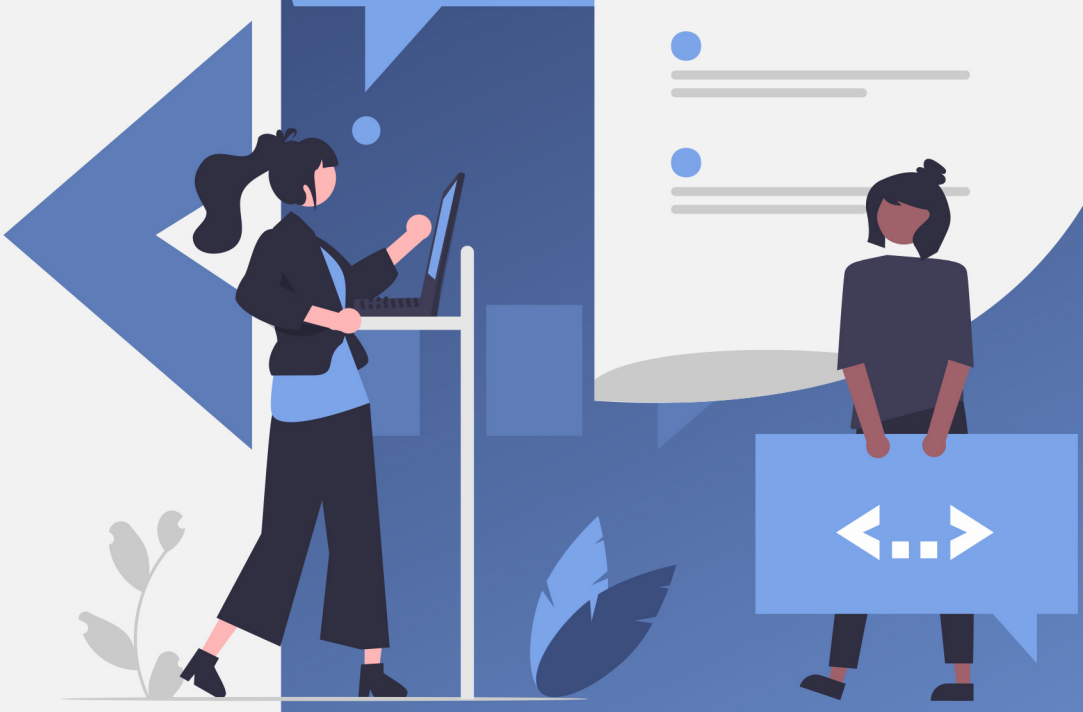
mezzanine

SLAVISTIČNI ZNANSTVENI PREMISLEKI

Univerzitetna založba Univerze v Mariboru



Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah

6. mednarodna znanstvena konferenca
Slavistični znanstveni premisleki

ZBORNIK POVZETKOV

Mira Krajnc Ivič
urednica

Univerza v Mariboru

Filozofska fakulteta

Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah

Zbornik povzetkov

Urednica

Mira Krajnc Ivič

Maj 2023

Naslov
Title
Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah
Infrastructure for Speech Research in the Humanities and Language Technologies

Podnaslov
Subtitle
Zbornik povzetkov
Book of Abstracts

Urednica
Editor
Mira Krajnc Ivič
(Univerza v Mariboru, Filozofska fakulteta)

Jezikovni pregled
Language editing
Darinka Verdonik (slovanski jeziki)
Melita Zemljak Jontes (angleški jezik)

Tehnična urednika
Technical editors
Dunja Legat
(Univerza v Mariboru, Univerzitetna založba)

Jan Perša
(Univerza v Mariboru, Univerzitetna založba)

Oblikovanje ovitka
Cover designer
Samo Kramberger

Jan Perša
(Univerza v Mariboru, Univerzitetna založba)

Grafika na ovitku
Cover graphic
Avtor Samo Kramberger, 2023

Grafične priloge
Graphic material
Avtorji prispevkov in Krajnc Ivič, 2023

Konferenca
Conference
6. mednarodna znanstvena konferenca Slavistični znanstveni premisleki z naslovom *Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah*

Datum in kraj
Date and place
18. 5.–19. 5. 2023, Maribor, Slovenija

Organizacijski odbor
Organizing committee
Mira Krajnc Ivič (predsednica, Univerza v Mariboru, Filozofska fakulteta), Gjoko Nikolovski (Univerza v Mariboru, Filozofska fakulteta), Natalija Ulčnik (Univerza v Mariboru, Filozofska fakulteta), Darinka Verdonik (Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko), Melita Zemljak Jontes (Univerza v Mariboru, 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
Filozofska fakulteta
Koroška cesta 160, 2000 Maribor, Slovenija
https://ff.um.si, ff@um.si

Izdaja
Edition
Prva izdaja

Vrsta publikacije
Publication type
E-knjiga

Dostopno na
Available at
http://press.um.si/index.php/ump/catalog/book/774

Izdano
Published at
Maribor, maj 2023

Naslov projekta: Temeljne raziskave za razvoj govornih virov in tehnologij za slovenščino – Mezzanine

Šifra projekta: J7-4642 Temeljne raziskave za razvoj govornih virov in tehnologij za slovenščino

Sofinancirano s strani Javne agencije za raziskovalno dejavnost Republike Slovenije.

Filozofska fakulteta

SLAVISTIČNO DRUŠTVO MARIBOR

Fakulteta za elektrotehniko, računalništvo in informatiko

Avtorice in avtorji prispevkov so odgovorni za jezikovno pravilnost svojega prispevka.

CIP - Kataložni zapis o publikaciji
Univerzitetna knjižnica Maribor

81'342(082)(0.034.2)

MEDNARODNA znanstvena konferenca Slavistični znanstveni premisleki (6 ; 2023 ; Maribor)
Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah [Elektronski vir] : zbornik povzetkov : [6. mednarodna znanstvena konferenca Slavistični znanstveni premisleki : 18. 5.-19. 5. 2023, Maribor, Slovenija] / urednica Mira Krajnc Ivič. - 1. izd. - E-zbornik. - Maribor : Univerza v Mariboru, Univerzitetna založba, 2023

Način dostopa (URL): https://press.um.si/index.php/ump/catalog/book/774
ISBN 978-961-286-735-5 (PDF)
doi: 10.18690/um.ff.5.2023
COBISS.SI-ID 150988291

ISBN 978-961-286-735-5 (pdf)

DOI https://doi.org/10.18690/um.ff.5.2023

Cena
Price Brezplačni izvod

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

Citiranje
Attribution Krajnc Ivič, M. (ur.). (2023). *Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: zbornik povzetkov*. Univerza v Mariboru, Univerzitetna založba. doi: 10.18690/um.ff.5.2023

Univerzitetna založba Univerze v Mariboru

Kazalo

POVZETKI PLENARNIH PREDAVANJ

Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: Zbornik povzetkov
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Slovenský hovorený korpus

Radovan Garabík
Slovenská akadémia vied, Jazykovedný ústav Ľ. Štúra, Bratislava, Slovensko
radovan.garabik@kassiopeia.juls.savba.sk

Slovenský hovorený korpus je korpusom súčasnej hovorenej slovenčiny, ktorý obsahuje manuálny prepis nahrávok, široký súbor štrukturálnych značiek označujúcich extralingvistické črty (ako sú pauzy, chyby, koktanie, smiech, externé zvuky atď.), odkazy na zvukové nahrávky a anotáciu metadát, ktorá obsahuje základné informácie o nahrávke a hovoriacich. Korpus je automaticky lematizovaný a morfologicky anotovaný a je prístupný prostredníctvom korpusového manažéra NoSketch Engine na webovej stránke ústavu.

Koncepcia „veľkého" reprezentatívneho korpusu hovoreného jazyka sa objavila prakticky hneď pri tvorbe koncepcie projektu Slovenského národného korpusu (t. j. v roku 2003) v nadväznosti na hlavný korpus písaného jazyka, ale k realizácii sme pristúpili až neskôr, prvá verzia hovoreného korpusu bola vytvorená v roku 2008.

Aktuálne je sprístupnená verzia korpusu *s-hovor-7.0*, ktorá obsahuje 869 zvukových nahrávok v celkovom rozsahu 851 hodín, čo predstavuje 7,8 milióna tokenov. Korpus sa skladá z dvoch hlavných častí – z primárnych nahrávok, a z nahrávok, ktoré poskytol Ústav pamäti národa (ÚPN). Časť korpusu, ktorá nepochádza z nahrávok ÚPN, obsahuje 4,2 milióna tokenov.

Prístup ku korpusu je možný prostredníctvom webového rozhrania NoSketch Engine (vyžaduje sa registrácia). V predchádzajúcich verziách bol korpus prístupný aj prostredníctvom osobitného webového rozhrania založeného na vlastnom

korpusovom manažéri, ktorého cieľom bolo vizuálne zlepšiť použiteľnosť poskytovaním farebných a graficky kódovaných značiek pre štrukturálne značky v korpuse (napríklad odkaz na zvukovú nahrávku bol zobrazený ako hudobná nota na začiatku výpovede). V neskorších revíziách sme však toto rozhranie nahradili nástrojom NoSketch Engine, ktorý používa štandardné štrukturálne značky v štýle XML, čo priblížilo rozhranie typickým používateľom našich korpusov.

Korpus je založený na podobných kritériách, aké sa používajú pri písaných korpusoch. Základnou jednotkou korpusu je dokument, ktorý zodpovedá jednej ucelenej nahrávke. Každý dokument sa ďalej delí na repliky (*turns*), ktoré predstavujú jednotlivé prehovory nahraných hovoriacich. Tieto zvyčajne zodpovedajú replikám v dialógu medzi hovoriacimi, ale dlhšie neprerušované prehovory môžu byť rozdelené do viacerých replík, ktoré sú základnými štrukturálnymi blokmi korpusu. Tieto repliky sú prepojené so zvukovými nahrávkami, zvyčajne v dĺžke niekoľkých sekúnd. Na rozdiel od písaných korpusov korpus nie je rozdelený na vety; namiesto toho jazykovú funkciu viet plnia práve repliky.

