteworthy for its alignment with practical annotation realities and its emphasis on event-level metrics over lower-level frame alignment.

3 Results

The evaluation results demonstrate strong performance of the model, particularly in-language (Slovenian), where it achieves an F1 score of 0.94 with post-processing. Cross-lingual performance is slightly reduced, with F1 scores ranging from 0.87 (Czech) to 0.94 (Serbian), showing the model's strong generalization capabilities across South and West Slavic languages. Post-processing consistently improves precision across languages by reducing false positives from boundary noise and short fragments.

Comparatively, the Slovenian results are on par with or exceed those reported for English in previous studies (e.g., Romana et al., 2023, Bayerl et al., 2022), despite the novelty of such work in Slavic languages. Cross-lingual drops in performance are acceptable and still above the level of random or transcript-based detection, with

machine performance occasionally surpassing inter-annotator agreement in Croatian and Serbian.

A detailed error analysis was conducted through manual review of disagreement cases between human annotations and model predictions. A phonetician assessed 20 such instances per language and found that humans more often missed actual filled pauses (false negatives), while the model tended to over-predict (false positives). Notably, the model's outputs were more consistent and stable than human annotations, suggesting its potential superiority in recall and F1.

Qualitatively, the model successfully detected subtle, low-energy filled pauses that humans sometimes overlooked, especially voiced but weakly articulated sounds. False positives were most commonly triggered by nasals or prolonged vowels that resemble filled pauses acoustically.

An acoustic analysis using vowel space diagrams revealed that while filled pauses in Slovenian, Croatian, Serbian, and Czech are acoustically similar (clustered near /ə/), Polish filled pauses diverge, approximating /e/. Despite this, the model performed very well on Polish, indicating it leverages contextual and general speech features beyond raw acoustic similarity.

Keywords: filled pauses, speech transformers, cross-language predictions

Literature

Barrault, L., Chung, Y.-A., Meglioli, M. C., Dale, D., Dong, N., & al, et. (2023). Seamless: Multilingual Expressive and Streaming Speech Translation. arXiv. Retrieved from https://arxiv.org/abs/2312.05187

Bayerl, S. P., Wagner, D., Noth, E., & Riedhammer, K. (2022). Detecting dysfluencies in stuttering therapy using wav2vec 2.0. Interspeech.

Ljubešić, N., Rupnik, P., & Koržinek, D. (2024). The ParlaSpeech Collection of Automatically Generated Speech and Text Datasets from Parliamentary Proceedings. International Conference on Speech and Computer (pp. 137–150). Springer.

Romana, A., Koishida, K., & Provost, E. M. (2023). Automatic Disfluency Detection from Untranscribed Speech. Retrieved from https://arxiv.org/abs/2311.00867

Verdonik, D., Dobrovoljc, K., Rupnik, P., Ljubešić, N., Majhenič, S., Čibej, J., & Schmidt, T. (2024). Training corpus of spoken Slovenian ROG 1.0. Retrieved from http://hdl.handle.net/11356/1992

Razvoj modela za zaznavanje zapolnjenih premorov za slovenščino in vrednotenje njegove uporabnosti na naboru slovanskih jezikov

Čeprav so zapolnjeni premori med najpogostejšimi parajezikovnimi prvinami govora, so v transkripcijah večinoma izpuščeni. V tem prispevku predstavimo metodo za zaznavanje zapolnjenih premorov neposredno iz govornega signala, ki temelji na transformerjih, doučenih na slovenščini ter ovrednotenih na južnoslovanskih in zahodnoslovanskih jezikih. Vrednotenje metode pokaže, da so govorni transformerji izredno uspešni pri zaznavanju zapolnjenih premorov v jeziku, na katerem so bili doučeni. Nato ovrednotimo tudi čezjezične zmožnosti modela na dveh zelo sorodnih, južnoslovanskih jezikih (hrvaščina in srbščina) ter dveh manj sorodnih zahodnoslovanskih jezikih (češčina in poljščina). Rezultati razkrijejo močne čezjezične zmožnosti posploševanja modela z majhnim padcem uspešnosti. Poleg tega analiza napak pokaže, da model glede na vrednosti priklica in mere F1 presega človeške označevalce, medtem ko je po natančnosti nekoliko slabši. Poleg vrednotenja zmožnosti govornih transformerjev za zaznavanje zapolnjenih premorov v slovanskih jezikih objavimo tudi nove večjezične testne podatkovne množice in javno dostopen doučeni model, s čimer želimo podpreti nadaljnje raziskave in razvoj na področju obdelave govorjenega jezika.