Prepis v korpuse obsahuje dve vrstvy – slovo (*word*) a výslovnosť (*pron*). Vrstva slova je prepisom do štandardnej slovenčiny bez ohľadu na prípadné výslovnostné odlišnosti; vrstva výslovnosti je tzv. ortografická transkripcia štandardným slovenským pravopisom s malým súborom značiek používaných na zachytenie odchýlok od štandardnej výslovnosti. V prepise nepoužívame fonetickú/fonematickú abecedu, pretože medzi slovenskými lingvistami nie je všeobecne prijímaný konsenzus fonematického modelu slovenčiny. Navyše slovenská fonetická abeceda (ktorá je vlastne fonematická) sa výrazne líši od medzinárodnej fonetickej abecedy. Vďaka tomuto prístupu je prepis ľahšie čitateľný používateľmi korpusu a umožňuje použiť bežné nástroje počítačového spracovania prirodzeného jazyka (lematizáciu a morfologickú analýzu), zároveň to urýchlilo celý proces ručného prepisu nahrávok.

Kľúčové slová: slovenčina, korpus, hovorený korpus, NLP, prepis

Corpus of Spoken Slovak

The *Corpus of Spoken Slovak* is a corpus of contemporary spoken Slovak. The corpus is accessible via a NoSketch Engine web interface on the Institute's webpage. The corpus includes manual transcriptions of the speech, a rich set of structural tags marking extralinguistic features (such as pauses, mistakes, stuttering, laughter, external sounds, etc.), links to the sound recordings, and metadata annotation that includes basic information about the recording and the speakers. The corpus is automatically lemmatized and morphologically annotated.

The concept of a "big" representative corpus of spoken language appeared practically at the beginning of the Slovak National Corpus project (i.e. since 2003), following the main written language corpus. However, it was not until 2008 that the first version of the spoken corpus appeared.

The current version *s-hovor-7.0* contains 869 records totaling 851 hours of audio recordings, equalling 7.8 million tokens. A particularly sizable part of the corpus comes from recordings provided by the Nation's Memory Institute (ÚPN) – an institution dedicated to the evaluation of the historical period marked by oppression. The non-ÚPN part of the corpus contains 4.2 million tokens.

The corpus can be accessed via a NoSketch Engine interface (registration required). Previously, we maintained a separate interface built upon a homebrew corpus manager, which aimed to improve usability by providing colour and graphic-coded visual clues for the structural tags in the corpus (for example, a clickable musical note at the beginning of the utterance). However, in later revisions, we replaced this interface with a NoSketch Engine that uses standard XML-like structural tags. This was done to make the interface more familiar to typical users of our corpora.

The corpus design was based on criteria similar to those used for written corpora. The basic unit of the corpus is the *document*, which corresponds to an individual recording. Each document is further divided into *turns*, which represent individual utterances made by recorded speakers. The turns typically correspond to the back-and-forth dialogue between speakers, but longer uninterrupted utterances may be

divided into multiple turns. The *turn* is the primary structural block of the corpus linked to its sound recording, typically lasting several seconds. Unlike written corpora, we eschew the division into sentences; instead, turns serve the linguistic function that sentences would in a written corpus.

The transcription in the corpus has two layers - the *word* and the *pronunciation*. The *word* is a transcription into a standard Slovak, disregarding any pronunciation differences; the *pronunciation* is the so-called orthographic transcription, using standard Slovak orthography, with a small controlled set of marks to indicate deviations from the standard pronunciation. We did not use a phonetic/phonemic alphabet, as there is no consensus among Slovak linguists about a general model of phonemic analysis of Slovak. Furthermore, the so-called Slovak phonetic alphabet (which is actually phonemic) differs significantly from the IPA. This makes the transcription easier to read and allows us to deploy usual NLP tools (such as morphology analysis, lemmatization). This also has made the whole transcription process significantly easier and faster.

Keywords: Slovak, speech, NLP, transcription, corpus

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Fonološka variantnost in korpus govorjene slovenščine

PETER JURGEC
Univerza v Torontu, Toronto, Kanada
peter.jurgec@utoronto.ca

Slovenska glasoslovna tradicija je bila vsaj v 20. stoletju skoraj izključno normativno naravnana: glavni namen je bilo izoblikovanje glasoslovnih značilnosti govorjene standardne slovenščine. Pri odločanju, kaj je standardno, se je vzpostavil kompromis med posplošitvami na podlagi ugotovitev teorije zgodovinskega ali strukturalnega jezikoslovja na eni strani in empiričnimi raziskavami na drugi (Dobrovoljc in Lengar Verovnik 2022). Težava je v tem, da so pri teh razpravah ključno vlogo igrale samo določene glasoslovne teme (npr. kvaliteta in kvantiteta naglašenih samoglasnikov, prim. Toporišič 2003), medtem ko so bile druge skoraj v celoti spregledane. Slednji primer je izgovarjava nosnikov v besedah *panj* ali *konjski*, kjer je dovoljen variantni izgovor, čeprav ni jasno, kakšna sploh je nezadlesničniška realizacija (daljša, palatalizirana ali palatalna). Empirična vrednost dejanske realizacije standardne slovenščine je delu jezikoslovcev v celoti nepomembna (Šekli 2022).

Glasoslovni opis slovenščine omejuje precejšnja zemljepisna variantnost, ki se zrcali tudi v realizaciji nadnarečne in splošnopogovorne variante. To omejuje ne le fonetični opis slovenščine, ampak tudi njene fonološke značilnosti. Takšna variantnost bo očitna tudi v še tako reprezentativnem govorjenem korpusu slovenščine.

V tem prispevku s primeri ponazorim, zakaj bi bil korpus govorjene slovenščine kljub omenjeni sociolingvistični pogojenosti vseeno uporaben za fonološke, lahko pa tudi fonetične, raziskave slovenščine.

Tak primer je naglasno mesto v slovenščini, ki je v slovenistični literaturi večinoma opisano kot razlikovalno, torej sinhrono večinoma nepredvidljivo (Toporišič 2000: 66, 112). Vendar pa novejša dognanja kažejo, da je naglas v veliki meri morfološko predvidljiv. V netvorjenih samostalnikih je daleč najpogostejši naglas na zadnjem zlogu korena, široki samoglasniki in polglasnik pa se naglasa izogibajo (Becker in Jurgec 2000). Slovenska povezava med samoglasniško kvaliteto in naglašenostjo je odločilno informirala splošno fonološko teorijo vokalne redukcije (Crosswhite 2001). Dodatna kompleksnost so tudi stranski naglasi v prevzetih besedah in nekaterih tvorjenkah (Jošt 2015), kar je podobno ruščini (Gouskva 2010). Čeprav se posamezni slovenski govori razlikujejo glede števila samoglasnika in naglasnega mesta, je skoraj vsem skupna povezava med samoglasniško kvaliteto in naglašenostjo.

V prispevku ponazorim še več takih primerov, ki jih bo mogoče podrobno raziskati šele v dovolj velikem korpusu slovenščine: (i) izgovor prevzetih besed, zlasti če so morfološko kompleksne, (ii) končna nezvenečnost, (iii) realizacija sičnikov in šumnikov znotraj iste besede in (iv) izgovarjava posameznih glasov. V tem smislu bi korpus govorjene slovenščine lahko bistveno prispeval k preseganju preteklih glasoslovnih razhajanj.

Ključne besede: fonologija, variantnost, govorni korpus, slovenščina

Phonological Variation and the Corpus of Spoken Slovenian

A key question among Slovenian linguists in the 20th century was to determine the phonetic and phonological properties of Standard Slovenian. In deciding what this prestigous variety of Slovenian should be, there was tension between prescription

(based on findings of historical and structuralist linguistics) and actual realization among the speakers (Dobrovoljc and Lengar Verovnik 2022). The challenge of this situation was that only certain topics were extensively discussed (e.g. quality and quantity of stressed vowels, Toporišič 2003), while others were almost entirely ignored. For example, the non-alveolar relization of the nasals in words like *pa̲n̲j* 'hive' and *ko̲n̲jski* 'equine' has never been accurately described (»longer«, »palatalized« or »palatal«). Some linguists even question the value of empirical studies of Standard Slovenian at all (Šekli 2022).

The challenge of Slovenian phonetic and phonological description has to do with its substantial geographical and other sociolinguistic variation. This will be a persistent question in any spoken corpus of Slovenian.

This paper illustrates how the corpus of spoken Slovenian can nevertheless address key phonological and phonetic questions, considering stress, which is largely described as constrastive and unpredictable (Tororišič 2000: 66, 122). Recent studies, however, suggest that stress in Slovenian is laregly morphologically predictable. Stress in bare nouns falls mostly on the final syllable of the root, while vowel quality also has a significant role, with lax vowels and schwa avoiding stress (Becker and Jurgec 2000). The Slovenian connection between vowel quality and stress has played a central role in the theories of vowel reduction (Crosswhite 2001). An additional complication concerns secondary stress in loanwords and morphologically complex words (Jošt 2015), where Slovenian appears to be remarkably similar to Russian (Gouskova 2010). While the dialects of Slovenian differ significantly in terms of vowel quality and stress, almost all show a connection between the two.

I extend this reasoning to several other phenomena: (i) the phonological properties of loanwords, (ii) final devoicing, (iii) sibilant harmony, and (iv) phonetic realization of inidividual sounds. These questions can be accurately answered by studying a representative corpus of spoken Slovenian. I hope this would help shift the debate among Slovenian linguists.

Keywords: phonology, variation, spoken corpus, Slovenian

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Intenziteta jezika – večpredstavna upovedovalna določitev

VESNA MIKOLIČ
Znanstveno-raziskovalno središče Koper, Inštitut za jezikoslovne študije, Koper, Slovenija
vesna.mikolic@zrs-kp.si
Univerza v Trstu, Trst, Italija
vmikolic@units.it

V predavanju bomo izhajali iz analize procesa upovedovanja, ki kaže na to, da vsak govor vključuje referenčni pomen, osnovno vsebino, propozicijo na eni strani in vrednotenje, naklonskost, modalnost, modus na drugi. V slovenskem jezikoslovju je razmerje med propozicijo in upovedovalnimi določitvami/modifikacijami v svoji slovnici temeljito predstavil Jože Toporišič.

Ko se iz propozicije tvori poved in se torej upoveduje odnos sporočevalca do zunajjezikovne dejanskosti in do prejemnika, vedno prihaja do subjektivne izbire jezikovnih sredstev in prilagoditev pomena govorčevemu videnju stvarne vsebine in njegovemu odnosu do sogovorca. Po Michaelu Stubbsu bi tako naklonskost morali obravnavati kot posebno jezikovno ravnino, modalna slovnica pa bi po njegovem mnenju morala biti v središču jezikoslovnega zanimanja.

Ta izraženi odnos sporočevalca do zunajjezikovne dejanskosti ali prejemnika pa je lahko bolj ali manj nevtralen, blizu stvarnemu stanju zunajjezikovne dejanskosti in prejemnika, lahko pa pride v tem okviru tudi do odmika od nevtralne stopnje intenzitete propozicijskega pomena, bodisi da sporočevalec svoj odnos do zunajjezikovne dejanskosti ali do prejemnika stopnjuje navzgor ali navzdol. V okvir

naklonskosti sodi torej tudi intenzifikacija ali modifikacija intenzitete, bodisi da gre za krepitev ali šibitev, krepilce ali šibilce posameznih sestavin ali govora kot celote.

Pri intenziteti jezika gre za modifikacijo tako propozicijskih kot naklonskih sestavin, kamor sodi tudi sporočevalčev odnos do naslovnika. Kompleksnost pojava pa je še večja, če upoštevamo, da je modifikacija intenzitete določena tudi z lastnostmi posameznega jezika in kulture ter različnimi komunikacijskimi praksami. Te pa se spet ločujejo glede na prenosnik, tako se označevalci intenzitete razlikujejo glede na to, ali gre za vidni in slušni prenosnik oz. ali gre za pisni ali govorjeni diskurz.

V predavanju bomo tako predstavili razvoj koncepta intenzitete jezika in podali slovnični in slovarski opis intenzitete v slovenskem jeziku. Prav tako bomo predstavili model za analizo intenzitete jezika, v kateri so zajeta jezikovna sredstva za izražanje intenzitete na vseh jezikovnih ravninah, to so: 1. glasoslovni in pravopisni izrazi intenzitete ter sredstva neverbalne komunikacije, 2. oblikoslovni izrazi intenzitete, 3. skladenjski izrazi intenzitete, 4. pomenoslovni ali besedoslovni izrazi intenzitete, 5. besediloslovni ali besedilni izrazi intenzitete, 6. pragmatični ali diskurzni izrazi intenzitete. Poseben poudarek bomo posvetili označevalcem intenzitete, ki izstopajo v govorjenem diskurzu, in sicer so to predvsem izrazi intenzitete na ravni neverbalnega sporazumevanja in glasoslovni ravni, kot so: stavčni poudarek, tonski potek ali intonacija, premori, hitrost, register, barva glasu, glasovne figure, mašila, očesni stik, gestikulacija, obrazna mimika, glasovni razpon, telesna drža, premiki telesa ipd.

Tako se prav intenziteta jezika pokaže kot tista plast govora, zaradi katere je nujno potrebno upoštevati večpredstavnost govorne komunikacije, ko načrtujemo analizo govorjenega diskurza.

Ključne besede: naklonskost, intenziteta jezika, označevalci intenzitete, krepilec, šibilec

Language Intensity – A Multi-Representative Modification

The lecture begins with the analysis of the verbalisation process, which shows that every discourse contains, on the one hand, a referential meaning, a basic content, a proposition, and, on the other hand, evaluation, modality, mode. In Slovenian linguistics, Jože Toporišič described the relationship between propositions and modifications in detail in his grammar.

When a sentence is formed from proposition and at the same time the attitude of the speaker towards the extra-linguistic reality and the receiver is established, there is always a subjective choice of linguistic means and an adaptation of the meaning to the speaker's view of the real content and his/her attitude towards the interlocutor. Therefore, according to Michael Stubbs, modality should be considered as a special linguistic level and a modal grammar should, in his opinion, be the focus of linguistic interest.

This expressed attitude of the communicator towards the extra-linguistic reality or the receiver can be more or less neutral, close to the real state of the extra-linguistic reality and the receiver, but in this context there can also be a deviation from the neutral level of intensity of the propositional meaning, either that the communicator intensifies or weakens his attitude towards the extra-linguistic reality or the receiver. Therefore, intensification or modification of intensity is also part of the framework of modality, be it amplification or attenuation, the use of intensifier or mitigators of individual components or of a discourse as a whole.

The language intensity is a modification of both, the proposition and the modification itself, which also includes the communicator's attitude towards the addressee. The complexity of the phenomenon becomes even greater when we consider that the modification of intensity is also determined by the characteristics of the language and culture in question, as well as by various communication practises. These are in turn differentiated according to the communication channel,

so that intensity markers differ depending on whether the channel is visual or auditory, or whether the discourse is written or oral.

In the presentation we will introduce the development of the concept of language intensity and give a grammatical and dictionary description of intensity in Slovenian. We will also present a model for the analysis of language intensity, which includes linguistic means of expressing intensity at all linguistic levels, i.e.: 1. phonetic and orthographic expressions of intensity and means of non-verbal communication, 2. morphological expressions of intensity, 3. syntactic expressions of intensity, 4. semantic or lexical expressions of intensity, 5. lexical or textual expressions of intensity, 6. pragmatic or discursive expressions of intensity. We pay particular attention to the intensity features that are salient in spoken discourse, i.e. the intensity expressions at the level of non-verbal communication and the phonological level, such as: sentence stress, tone of voice or intonation, pauses, speed, register, voice colour, vocal figures, exclamations, eye contact, gestures, facial expressions, vocal range, posture, body movements, etc.

Thus, it is precisely the intensity of speech that proves to be the level of language that makes it absolutely necessary to take into account the multimedia character of spoken communication when planning the analysis of spoken discourse.

Keywords: modality, intensity of language, intensity markers, intensifier, mitigator

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

60 let pozneje – pomen analize govorjenega diskurza

MOJCA SMOLEJ
Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija
mojca.smolej@ff.uni-lj.si

V prispevku bo najprej podan kratek pregled začetkov raziskav spontano govorjenega jezika na Slovenskem. Pred slabimi 60 leti je enega prvih prispevkov objavila Breda Pogorelec. Gre za prispevek *Vprašanja govorjenega jezika*, objavljenega v *Jezikovnih pogovorih*. Navedeni prispevek ni le pomemben za razvoj metodoloških in teoretičnih pristopov k analizam spontano govorjenega diskurza, pač pa pomeni tudi pomembno prelomnico za jezikoslovne vede nasploh. V 60. letih prejšnjega stoletja so bili namreč v ospredju še vedno predvsem historična slovnica, dialektologija in knjižni, torej pisni jezik. Mejnik k preučevanju zakonitosti govorjenega diskurza pomenita npr. tudi deli *Interesne govorice sleng, žargon, argo* avtorja Velemirja Gjurina in *Poglavje iz govornega jezika tržaških Slovencev: govorni signali* avtorice Žive Gruden.

Izhajajoč iz začetkov preučevanja govorjenega diskurza, bodo v prispevku podani tudi nekateri premisleki, zakaj je nujno, da se s preučevanjem nadaljuje in nadgrajuje predhodne raziskave. Podani bodo s stališča opisne slovnice tako govorjenega kot pisnega (knjižnega) jezika. Razumevanje zakonitosti govorjenega jezika lahko namreč pomaga pri razumevanju in reševanju nekaterih nejasnih ali pomanjkljivo opredeljenih slovničnih opisov (predpisov) knjižnega jezika. Nezadostno opredeljene ali opisane slovnične zakonitosti knjižnega jezika lahko posledično povzročajo tako stisko na ravni poučevanja slovenščine kot prvega in drugega/tujega jezika, prav tako pa tudi na ravni rabe jezika pri splošnih govorcih (torej

nejezikoslovcih). Ena izmed pomanjkljivo in obenem celo napačno predstavljenih slovničnih kategorij v SS je npr. kategorija skladenjskih razmerij in s tem povezanih pomenskih razmerij. Vzporedno z analizo pisnega (knjižnega) jezika bi morali nujno opraviti tudi ekvivalentno raziskavo v govorjenih korpusih. Raven izražanja podrednih medstavčnih skladenjskih razmerij v govorjenem jeziku je namreč slabo raziskana, kar je velika pomanjkljivost, saj ravno poznavanje zakonitosti govorjenega jezika lahko olajša ali pojasni marsikatero težavo, povezano s pisnim (knjižnim) jezikom. Časovni, vzročni in načinovni odvisniki, katerih propozicije niso del propozicije matičnega stavka, so tipični predvsem za knjižni jezik, v spontano govorjenem jeziku so ta razmerja izražena največkrat s prirednimi skladenjskimi sredstvi.

Na nujnost in smiselnost analize govorjenega diskurza bomo tako pogledali predvsem z zornega kota slovničnega opisa, brez katerega si dandanes ne smemo in ne moremo več predstavljati kateregakoli preučevanja slovničnih zakonitosti pisnega oz. knjižnega jezika.

Ključne besede: govorjeni diskurz, Breda Pogorelec, skladnja, diskurzni označevalci, knjižni (standardni) jezik

60 Years Later – The Importance Of Spoken Discourse Analysis

The paper will first give a brief overview of the beginnings of research on spontaneous spoken language in Slovenia. One of the first papers was published by Breda Pogorelec almost 60 years ago. This is an article *V prašanja govorjenega jezika* published in Jezikovni pogovori. This paper is not only important for the development of methodological and theoretical approaches to the analysis of spontaneous spoken discourse, but also represents an important turning point for linguistics in general. In the 1960s, the focus was still primarily on historical grammar, dialectology and the written (standard) language. Important contributions

to the study of the characteristics of spoken discourse are also made, for example, by Velemir Gjurin's article *Interesne govorice sleng, žargon, argo* and Živa Gruden's *Poglavje iz govornega jezika tržaških Slovencev: govorni signali.*

Starting from the beginnings of the study of spoken discourse, the paper will also provide some reflections on why it is necessary to continue and build on previous research. They will be presented from the point of view of descriptive grammar of both spoken and written (standard) language. Understanding the regularities of spoken language can help understand and resolve some unclear or insufficiently-defined grammatical descriptions (rules) of written language. The insufficiently defined or described grammatical regularities of the spoken language can consequently cause distress both at the level of teaching Slovene as a first and second/foreign language, as well as at the level of language use by general speakers (i.e. non-linguists). One of the grammatical categories in the *Slovenska slovnica* that is not well represented and even misrepresented is, for example, the category of syntactic relations and related semantic relations. In parallel to the analysis of the written (standard) language, an equivalent study in spoken corpora should necessarily be carried out. The level of expression of subordinate syntactic relationships in spoken language is poorly understood, which is a major disadvantage, since knowledge of the regularities of spoken language can clarify many problems related to written language. Subordinate temporal and causal sentences are typical of written language, whereas in spontaneous spoken language these relations are most often expressed by means of coordinate syntactic sentences.