Ključne besede: zapolnjeni premori, govorni transformerji, čezjezične napovedi

GOVORJENI JEZIK MED RAZISKOVANJEM IN TEHNOLOGIJO:
ZBORNIK POVZETKOV
D. Verdonik, N. Ljubešić (ur.)

Univerzitetna založba
Univerze v Mariboru

Prozodične in stavčne enote v govoru

DARINKA VERDONIK, JASNA VIDINIĆ
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor, Slovenija
darinka.verdonik@um.si, jasna.vidinic1@um.si

Raziskava preučuje razmerje med prozodičnimi in stavčnimi enotami v slovenskem spontanem govoru na podlagi ročno označenih primerov iz učnega korpusa ROG-Artur. Rezultati kažejo, da prozodične enote pogosto ne sovpadajo z mejo stavka in da je prozodično členjenje govora izredno kompleksen preplet različnih faktorjev.

1 Izhodišča raziskave

Raziskava se osredotoča na razmerje med prozodičnimi in stavčnimi enotami v slovenskem spontanem govoru, temelječ na učnem korpusu ROG-Artur (Verdonik et al., 2024). Prozodija preučuje lastnosti govora, kot so jakost, višina, hitrost, ritem ipd., ki pomembno vplivajo na členjenje in razumevanje govora (Peterek, 2000; Zwitter Vitez, 2018). Slovenska slovnica izpostavlja pomen členitve govora za preglednost in jasnost sporočila, vendar razlaga to členjenje precej konvencionalno: povedi naj bi bile členjene s premori na mejah med povedmi, med relativno samostojnimi deli iste povedi, po spremnem stavku premega govora, med posameznimi prirednimi deli, pred dostavki in tudi pred odvisniki (Toporišič 2000, str. 537–538). Velja torej prepričanje, da je govor s premori členjen v prozodične enote na način, da te enote večinoma sovpadajo z mejami stavkov. To prepričanje smo preverili na avtentičnem govornem gradivu in raziskali, (1) v kolikšni meri začetki in konci prozodičnih enot sovpadajo z mejami med stavčnimi enotami in (2) ali lahko prepoznamo skupne tipe takšnih primerov.

2 Metodologija

Raziskava je temeljila na ročnem označevanju 5337 prozodičnih enot (Beňuš, 2021; Fox, 2000; Izre'el in sod., 2020), pri katerih smo ročno označili enote, kjer konec enote ni svopadal z mejo stavka, enote, kjer začetek ni sovpadal z mejo stavka, in enote, kjer niti začetek niti konec nista sovpadala z mejo stavka. Označevanje je bilo narejeno s pomočjo orodij Praat (Boersma in Weenink, 2001), ki je omogočalo akustično analizo, in EXMARaLDA (Schmidt in Wörner, 2009), ki omogoča korpusno iskanje po dodanih oznakah.

3 Rezultati

Kvantitativni rezultati kažejo, da več kot polovica prozodičnih enot ne sovpada z začetkom in/ali koncem stavka. V kvalitativni analizi smo opredelili tri ključne vzorce prozodičnega členjenja znotraj stavkov: obotavljanje, členjenje govora in poudarjanje. Govorci si pogosto vzamejo dodaten čas za tvorjenje (ko iščejo ustrezno besedo ali nadaljevanje stavka, pri samopopravljanjih, da zajamejo sapo), ločujejo posamezne dele povedi po aktualnosti informacij, med nestavčnimi enakovrednimi deli, pred diskurznimi označevalci in presenetljivo pogosto tudi priključujejo veznik k predhodnemu, ne k naslednjemu stavku, in s premori in drugimi prozodičnimi sredstvi poudarjajo določeno vsebino.