The necessity and meaningfulness of the analysis of spoken discourse will thus be looked at primarily from the point of view of grammatical description, without which any study of the grammatical regularities of written language should not be, and can no longer be, imagined nowadays.

Keywords: spoken discourse, Breda Pogorelec, syntax, discourse markers, the standard language

POVZETKI PRISPEVKOV

Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: Zbornik povzetkov
M. Krajnc Ivič (ur.)

Univerzitetna založba Univerze v Mariboru

Oslovljavanje članova najbliže rodbine nekad i danas

Marko Alerić
Sveučilište u Zagrebu, Filozofski fakultet, Zagreb, Hrvatska
maleric@ffzg.hr

U oslovljavanju se, osim imenom i/ili prezimenom, odnosno nadimkom, služimo i ličnim zamjenicama. Upotreba ličnih (osobnih) zamjenica u oslovljavanju podrazumijeva usvojenost i primjenu socijalne kompetencije koja prevladava u odgovarajućoj užoj ili široj društvenoj zajednici. U istraživanju će, na temelju izjava kazivača, biti utvrđen i protumačen način oslovljavanja članova najbliže rodbine (majka, otac, djeca, baka, djed, unuci) upotrebom ličnih (osobnih) zamjenica u hrvatskim mjesnim govorima u tri razdoblja: 1. od otprilike 1900. do 1940. (bake i djedovi) 2. od 1941. do 1980. (majke i očevi) i 3. od 1981. do danas (djeca). Istraživanje će biti provedeno na pisanim izvorima i na terenskim istraživanjima. Bit će utvrđeno i protumačeno dolazi li s vremenom do promjena u načinu oslovljavanja upotrebom ličnih (osobnih) zamjenica i čime su te promjene uzrokovane.

Važnost istraživanja je u utvrđivanju i praćenju načina oslovljavanja članova najbliže rodbine i u utvrđivanju lingvističkih i sociolingvistički razloga koji uzrokuju promjene. Rezultati istraživanja omogućit će utvrđivanje načina oslovljavanja članova najbliže rodbine u Hrvatskoj nekad i danas, kao i daljnja lingvistička, sociolingvistička, dijalektološka, ali i sociološka istraživanja.

Ključne riječi: hrvatski govori, načini oslovljavanja, lične zamjenice, jezične promjene

Addresing Members of the Closest Relatives in the Past and Today

In addressing each other, in addition to the first and/or last name, or nickname, we also use personal pronouns. The use of personal pronouns in addressing implies the adoption and application of social competences that prevail in the respective close or wider social community. In the research, based on the statements of the tellers, the way of addressing members of the closest relatives (mother, father, children, grandmother, grandfather, grandchildren) will be determined and interpreted using personal pronouns in Croatian local dialects in three periods: 1. from approximately 1900 to 1940 (grandparents) 2. from 1941 to 1980 (mothers and fathers) and 3. from 1981 to the present (children). The research will be conducted on written sources and on field research. It will be determined and interpreted whether there are changes in the way of addressing with the use of personal pronouns over time and what caused these changes.

The importance of the research is in determining and monitoring the way members of the closest relatives are addressed and in determining the linguistic and sociolinguistic reasons that cause changes. The results of the research will make it possible to determine the way of addressing members of the closest relatives in Croatia then and now, as well as further linguistic, sociolinguistic, dialectological, and sociological research.

Keywords: Croatian dialects, ways of addressing, personal pronouns, language changes

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Jezik influencera u kontekstu novih, novih medija

BORKO BARABAN, SNJEŽANA BARIČ-ŠELMIĆ
Sveučilište Josipa Juraja Strossmayera u Osijeku, Akademija za umjetnost i kulturu, Osijek, Hrvatska
borko.baraban@aukos.hr, sbaric@aukos.hr

Velik tehnološki napredak digitalnog i umreženoga društva omogućilo je stvaranje novih, novih medija koji su iznjedrili novu publiku – *prosumere*. Novi, novi mediji omogućuju njihovu korisniku istovremeno stvaranje i konzumiranje sadržaja. Upravo taj trenutak obilježit će uspon *influencera* – utjecajnih osoba. Utjecajne osobe odnosno kreatori javnoga mišljenja nisu novum, no za razliku od utjecajnih osoba u prošlosti, kraljevskih obitelji, plemstva, političke elite, sportaša i drugih, današnje utjecajne osobe svoju popularnost i utjecaj stječu prvenstveno zahvaljujući razvoju tehnologije, a stavove publike oblikuju stvaranjem objava, *vblogova*, *tweetova* i drugih kanala mrežnih društvenih medija.

Ovaj rad bavit će se opisom jezika influencera i to na dvjema razinama: leksičkoj i sintaktičkoj. Cilj je rada utvrditi utjecaj novih, novih medija na strukturu jezika. Metodološki okvir rada usmjeren je kvalitativnoj metodi analize sadržaja koja se može kvantificirati. To je istraživačka tehnika kojom se klasificira i opisuje komunikacijski sadržaj prema unaprijed određenim kategorijama, odnosno komunikacija će se analizirati na sustavan i kvantitativan način. Nastavno istraživanju koje je provela Styria i agencija Nielsen, prema metodologiji NielsenMedia i alatu InfluenceScope (2021), određen je uzorak istraživanja.

Ključne riječi: jezik influencera; kvalitativna analiza sadržaja; leksik; novi, novi mediji; sintaksa

The Language Of Influencers In The Context Of New, New Media

The tremendous technological advancement of the digital and networked society enabled the creation of new, new media that create a new audience - *prosumers*. New, new media allow their users to produce and consume content simultaneously. It is precisely this momentum that will mark the rise of influencers. Influential persons or creators of public opinion are not a novelty, but unlike influential persons in the past, royal families, nobility, political elite, athletes and others, today's influential persons gain their popularity and influence primarily thanks to the development of technology, and they shape the attitudes of the audience by creating announcements, vblogs, tweets and other online social media channels.

This paper will describe influencers' language at the following two levels: lexical and syntactic. The methodological framework has been directed towards the qualitative method of content analysis that can be quantified. It is a research technique used to classify and describe communication content according to predetermined categories, that is, communication will be analysed in a systematic and quantitative way. Following the research conducted by Styria and the Nielsen agency, according to the NielsenMedia methodology and the InfluenceScope tool (2021), a research sample was determined.

Keywords: influencers' language; lexis; new, new media; qualitative content analysis; syntax

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)

Univerzitetna založba Univerze v Mariboru

Konteksti snemanja govorjenega diskurza v sociolingvistiki

MAJA BITENC
Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija
maja.bitenc@ff.uni-lj.si

Prispevek se osredotoča na kontekste snemanja govorjenega diskurza, predvsem s sociolingvističnega vidika, pri čemer vključuje pregled metodoloških pristopov v različnih tujih in slovenskih raziskavah.

Sociolingvistika od vsega začetka oz. od pionirskih raziskav Williama Labova v ZDA in Petra Trudgilla v Veliki Britaniji v 60. in 70. letih 20. stoletja poudarja vpliv situacijskega konteksta na posameznikov govor: govorci izbirajo različne variante in varietete iz svojega govornega repertoarja glede na številne pragmatične dejavnike, npr. sogovorca, morebitne druge udeležence, kraj, temo pogovora, formalnost situacije, želen učinek na naslovnika in pozornost, posvečeno govoru. V različnih situacijah govorci uporabljajo različne deleže jezikovnih variant posameznih variabel; tako je pri vsaki analizi jezikovnega sistema in njegove variantnosti ključen podatek o tem, v kateri situaciji oz. s kom se določene variante uporablja in v kolikšni meri. Na kontinuumu med narečjem in standardnim jezikom se načeloma z naraščajočo formalnostjo in javnostjo situacije ter večjo pozornostjo, posvečeno govoru, viša delež standardnih variant. V sociolingvističnih raziskavah se posebno pozornost namenja iskanju načinov, kako presegati t. i. opazovalčev paradoks, da bi pridobili posnetke čim bolj avtentičnega govora in analizirali, kako ljudje govorijo, kadar niso opazovani.

V slovenskem jezikoslovju se je večina raziskav govorjene slovenščine osredotočala na eno od govornih varietet v govornem repertoarju izbranih govorcev, redki pa so primeri analize govorne variantnosti glede na okoliščine. Slovenska dialektologija se večinoma posveča raziskavam čim bolj tradicionalnih narečij. Osrednjeslovenska varieteta Ljubljane in okolice je bila predmet fonoloških raziskav na nereprezentativnih vzorcih govorcev z namenom standardizacije in načrtovanja govorjenega standarda od druge polovice 20. stoletja naprej. Tudi gradivo za Slovenski govorni korpus GOS večinoma vključuje govor posameznega govorca samo v eni izbrani situaciji. Variantnosti pri posameznih govorcih v različnih okoljih in okoliščinah so se z različnimi pristopi posvečale npr. Jožica Škofic (Guzej), Irina Makarova, Melita Zemljak Jontes in Simona Pulko. Prvi primer variantnostne analize na podlagi izbranih jezikovnih variabel so raziskave Maje Bitenc, pri katerih so se z namenom proučevanja govorne variantnosti geografsko mobilni govorci iz Idrije, Ribnice in Maribora in okolice samosnemali v različnih vsakodnevnih okoliščinah. Predstavljeni so kritični pomisleki glede posameznih pristopov.

Prispevek prinaša tudi pregled načinov in kontekstov pridobivanja posnetkov v različnih tujih raziskavah – od pionirskih sociolingvističnih intervjujev Williama Labova v ZDA z izvabljanjem različnih govornih stilov (običajni stil pri vsakdanjem govoru, pazljivi stil pri odgovorih na zastavljena vprašanja, branje besedila, seznama besed in minimalnih parov) do metodoloških pristopov v sodobnih sociolingvističnih projektih v jezikovnih skupnostih, katerih sociolingvistični profili so bolj sorodni slovenskemu, npr. v Avstriji, Nemčiji, Belgiji in na Madžarskem. Ti vključujejo npr. branje besedila oz. seznama besed, prevod iz narečja v standard, prevod iz standarda v narečje, neformalni pogovor s prijateljem iz istega kraja ali z drugega narečnega področja, formalni sociolingvistični intervju z raziskovalcem, posnetke skupinskih srečanj, posnetke, pridobljene za druge namene, ter opazovanje z udeležbo.

Ključne besede: snemanje govora, konteksti jezikovne rabe, jezikovna variantnost, sociolingvistični intervju, opazovalčev paradoks

Contexts of Speech Recording in Sociolinguistics

The contribution focuses on the contexts of eliciting and recording spoken discourse, especially from the sociolinguistic point of view, and includes a presentation of methodological approaches in various Slovene and foreign studies and research projects.

Sociolinguistics points out (beginning in the 1960s and 1970s with Willilam Labov's research in the USA and Peter Trudgill's in Great Britain) to the influence of the situational context for an individual's speech. Speakers choose various variants and varieties from their speech repertoires depending on numerous pragmatic factors, e.g. the interlocutor, potential other participants, the setting, the topic, the intended effect on the addressee, and the attention, paid to speech. Different situational settings yield different amounts of existing language variants; therefore, in any analysis of the language system and its variation, it is significant in which situation and with whom certain variants are used and to what degree. For dialect–standard continua in particular, with an increasing public orientation and formality of the situation as well as attention, paid to speech, generally a shift towards the standard pole is implied. In sociolinguistic research, special attention is paid to finding ways to overcome the so called observer's paradox, in order to obtain authentic samples of speech and find out how people speak when we they are not observed.

In Slovene linguistics, the majority of research into spoken Slovene has focused on one of the speech varieties in the speech repertoire of selected speakers, whereas examples of variation analysis according to different circumstances are rare. Slovene dialectology has mainly dealt with traditional dialects. The central Slovene variety of Ljubljana and its surroundings has been the object of phonological researches on unrepresentative samples of speakers, connected with standardisation and spoken standard planning since the second half of the 20th century. Also the Corpus of spoken Slovene GOS mostly includes the speech of an individual speaker only in one particular situation. Language variation of individual speakers in different environments and circumstances has been studied by e.g. Jožica Škofic (Guzej), Irina

Makarova, Simona Pulko and Melita Zemljak Jontes. The first example of the variation analysis based on selected linguistic variables are studies by Maja Bitenc, in which geographically mobile speakers from Idrija, Ribnica and Maribor and the surroundings self-recorded their speech in various everyday situations. Critical evaluation regarding different approaches is presented.

In addition, the contribution provides an overview of the methods and contexts of obtaining speech samples in various foreign studies – from William Labov's pioneering sociolinguistic interviews in the USA with the elicitation of different speech styles (common style in everyday speech, careful style in answering questions, reading a text, a word list and minimal pairs) to methodological approaches in modern sociolinguistic projects in language communities, whose sociolinguistic profiles are more similar to Slovene, e.g. in Austria, Germany, Belgium and Hungary. These include e.g. reading a text or a word list, translation from dialect to standard, translation from standard to dialect, informal conversation with a friend from the same town or from another dialect area, formal sociolinguistic interview with the researcher, recording of group meetings, recordings obtained for other purposes, and participant observation.

Keywords: speech recording, contexts of language use, linguistic variation, sociolinguistic interview, observer's paradox

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)

Univerzitetna založba Univerze v Mariboru

Transkribiranje v sociolingvističnih raziskavah

MAJA BITENC
Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija
maja.bitenc@ff.uni-lj.si

Transkribiranje je kompleksen, interpretativen in selektiven proces, pri katerem pretvarjanje iz govorjenega v zapisano besedilo odpira številna temeljna in praktična vprašanja. Pri transkripciji gre vedno za interpretacijo, saj nikoli ne more biti povsem zanesljiva predstavitev originalne interakcije; pri pretvarjanju besedila iz posnetka v zapis ni mogoče vsega istočasno in izčrpno zabeležiti. V sociolingvističnih raziskavah ni standardnih transkripcijskih načel, temveč so ta vedno odvisna od teoretske usmeritve raziskovalca, predmeta in namena raziskave – torej kaj in zakaj se proučuje. Selektivnost je tako tesno povezana s subjektivnostjo. Osnovni vodili pri transkripciji za variantnostne raziskave govora sta, naj bodo transkripcije dovolj natančne, da učinkovito ohranijo dovolj podatkov za želeno jezikoslovno analizo, in dovolj preproste, da jih je mogoče sorazmerno enostavno brati in zapisovati. Raziskovalec se torej odloči, koliko in katere vrste informacij zapisovati, kaj potrebuje in kaj je relevantno za namen aktualne oz. morebitne kasnejše raziskave.

Osrednji del prispevka se osredotoča na izkušnje in različne prakse transkribiranja v avtoričinih sociolingvističnih raziskavah variantnosti govorjene slovenščine in relevantnih socialnopsiholoških tem pri govorcih iz različnih narečnih skupin – od rovtarske (govorci iz Idrije in okolice) do dolenjske (Ribnica) in štajerske (Maribor). Avtorica je prvotno uporabljala natančne fonetične transkripcije z vnašalnim sistemom ZRCola, ki ga je razvil Peter Weiss, potem pa se odločila za različne

stopnje poenostavljanja in prehajanje k ortografskemu zapisu s posebnimi znaki za posamezne variable in njihove variante, ki so relevantne za analizo variantnosti govora posameznih informantov v različnih okoliščinah, z različnimi sogovorci. Variable so seveda različne v posameznih proučevanih narečnih skupinah. Primerjalno so predstavljena tudi transkripcijska načela v nekaterih drugih sorodnih raziskavah in projektih, od dialektoloških raziskav do različnih študij govorjenega jezika, ki se osredotočajo na različna raziskovalna vprašanja (npr. raziskave Jožice Škofic (Guzej), Irine Makarove, Melite Zemljak Jontes in Simone Pulko, Mojce Smolej, Darinke Verdonik), in slovenskega govornega korpusa GOS. Posebna pozornost je posvečena problematiki zapisovanja polglasnika in variant fonema /v/. Omenjene so tudi transkripcijske prakse in načela v nekaterih tujih sociolingvističnih raziskavah.

Ključne besede: transkribiranje, fonetična in ortografska načela, sociolingvistika, jezikovna variantnost, polglasnik, fonem /v/

Transcription in Sociolinguistic Research

Transcription is a complex, interpretative and selective process, where the conversion from spoken to written text raises a number of fundamental and practical questions. Transcription is always an interpretation, as it can never be a completely reliable representation of the original interaction; when converting text from an audio recording to writing, it is not possible to capture everything simultaneously and comprehensively. In sociolinguistic research, there are no standard transcription conventions; they always depend on the theoretical orientation of the researcher, the object and purpose of the research, i.e. what and why is being studied. Selectivity is thus closely related to subjectivity. The basic goals for transcription conventions for the study of language variation are that transcriptions should be detailed enough to retain enough data to conduct the desired linguistic analysis in an efficient way and, and simple enough to be relatively easily readable and transcribed. The researcher

therefore decides how much and what type of information to write down, what (s)he needs and what is relevant for the purpose of the current or possible later research.

The central part of the contribution focuses on experiences and various transcribing practices in the author's sociolinguistic research of variation of spoken Slovene and relevant social-psychological issues among speakers from different Slovene dialect groups – from Rovte (speakers from Idrija and the surroundings) to Lower Carinola (Ribnica) and Styria (Maribor). First, the author used precise phonetic transcriptions with the ZRCola font, developed by Peter Weiss, and later she decided for various levels of simplification and transition to orthographic transcriptions with special characters for different variants of chosen variables, which are relevant for the analysis of speech variation of individual informants in different circumstances, with different interlocutors. The variables are of course different in different dialect groups under investigation. Transcription conventions in some other related studies and projects are presented comparatively, from dialectological research to various studies of the spoken language, focusing on different research questions (e.g. studies by Jožica Škofic (Guzej), Irina Makarova, Melita Zemljak Jontes and Simona Pulko, Mojca Smolej, Darinka Verdonik), and the Corpus of spoken Slovene GOS. Special attention is paid to the problem of writing the schwa and variants of the phoneme /v/. Transcription practices and conventions from some foreign sociolinguistic studies are also mentioned.

Keywords: transcription, phonetic and orthographic conventions, sociolinguistics, language variation, schwa, phoneme /v/

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Mi i *naši, oni* i *njihovi* u politici: Osobne deikse u govorima hrvatskih saborskih zastupnika

GORANKA BLAGUS BARTOLEC
Institut za hrvatski jezik i jezikoslovlje, Zagreb, Hrvatska
gblagus@ihjj.hr

Značenje deiksa uvjetovano je konkretnim jezičnim iskazom u kojemu se ostvaruju. Kao upućivačke jedinice kojima se referira na stvarne sadržaje i okolnosti (osobe, društvo, prostor, vrijeme, diskurs) deikse su dio jezične strukture koje sudionicima komunikacijskoga događaja omogućuju da se postave prema okolnostima iskaza čiji sadržaj prenose kao govornici ili ga usvajaju kao slušatelji (Levinson 1983.). U radu će se analizirati upotreba zamjenica *mi* i *naš, oni* i *njihov* kao osobnih (personalnih) deiksa u govoru hrvatskih saborskih zastupnika prema potvrdama iz korpusa *ParlaMint-HR 2.0* (*Croatian parliament*) 2016. – 2020. dostupnom na korpusnoj platformi (No)Sketch Engine. Deiktičnost osobnih zamjenica kojima se izražava kategorija lica, odnosno upućuje na sudionike govornoga čina (Levinson 1983., Pranjković 2013.) u takvim iskazima: 1. ima primarno pragmatičku funkciju jer se s pomoću njih uspostavlja odnos *govornik/pošiljatelj poruke – iskaz – sugovornik/primatelj poruke*, 2. u okvirima kritičke analize diskursa može se promatrati kao sredstvo povezivanja jezika i društvenoga konteksta unutar kojega politički diskurs djeluje. Političar kao pojedinac u političkim govorima najčešće govori u prvom licu množine te se persuazivnost i intencionalnost njegove političke argumenatacije u načelu temelji na stavovima društvene skupine koju predstavlja ili kojoj se ideološki priklanja. Iskazi saborskih zastupnika pripadaju političkomu diskursu, ali u odnosu

na politički govor koji je monološka forma, manifestiraju se kao interaktivna forma, najčešće kao replika na prethodno rečeno ili pitano. Cilj je rada opisati sintaktička i značenjska obilježja osobnih deiksa u govoru saborskih zastupnika s obzirom na referente na koje upućuju te utvrditi u kojoj se mjeri upotreba osobnih deiksa temelji na prototipnoj slici *nas* i *njih* kao polariziranih strana (*Oni ne razumiju da potičemo zapravo one koji znaju raditi...; Štite sebe i svoje ljude* , a *mi štitimo građane RH*., *Mi za njihovog mandata nažalost nismo uspjeli ništa*...), a u kojoj su *mi/naši* i *oni/njihovi* ravnopravni, odnosno *mi* se postavlja kao subjekt koji štiti ili je na strani referenata obuhvaćenih deiksom *oni* (... *oni su naši heroji i mi prema njima imamo moralni dug*., *Na taj način bismo mogli i bolje braniti njihove želje, njihova prava, njihove zahtjeve*.).