Raziskava kaže, da členjenje govora ni zgolj zunanji pokazatelj slovničnih struktur, ampak kompleksen preplet kognitivnih omejitev ter diskurzivnih in informacijskih strategij govorca. Ugotovitve dopolnjujejo obstoječe slovenske raziskave o govoru in prozodiji ter poudarjajo potrebo po nadaljnjih študijah.

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta ARIS Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik (J7-4642).

Literatura

Beňuš, Š. (2021). *Investigating spoken English: A practical guide to phonetics and phonology using Praat*. Palgrave Macmillan.

Boersma, P. in Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *Glot international, 5*, 341–345.

Fox, A. (2000). *Prosodic Features and Prosodic Structure: The Phonology of Suprasegmentals* (1. izd.). Oxford University Press. https://doi.org/10.1093/oso/9780198237853.001.0001.

Izre'el, S. (2020). The basic unit of spoken language and the interfaces between prosody, discourse and syntax: A view from spontaneous spoken Hebrew. V Izre'el, S., Mello, H., Panunzi, A. in Raso, T. (Ur.), *In Search of Basic Units of Spoken Language. A corpus-driven approach* (str. 77–106). John Benjamins Publishing Company. 10.1075/scl.94.02izr.

Peterek, N. (2001). Recent Methods of Prosody Analysis. *The Prague Bulletin of Mathematical Linguistics 76*, 45–54. https://ufal.mff.cuni.cz/~peterek/ons/5-pbml01.pdf.

Schmidt, T. in Wörner, K. (2009). EXMARaLDA – Creating, Analysing and Sharing Spoken Language Corpora for Pragmatic Research. *International Pragmatics Association. Pragmatics 19(4),* 565–582. 10.1075/prag.19.4.06sch.

Toporišič, J. (2000). *Slovenska slovnica.* Založba Obzorja.

Verdonik, D., Ljubešić, N., Rupnik, P., Dobrovoljc, K., in Čibej, J. (2024). Izbor in urejanje gradiv za učni korpus govorjene slovenščine ROG. V Š. Arhar Holdt, T. Erjavec (Ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (JT&DH 2024) (str. 469-484). Inštitut za novejšo zgodovino. https://doi.org/10.5281/zenodo.13936426.

Zwitter Vitez, A. (2018). Enota analize spontanega govora: interakcija prozodije, pragmatike in skladnje. *Jezik in slovstvo 63*(2–3), 158–175. https://doi.org/10.4312/jis.63.2-3.157-175.

Prosodic and Syntactic Units in Speech

This study investigates the relationship between prosodic and syntactic units in Slovenian spontaneous speech, based on manually annotated samples from the training corpus ROG-Artur. The findings indicate that prosodic units often do not align with syntactic boundaries and that prosodic segmentation of speech constitutes a highly complex interplay of various factors.

Keywords: prosodic units, syntactic structure, spontaneous speech

GOVORJENI JEZIK MED RAZISKOVANJEM IN TEHNOLOGIJO:
ZBORNIK POVZETKOV
D. Verdonik, N. Ljubešić (ur.)

Univerzitetna založba
Univerze v Mariboru

Slovenian Parent-Child Communication Corpus EPIC-SI

AMANDA SAKSIDA,[1] MATIC PAVLIČ,[2] JONA JAVORŠEK,[2]
NIKOLA LJUBEŠIĆ,[3] MOJCA BRGLEZ,[2,3] GREGOR STRLE,[3] ŠPELA VINTAR[2,3]

[1] Educational Research Institute, Ljubljana, Slovenia
amanda.saksida@pei.si
[2] University of Ljubljana, Ljubljana, Slovenia
Matic.Pavlic@pef.uni-lj.si, Jan.Jona.Javorsek@ijs.si, mojca.brglez@ff.uni-lj.si, Spela.Vintar@ijs.si
[3] Jožef Stefan Institute, Ljubljana, Slovenia
nikola.ljubesic@ijs.si, Mojca.Brglez@ijs.si, gregor.strle@fe.uni-lj.si, Spela.Vintar@ijs.si

We present EPIC-SI, a new project to create a corpus of Slovenian parent-child communication (ages 12–48 months) with audio, video, and manual transcriptions. The project will address challenges in child speech processing, multimodal annotation, and NLP for non-standard language. We will adapt and benchmark NLP pipelines for child-directed and child speech in a low-resource language. The corpus will support ASR development, linguistic analysis, and fine-tuning of LLMs.

1 EPIC-SI project

The EPIC-SI project (Early Parent-Child Communication in Slovenian: Corpus-based Insights) will aim to create the first comprehensive and publicly accessible corpus of early child language in Slovenian. The project will collect 150–300 hours of audio-visual recordings of naturalistic parent-child interactions (ages 12–48 months), enriched with manually transcribed and linguistically annotated speech. The corpus will be aligned at word and phoneme level and include gesture and

posture annotations, creating a multimodal benchmark for early communication research and child-directed language technology development.

2 Challenges of automatic analysis of child and child-directed speech

Despite notable advances in speech processing (Radford et al., 2023), the automatic analysis of child and child-directed speech remains a major challenge due to high variability in phonation, prosody, disfluency, and syntactic irregularity (Shivakumar & Narayanan, 2022). For under-resourced languages like Slovenian, this gap is critical. The EPIC-SI corpus will support domain-specific adaptation of existing pipelines such as CLASSLA-Stanza (Ljubešić, Terčon & Dobrovoljc, 2024), improve forced alignment techniques on child speech (McAuliffe et al., 2017), and enable fine-tuning of ASR systems using transformer-based architectures (Jain et al., 2023).

3 Planned studies

The corpus will also support interdisciplinary studies on phonological development, early syntax and lexical growth, and the temporal coordination of multimodal cues in early turn-taking (Casillas et al., 2016; Saksida et al., 2024), which will enable further research on predictive value of gesture, gaze, and posture in early communication development (Iverson & Goldin-Meadow, 2005; Gama et al., 2024). The project will offer a possibility to collaborate on the development and evaluation of domain-adapted NLP tools. EPIC-SI aims to contribute a high-quality, longitudinal dataset to both linguistic research and the development of speech and language technologies for children, offering the first benchmark for child speech processing in Slovenian.

Keywords: Child Speech Processing, Low-Resource Language NLP, Multimodal Language Corpora

Literature

Casillas, M., Bobb, S. C., & Clark, E. V. (2016). The development of turn-taking in early communication. *Journal of Child Language*, *43*(6), 1310–1337. https://doi.org/10.1017/S0305000915000689

Gama, F., Misar, M., Navara, L., Popescu, S. T., & Hoffmann, M. (2024). ViTPose: Transformer-based pose estimation for infants in real-world scenes. *arXiv preprint arXiv:2406.17382*.

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, *16*(5), 367–371. https://doi.org/10.1111/j.0956-7976.2005.01542.x

Jain, R., Barcovschi, A., Yiwere, M. Y., Bigioi, D., Corcoran, P., & Cucu, H. (2023). Fine-tuning Wav2Vec2 for children's speech recognition. *IEEE Access*, *11*, 46938–46948. https://doi.org/10.1109/ACCESS.2023.3275106

Ljubešić, N., Terčon, L., & Dobrovoljc, K. (2024). Domain-specific linguistic annotation of Slovenian child speech. In *Proceedings of the Conference on Language Technologies and Digital Humanities (JT-DH-2024)*. https://doi.org/10.5281/zenodo.13936406

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of Interspeech 2017* (pp. 498–502).

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via Whisper. In *International Conference on Machine Learning* (pp. 28492–28518).