Ključne riječi: hrvatski jezik, korpus, osobne deikse, parlamentarni govor, politički diskurs

We and *Our, They* and *Their* in Politics: Person Deixes in the Speeches of Croatian Parliamentarians

The meaning of deixes depends on the concrete speech in which they are realized. As reference units that refer to real contents and circumstances (persons, society, space, time, discourse), deixes are a part of the language structure that enable the participants of a communication event to position themselves according to the circumstances of the utterance the content of which they send as speakers or receive as recipients (Levinson 1983). The paper will analyze the use of the pronouns we and our, they and their as person deixes in the speech of Croatian parliamentarians according to the ParlaMint-HR 2.0 corpus 2016-2020 available on the corpus platform (No)Sketch Engine. The deicticity of personal pronouns that express the category of person, that is, refer to the participants of the speech act (Levinson 1983, Pranjković 2013) in such utterances: 1. has a primarily pragmatic function, as they

establish the relationship speaker/sender of the message - utterance - listener/recipient of the message, 2. within the framework of critical discourse analysis, it can be seen as a means of connecting language and the social context within which political discourse operates. In political speeches, a politician as an individual usually speaks in the first person plural, and the persuasiveness and intentionality of his/her political argumentation is, in principle, based on the views of the social group he represents or to which he/she is ideologically inclined. Utterances by parliamentarians belong to political discourse, but in relation to political speech, which is a monologue form, they manifest as an interactive form, most often as a reply to what has previously been said or asked. The aim of the paper is to describe the syntactic and semantic features of person deixes in the speeches of parliamentarians with regard to the referents they refer to, and to determine to what extent the use of person deixes is based on the prototypical image of us and them as polarized parties (1 *Oni ne razumiju da potičemo zapravo one koji znaju raditi ...;* 2 *Štite sebe i svoje ljude* , a *mi štitimo građane RH*.; 3 *Mi za njihovog mandata nažalost nismo uspjeli ništa* ...'1 They do not understand that we actually encourage those who know how to work ...; 2 They protect themselves and their people, and we protect the citizens of the Republic of Croatia; 3 Unfortunately, we did not achieve anything during their mandate ...'), and in which we/our and they/there are equal, i.e. deixis we sets itself as a subject that protects or is on the side of the referents covered by deixis they (4 ... oni su naši heroji i mi prema njima imamo moralni dug., 5 Na taj način bismo mogli i bolje braniti njihove želje, njihova prava, njihove zahtjeve. '4 ... they are our heroes and we owe them a moral debt; 5 In this way, we could better defend their wishes, their rights, their demands').

Keywords: Croatian, corpus, person deixes, parliamentary speech, political discourse

Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: Zbornik povzetkov
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Vključevanje nestandardnih vnosov v slovenske jezikovne vire z vidika jezikovnotehnoloških potreb

Jaka Čibej, Nejc Robida, Simon Krek
Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija
jaka.cibej@ff.uni-lj.si, nejc.robida@fri.uni-lj.si, simon.krek@guest.arnes.si

V zadnjem desetletju je konstanten napredek pri razvoju govornih tehnologij (npr. razpoznavalnikov in sintetizatorjev govora) vzrok za povečano potrebo po gradnji jezikovnih virov za govorjeno slovenščino, ki je bila v slovenskem korpusnem jezikoslovju v primerjavi s pisno (zlasti standardno) slovenščino deležna manj pozornosti. Na voljo so že nekateri korpusi in podatkovne baze govorjene slovenščine – npr. GOS v1.1 (Zwitter Vitez et al. 2015) in GOS-VL v4.2 (Verdonik et al. 2021) ter Artur v0.1 (Verdonik et al. 2022) in GOS v2.0 (Zwitter Vitez et al. 2023); zadnja dva sta bila izdelana v nedavno zaključenem projektu *Razvoj slovenščine v digitalnem okolju* (RSDO) –, precej manj pozornosti pa je bilo namenjene govorjeni slovenščini pri leksikonskih in leksikografskih virih, kot sta Slovenski oblikoslovni leksikon Sloleks (Čibej et al. 2022) in Digitalna slovarska baza slovenščine (Kosem et al. 2021). Sloleks, ki v različici 3.0 vsebuje iztočnice, njihove pregibne oblike in podatke o njihovih izgovorih (v mednarodnih fonetičnih abecedah IPA in SAMPA), še ni bil razširjen s podatki, ki so tipični za govorjeno slovenščino. Če želimo tovrstne pojave vključiti v leksikon, moramo proučiti, kako sistematično navajati podatke o

govorjeni slovenščini v pisni obliki tako, da so intuitivni uporabnikom in čim bolje odsevajo jezikovno rabo, hkrati pa so strojno berljivi ter neposredno uporabni za razvoj jezikovnih tehnologij.

V okviru projekta MEZZANINE se zato med drugim osredotočamo na leksikonsko obravnavo besedišča, ki se tipično pojavlja v govorjeni slovenščini, ne (oz. zelo redko) pa v pisni. Pri tem pogosto dilemo predstavlja variantnost zapisa (*šraufati/šravfati*, *cukr/cuker*, *miljavžent/miljavžnt/miljaužnt/* …), pri čemer je treba določiti, kateri zapis obravnavamo kot kanonični (in ima kot tak npr. prednost pri izpisu ob nareku razpoznavalniku govora). Pojavi se tudi vprašanje, v katerih primerih v leksikonu zadošča zgolj dodaten nestandarden izgovor pri standardni ustreznici, kdaj pa bi morali izgovor pripisati ločeni naglašeni obliki (npr. izgovor [ˈdoːbu] za deležnik na -l moškega spola pri naglašeni obliki *dobíl*; kam uvrstiti obliko rodilnika ednine *Márkota* [ˈmaːɾkɔta], ki se oblikoslovno razlikuje od standardne ustreznice *Márka* [ˈmaːɾka]). Kadar dodajamo nestandardne izgovore, moramo določiti tudi nabor fonemov, ki bo uporabljen. Preverili bomo, med katerimi variantami fonemov je treba za učinkovit razvoj jezikovnotehnoloških aplikacij razlikovati – npr. ali je za razpoznavo govora pomembno, da v leksikonu razlikujemo med [ˈmaːɾka] in [ˈmaːʁka] ali pa med [ˈgɾɔːza] in [ˈɦɾɔːza]?

Dileme bomo razrešili tako, da bomo s pomočjo analize besedišča, izluščenega iz korpusov govorjene slovenščine in korpusov nestandardne spletne slovenščine, kot je JANES v1.0 (Ljubešić et al. 2017), določili smernice, kako vključevati nestandardno gradivo v jezikovne vire in kako ga obravnavati z leksikografskega vidika. Ugotovili bomo, kateri pojavi so v govorjeni slovenščini najpogostejši, najbolj razširjeni in najbolj sistematični (npr. izpust končnega /i/ v deležniku na -l moškega spola množine, *(oni so) začéli* [zaˈʧeːl]), ter določili kriterije, s katerimi bomo v leksikonu zajeli čim širši nabor za jezikovne tehnologije relevantnih pojavov govorjenega jezika, po drugi strani pa omejili količino gradiva, da leksikon z dodajanjem nestandardnih prvin ne bo prenapihnjen in težko obvladljiv. Ob nedavnem izidu novih poglavij Pravopisa 8.0 bomo preverili tudi usklajenost novih pravil v Pravopisu 8.0 s svojo grafemsko-fonemsko pretvorbo, trenutnim zapisom IPA v Sloleksu itn. Poleg smernic bo rezultat raziskav tudi Sloleks, obogaten s podatki o govorjeni slovenščini.

Ključne besede: Sloleks, leksikon, govorjena slovenščina, nestandardno besedišče, korpusi govorjene slovenščine

Inclusion Of Non-Standard Entries In Slovene Language Resources With Regard To Language Technology Needs

In the last decade, the constant progress in the development of speech technologies (such as speech recognition and speech synthesis systems) has been the reason for an increased need for the compilation of language resources for spoken Slovene. Slovene corpus linguistics has paid relatively little attention to spoken Slovene so far, particularly if compared to written (especially standard) Slovene. Several corpora and databases of spoken Slovene have been compiled, such as GOS v1.1 (Zwitter Vitez et al. 2015) and GOS-VL v4.2 (Verdonik et al. 2021) as well as Artur v0.1 (Verdonik et al. 2022) and GOS v2.0 (Zwitter Vitez et al. 2023); the last two were compiled in the recently concluded project titled *Development of Slovene in a Digital Environment* (DSDE). However, much less focus was put on spoken Slovene in the compilation of lexica and other lexicographic resources, such as the Sloleks Morphological Lexicon of Slovene (Čibej et al. 2022) and the Digital Dictionary Database of Slovene (Kosem et al. 2021). Sloleks, which in version 3.0 contains entry words, their inflected forms and information on their pronunciations (encoded in the IPA and SAMPA international phonetic alphabets), has not been expanded with data typical of spoken Slovene. In order to include such phenomena in the lexicon, a study needs to be conducted on how to systematically include spoken Slovene data in written form to make them intuitive to users and reflective of real language use, but also machine-readable and directly applicable to language technology development.

Among other things, the MEZZANINE project focuses on the study of vocabulary that is typical of spoken Slovene, but does not occur in written form (or very rarely does). A frequent dilemma in this regard is the presence of spelling variance (*šraufati/šravfati*, *cukr/cuker*, *miljavžent/miljavžnt/miljaužnt/* …), where a canonical form needs to be determined (so that it is treated as preferred in speech recognition transcripts). Another dilemma concerns the decision on when to simply add a non-standard pronunciation to a standard equivalent and when the pronunciation should be linked to a separate accentuated form (e.g. the pronunciation [ˈdoːbu] for the masculine -l participle for the accentuated form *dobíl*; where to include the singular genitive form *Márkota* [ˈmaːɾkɔta], which is morphologically different from its standard equivalent *Márka* [ˈmaːɾka]). When adding non-standard pronunciations, a phoneme inventory needs to be determined as well. Our goal is to determine which phoneme variants should be differentiated in order to facilitate the development of language technology applications. For instance, is it important for speech recognition to differentiate between [ˈmaːɾka] and [ˈmaːʁka] or between [ˈgɾɔːza] in [ˈɦɾɔːza]?

We will resolve these questions through an analysis of vocabulary extracted from corpora of spoken Slovene and corpora of non-standard Internet Slovene, such as JANES v1.0 (Ljubešić et al. 2017), which will help us compile guidelines on how to include non-standard entries in language resources and analyze them from a lexicographic perspective. We will investigate which phenomena in spoken Slovene are the most frequent, the most widely distributed and the most systematic (e.g. the omission of end-position /i/ in the plural masculine -l participle, *(oni so) začéli* [zaˈʧeːl]), and determine the criteria to ensure that the lexicon includes a wide range of spoken language phenomena that are relevant for language technologies on one hand; and on the other hand, that the addition of non-standard elements keeps the lexicon manageable in terms of its size. With the recent publication of new chapters of the new Slovene orthography, *Pravopis 8.0*, we will also check the consistency of orthographical rules with the grapheme-to-phoneme conversion used in Sloleks, the current IPA conventions, etc. In addition to the guidelines, the study will also result in a new version of Sloleks enriched with data on spoken Slovene.

Keywords: Sloleks, lexicon, spoken Slovene, non-standard vocabulary, corpora of spoken Slovene

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Skladenjska drevesnica govorjene slovenščine: stanje in perspektive

KAJA DOBROVOLJC
Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija
kaja.dobrovoljc@ff.uni-lj.si

Skladenjsko razčlenjeni korpusi govorjenega jezika, t. i. skladenjske drevesnice, predstavljajo enega temeljnih jezikovnih virov za strojno obdelavo govorjenega jezika, denimo za razvoj skladenjskih razčlenjevalnikov ali drugih orodij za priklic specifičnih jezikovnih podatkov. Obenem so tudi nepogrešljiv gradivni vir za jezikoslovne raziskave, saj omogočajo sistematične kvalitativne in kvantitativne analize skladenjskih značilnosti govorjenega jezika ter njihovo sopostavljanje z drugimi ravnmi jezikovnega opisa, od oblikoslovja do pragmatike. V prispevku bomo predstavili drevesnico SST (angl. *Spoken Slovenian Treebank*), prvi skladenjsko razčlenjeni korpus govorjene slovenščine, orisali smernice njenih nadaljnjih izboljšav in predstavili z njo povezano raziskovalno infrastrukturo.

Drevesnica SST (Dobrovoljc in Nivre 2016), ki je edini tovrstni govorni vir v slovenskem prostoru, je bila zasnovana kot reprezentativni ročno označeni vzorec referenčnega korpusa GOS (Verdonik et al. 2013) in obsega približno 30.000 besed. Poleg podedovanih ročnih transkripcij in segmentacij izvornega korpusa so bili (standardiziranim) pojavnicam korpusa ročno pripisani še podatki o osnovni obliki, besedni vrsti in drugih oblikoslovnih lastnostih, izjave pa so bile skladenjsko razčlenjene po načelih odvisnostne slovnice, ki skladenjsko strukturo povedi člení glede na binarne asimetrične relacije (odvisnosti) med posameznimi besedami.

Konkretno drevesnica sledi označevalni shemi Universal Dependencies (de Marneffe et al. 2021), ki si prizadeva za mednarodno poenoteno slovnično označevanje na podlagi enotnega nabora oblikoslovnih in skladenjskih kategorij (oznak) ter usklajenih smernic za njihovo uporabo. Z vidika govorjenega jezika sta najpomembnejši prednosti te sheme zlasti visoka stopnja interoperabilnosti, saj omogoča neposredne konstrastivne raziskave med drevesnicami različnih jezikov ali jezikovnih zvrsti, ter celosten, enonivojski pristop k označevanju jezika, v skladu s katerim se oblikoslovne oz. skladenjske oznake pripišejo vsem izgovorjenim pojavom, brez kakršnegakoli izključevanja netekočnosti in drugih strukturnih posebnosti govora. Shema UD je bila za razčlenjevanje govorjenega jezika prvič preizkušena prav na slovenski drevesnici SST, odtlej pa je temu vzoru sledilo že več kot 15 drugih drevesnic govorjenega jezika po vsem svetu.

V okviru nacionalnega projekta *Na drevesnici temelječ pristop k raziskavam govorjene slovenščine* (ARRS Z6-4617) nameravamo ta korpus bistveno nadgraditi z več vidikov. Z vidika obsega bo korpus povečan za dodatnih 50.000 ročno razčlenjenih pojavnic, da bi s tem zagotovili trdnejšo empirično podlago za nadaljnje raziskave ter ga približali nedavno objavljeni novi različici referenčnega korpusa GOS 2.0, ki se od prve razlikuje tako z vidika sestave kot načina zapisovanja govora. Z vsebinskega vidika bo revidirana tudi prvotna različica označevalnih smernic, tako z vidika sistematičnejšega naslavljanja v slovenistični literaturi izpostavljenih skladenjskih specifik govorjene slovenščine kot z vidika poenotenja z drevesnicami govorjenega jezika v drugih jezikih (Kahane et al. 2021, Dobrovoljc 2022).

Po zgoraj orisani predstavitvi drevesnice SST bomo prispevek sklenili še s kratko predstavitvijo povezane raziskovalne infrastrukture, s pomočjo katere lahko jezikoslovci in drugi raziskovalci po tem viru tudi iščejo in ga analizirajo, kot so splošni konkordančnik noSketchEngine, označevalno orodje Q-CAT in specializirano spletno orodje Drevesnik.

Ključne besede: slovnično označevanje, skladenjsko razčlenjeni korpusi, odvisnostna skladnja, govorjeni jezik

Spoken Slovenian Treebank: Current Situation and Perspectives

Spoken language treebanks, i.e. syntactically annotated collections of transcribed speech, represent one of the fundamental language resources for spoken language processing tasks, such as syntactic parsing and information retrieval. Spoken language treebanks represent an equally valuable resource in linguistics, enabling researchers to carry out systematic qualitative and quantitative data-based investigations of spoken language syntax and its interaction with other layers of linguistic description, from morphology to pragmatics. This paper presents the Spoken Slovenian Treebank (SST), the first syntactically annotated collection of transcribed speech in Slovenian, with respect to its content, related infrastructure and perspectives for future development.

The SST Treebank (Dobrovoljc and Nivre 2016), the first language resource of this kind for Slovenian, is a representative manually annotated subset of the GOS reference corpus of Spoken Slovenian (Verdonik et al. 2013), amounting to approxicmately 30,000 words. In addition to the manual orthographic transcription and segmentation of the original GOS, each token in the SST treebank has been manually annotated with respect to lemma, part-of-speech class and morphological features. In addition, each utterance has been manually parsed following a dependency-based approach, in which syntactic structure is described as a set of binary asymmetrical relations (dependencies) that hold between words.

Specifically, the treebank employs the Universal Dependencies (UD) annotation scheme (de Marneffe et al. 2021) aimed at cross-linguistically consistent treebank annotation by providing a universal inventory of grammatical categories and guidelines for their application. From the perspective of spoken language annotation in particular, the two main advantages of the scheme include its high degree of interoperability (i.e. straightforward contrastive research across treebanks belonging to different languages or language modes, such as spoken and written) and its single-layer annotation approach, in which the morphosyntactic analysis is performed on all uttered phenomena, including disfluencies and other speech-specific structural

units. SST was the first treebank to apply the scheme to spoken data, but more than 15 other spoken UD treebanks have followed since worldwide.

In the framework of the national project *A Treebank-Driven Approach to the Study of Spoken Slovenian* (ARRS Z6-4617), we intend to significantly upgrade the SST treebank in several aspects. In terms of size, the corpus will be enlarged by additional 50,000 manually annotated tokens in order to provide a more solid empirical basis for further research and to better align it with the recently published new version of the GOS 2.0 reference corpus, which differs from the original version both in terms of genre distribution and transcription principles. From a content perspective, the original version of the annotation guidelines will also be revised, both in terms of systematically addressing syntactic characteristics of spoken Slovene, and in terms of its cross-lingual harmonization (Kahane et al. 2021, Dobrovoljc 2022).

After the SST presentation outlines above, we will also briefly present the related infrastructure that enables linguists and other researchers to analyse this language resource further, such as the noSketchEngine concordancer, the Q-CAT annotation tool, and the Drevesnik online service.

Keywords: linguistic annotation, syntactic treebanks, dependency syntax, spoken language

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)

Univerzitetna založba Univerze v Mariboru

Izazovi istraživanja multimodalnosti govornoga diskursa: aktualni pogledi iznutra

TAMARA GAZDIĆ-ALERIĆ
Sveučilište u Zagrebu, Učiteljski fakultet, Zagreb, Hrvatska
tamara.gazdic-aleric@ufzg.unizg.hr

Jezična je djelatnost bitna i jedna od najraširenijih osobina čovjekova djelovanja. Iz različitosti pojedinčeva djelovanja proizlaze i različiti tipovi diskursa, čija proučavanja jezika i stila zahtijevaju precizno definirane fenomenološke i metodološke posebnosti te predviđanje mogućih problema vezanih uz proučavani diskurs.

U radu se govori o izazovima koji proizlaze iz nedostatka resursa govornoga jezika i povezane istraživačke infrastrukture koji su trenutačno glavne prepreke u istraživanju govora i govorne komunikacije. To je uglavnom zbog činjenice da stvaranje govornih jezičnih resursa zahtijeva znatno više truda od stvaranja pisanih resursa, a dijelom i zbog činjenice da je u lingvistici pisana riječ bila, i još uvijek je, u središtu istraživačkoga interesa.

Jedna od mogućih teorijskih podjela suvremene znanosti o jeziku jest podjela na izvankontekstualnu i kontekstualnu lingvistiku. Prva se od njih, izvankontekstualna lingvistika, strukturalistički gledajući njezino polje djelovanja, bavi jezičnim sustavom, a druga, kontekstualna lingvistika, bavi jezikom ovisno o kontekstu, tj. ovisno o skupu konkretnih okolnosti u kojima se govorna djelatnost odvija.