Saksida, A., Rebesco, R., Colombani, A., Pintonello, S., Tonon, E., Santoro, A. M., & Orzan, E. (2024). Revisiting non-verbal markers of early communication: Audiovisual perspectives from infants with hearing loss. *Frontiers in Pediatrics*, *11*, Article 1209754. https://doi.org/10.3389/fped.2023.1209754

Shivakumar, P. G., & Narayanan, S. (2022). Automatic processing of child speech: Challenges and future directions. *Computer Speech & Language*, *72*, 101289. https://doi.org/10.1016/j.csl.2021.101289

Slovenski korpus komunikacije med starši in otroki EPIC-SI

Predstavili bomo EPIC-SI, projekt za izgradnjo novega korpusa slovenske komunikacije med starši in otroki (starimi 12–48 mesecev) z avdio in video posnetki in ročnimi transkripcijami. Projekt bo obravnaval izzive pri obdelavi otroškega govora, multimodalnem označevanju in NLP za nestandardni jezik. Prilagodili bomo in primerjali NLP za otroški govor in otroku namenjen govor v jeziku z omejenimi viri. Korpus bo podpiral razvoj ASR, jezikoslovno analizo in natančno prilagajanje LLM.

Keywords: obdelava otroškega govora, NLP v jezikih z omejenimi viri, multimodalni jezikovni korpusi

GOVORJENI JEZIK MED RAZISKOVANJEM IN TEHNOLOGIJO: ZBORNIK POVZETKOV

DARINKA VERDONIK,[1] NIKOLA LJUBEŠIĆ[2] (UR.)

[1] Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor, Slovenija
darinka.verdonik@um.si
[2] Institut Jožef Stefan, Ljubljana, Slovenija
nikola.ljubesic@inz.si

Zbornik povzetkov s konference Govorjeni jezik med raziskovanjem in tehnologijo prinaša aktualne prispevke s presečišča govorjenih jezikovnih virov, jezikoslovja in govornih tehnologij. Predstavljeni so javno dostopni hrvaški otroški korpusi v CHILDES/TalkBank ter zbirka ParlaSpeech V3. Več prispevkov obravnava gradnjo in obdelavo govornih virov za slovenščino: od strategij občanske znanosti in odprtokodnih orodij (poravnava, anonimizacija, validacija, normalizacija) do fonetičnega zapisa v Digitalni slovarski bazi ter širjenja slovarskih virov z govorjenim besediščem. Raziskave segajo od (ne)tekočnosti in detekcije zapolnjenih premorov do razmerja med prozodičnimi in stavčnimi enotami ter izzivov narečne transkripcije; napovedan je tudi novi korpus zgodnje komunikacije EPIC-SI. Zbornik je odprtodostopen pod licenco CC BY-SA in je namenjen raziskovalcem jezikoslovja in govornih tehnologij ter širši strokovni skupnosti.

DOI
https://doi.org/10.18690/um.feri.9.2025

ISBN
978-961-299-050-3

Ključne besede:
govorni viri,
govorne tehnologije,
korpusno jezikoslovje,
jezikovni korpus,
raziskave govora

Univerzitetna založba Univerze v Mariboru

DOI
https://doi.org/10.18690/um.feri.9.2025

ISBN
978-961-299-050-3

Spoken Language Between Research and Technology: Book of Abstracts

Darinka Verdonik,[1] Nikola Ljubešić[2] (eds.)

[1] University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia
darinka.verdonik@um.si
[2] Jožef Stefan Institute, Ljubljana, Slovenia
nikola.ljubesic@inz.si

Keywords:
spoken language resource, speech technology, corpus linguistics, language corpus, speech research

The book of abstracts from the conference Spoken Language between Research and Technology brings timely contributions at the intersection of spoken language resources, linguistics, and speech technologies. It features publicly available Croatian child-language corpora in CHILDES/TalkBank and the ParlaSpeech V3 collection. Several papers address the creation and processing of Slovenian speech resources: from citizen-science strategies and open-source tools (alignment, anonymization, validation, normalization) to phonetic transcription in the Digital Dictionary Database of Slovene and the expansion of lexical resources with typically spoken vocabulary. The research spans (dis)fluency and filled-pause detection, the relationship between prosodic and syntactic units, and challenges of dialect transcription; a new EPIC-SI early communication corpus is also announced. The volume is open access under the CC BY-SA license and is intended for researchers in linguistics, corpus studies, and speech technologies, as well as the broader professional community.

University of Maribor Press

Univerza v Mariboru

Fakulteta za elektrotehniko,
računalništvo in informatiko

18. september 2025, Ljubljana, Slovenija

<..> mezzanine