Uzimajući u obzir kibernetički model lanca komunikacijskoga procesa, kontekst kao skup konkretnih okolnosti u kojima se odvija govorna djelatnost, može se nazvati i komunikacijom koja se ostvaruje u pojedinim komunikacijskim činovima. Usto, analiza diskursa zahtijeva razumijevanje načina na koji jezik funkcionira, i gramatike i semantike, kako bi se identificiralo što bi iskaz mogao značiti. Zatim je potrebna sposobnost procjene značenja, ili praćenje sekvenci, kako bi se konstruiralo značenje u odnosu na društveni kontekst. Zato se u radu razmatraju neka od temeljnih pitanja vezana uz istraživanje govornoga diskursa, poput pitanja čiji tekst i govor istraživati, koliko su dostupni i relevantni izvori za jezičnu analizu, uloga medija, kako opisati jezik u kontekstu, multidisciplinarnost i interdisciplinarnost analize diskursa i drugo.

Težište rada bit će na metodološkim i fenomenološkim osobinama političkoga diskursa, iako su gotovo svi problemi jedinstveni i javljaju se prilikom istraživanja svih vrsta govornih diskursa.

Ključne riječi: analiza diskursa, govorni jezik, kontekstualna lingvistika, politički diskurs.

Challenges of Researching the Multimodality of Spoken Discourse: Current Views from the Inside

Linguistic activity is essential and one of the most widespread features of human activity. Different types of discourse arise from the diversity of individual actions, the study of language and style of which requires precisely defined phenomenological and methodological peculiarities and the prediction of possible problems related to the studied discourse.

The article reflects on current challenges emerging from the lack of spoken language resources and associated research infrastructure that are currently major obstacles in speech and speech communication research. This is due mainly to the fact that the

creation of spoken language resources requires significantly more effort than the creation of the written ones, and, partly, also due to the fact that, in linguistics, the written word was, and still is, at the centre of its interest.

One of the possible theoretical divisions of contemporary language science is the division into extra-contextual and contextual linguistics. The first extra-contextual linguistics, from a structuralist view of its field of activity, deals with the language system, while the second, contextual linguistics, deals with language depending on the context, i.e. depending on the set of concrete circumstances in which speech activity takes place. Considering the cybernetic model of the communication process chain, the context as a set of concrete circumstances in which speech activity takes place can also be called communication that is realized in individual communication acts. In addition, discourse analysis requires an understanding of how language works, both grammar and semantics, in order to identify what an utterance might mean. Then it requires the ability to assess the meaning of the uptake, or to follow a sequence, in order to construct meaning in relation to social context.

That is why the paper considers some of the fundamental questions related to the research of spoken discourse, such as the question of whose text and speech to research, how accessible and relevant sources are for linguistic analysis, the role of the media, how to describe language in context, the multidisciplinary and interdisciplinary nature of discourse analysis, and more.

The focus of the article will be on methodological and phenomenological features of political discourse. Some of these problems are unique and arise when researching all spoken discourses.

Keywords: discourse analysis, spoken language, contextual linguistics, political discourse

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Splošnoslovenski besedni nabor

JANUŠKA GOSTENČNIK, JANOŠ JEŽOVNIK
ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, Ljubljana, Slovenija
januska.gostencnik@zrc-sazu.si, janoš.jezovnik@zrc-sazu.si

Prispevek predstavlja poskus izdelave kratke splošne sondažne vprašalnice za določitev pripadnosti določenega krajevnega govora nekemu (slovenskemu) narečju. Vprašalnica temelji na do sedaj izdelanih področnih vprašalnicah in gramatičnem delu Vprašalnice za Slovenski lingvistični atlas (SLA).

Slovenska dialektologija za umestitev nekega krajevnega govora uporablja t. i. anketno metodo, tj. zbiranje narečnega gradiva na terenu neposredno od krajevnih govorcev na podlagi vnaprej pripravljenega nabora reprezentativnega besedja – vprašalnice. Tradicionalno se je za ta namen uporabljala Vprašalnica za SLA (1946, 1961), zlasti njen gramatični del, ki pa se je izkazala za preobsežno in mestoma premalo natančno. Tudi zaradi tega je do danes nastalo že več t. i. področnih (leksično-fonetičnih) vprašalnic, namenjenih določanju razlikovalnih fonetičnih, morfoloških in leksičnih lastnosti posameznih krajevnih govorov in narečij, ki so upoštevale leksikalne specifike posameznega narečja. Izdelane so bile področne vprašalnice za naslednja raziskovalna območja oz. narečja: 1) Gorski Kotar: za določitev pripadnosti določenega krajevnega govora čebranškemu oz. kostelskemu narečju dolenjske narečne skupine); 2) Dubravica z okolico: za določitev pripadnosti določenega krajevnega govora kozjansko-bizeljskemu narečju štajerske narečne skupine; 3) Hum na Sutli z okolico: za določitev pripadnosti določenega krajevnega govora srednještajerskemu narečju štajerske narečne skupine; 4) Radgonski kot in Gradiščanska: za določitev pripadnosti določenega krajevnega govora

prekmurskemu narečju panonske narečne skupine; 5) zahodna in osrednja Beneška Slovenija: za določitev pripadnosti določenega krajevnega govora terskemu narečju oziroma za identifikacijo potencialnih relevantnih razlikovalnih lastnosti krajevnih govorov znotraj narečja. Prednost področnih vprašalnic je njihova prilagojenost specifikam obravnavanih narečij; to pa je obenem tudi ovira za njihovo uporabo v splošnoslovenskem kontekstu.

V prispevku predstavljena vprašalnica zajema karseda splošen, vendar jedrnat nabor besedja, v največji možni meri zastopanega v vseh ali vsaj v pretežnem delu slovenskih narečij. Ker je v slovenski dialektologiji glavni kriterij za umestitev krajevnega govora v neko narečje odrazi (historično) dolgih in kratkih naglašenih samoglasnikov, se vprašalnica osredotoča zlasti na te. Vprašalnica pripadnosti obravnavanega krajevnega govora nekemu slovenskemu narečju ne predpostavlja vnaprej, saj je njen glavni cilj enotna uporaba znotraj celotnega slovenskega jezikovnega območja.

Ključne besede: slovenščina, dialektologija, slovenska narečja, narečna vprašalnica, besedni nabor

General Slovenian Lexical Set

This paper presents an attempt to create a short general probing questionnaire for determining the belonging of a given local dialect within a particular (Slovenian) dialect. The questionnaire is based on the so far developed area-specific questionnaires and the grammatical part of the Questionnaire for the Slovenian Linguistic Atlas (SLA).

Slovenian dialectology uses the so-called survey method to place a local dialect, i.e. collecting dialect data in the field directly from local speakers on the basis of a pre-prepared set of representative lexis – a questionnaire. Traditionally, the SLA Questionnaire (1946, 1961), especially its grammatical part, was used for this purpose, but it proved to be overly extensive and at times insufficiently precise. This is one of the reasons why several so-called area-specific (lexico-phonetic)

questionnaires have been developed to date, aimed at determining the distinctive phonetic, morphological and lexical features of individual local dialects and dialects, taking into account the lexical specificities of each dialect. Area-specific questionnaires have been developed for the following research areas or dialects: 1) Gorski Kotar: to determine whether a particular local dialect belongs to the Čebranka or Kostel dialect of the Lower Carniolan dialect group; 2) Dubravica and its surroundings: to determine whether a particular local dialect belongs to the Kozjansko-Bizeljsko dialect of the Styrian dialect group; 3) Hum na Sutli and its surroundings: to determine whether a particular local dialect belongs to the Central Štajerska dialect of the Styrian dialect group; 4) Radgonski Kot and Gradiška: to determine whether a particular local dialect belongs to the Prekmurje dialect of the Pannonian dialect group; 5) Western and Central Beneška Slovenija: to determine whether a particular local dialect belongs to the Ter dialect of the Littoral dialect group and to identify potentially relevant distinctive features of local dialects within the dialect, respectively. The advantage of area-specific questionnaires is that they are tailored to the characteristics of the dialects in question; however, this is also an obstacle to their use in a general-Slovenian context.

The questionnaire presented in this paper covers as general but concise a set of vocabulary as possible, present in all or at least in the majority of Slovenian dialects. Since in Slovenian dialectology the main criterion for the placement of local dialects within a given dialect are the reflexes of (historically) long and short stressed vowels, the questionnaire focuses on these in particular. The questionnaire does not presuppose the belonging of a local dialect in question to a particular Slovenian dialect, since its main aim is uniform usage within the whole Slovenian linguistic area.

Keywords: Slovenian language, dialectology, slovenian dialects, dialect questionnaire, lexical set

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)

Univerzitetna založba Univerze v Mariboru

Stari podatki za nove začetke

KARMEN KENDA-JEŽ
ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, Ljubljana, Slovenija
carmen.kenda-jez@zrc-sazu.si

Da so lahko starejši jezikovnogeografski podatki pomembno izhodišče za proučevanje sodobne prostorske variabilnosti govorjenega jezika, je na prelomu tisočletja pokazal zlasti razvoj nemške regionalne dialektologije in njenega projekta raziskav regiolektov / regionalnih jezikov (REDE, https://www.regionalsprache.de/). Šele podrobno poznavanje jezikovnih značilnosti narečne podlage namreč omogoča ustrezno opredelitev jezikovnih variabel, ki so odvisne od drugih dejavnikov (kot so npr. družbeni status, izobrazba, spol, starost, formalni : neformalni govorni položaji itd.). Toda medtem ko je mogoče v nemškem jezikovnem prostoru izhodiščne narečne podatke vrednotiti s primerjanjem nacionalnih in regionalnih gradivskih zbirk, ki so po različnih metodoloških postopkih nastale v različnih časovnih obdobjih in vsebujejo tako transkribirano kot zvočno gradivo, je za slovenske razmere značilna fragmentarnost ali odsotnost podatkov o govorjenem jeziku, ki je do neke mere oteževala tudi načrtovanje uravnoteženosti in prostorske strukturiranosti slovenskih govornih korpusov GOS in GOS 2.0, ki sta začela nastajati v tem tisočletju.

Tudi gradivo Slovenskega lingvističnega atlasa (SLA, 1946–), ki je zaenkrat edini vir sistematično zbranih narečnih podatkov iz celotnega slovenskega jezikovnega prostora, pokritega z gosto krajevno mrežo 404 (danes 417) raziskovalnih točk, je v svoji današnji podobi primerno predvsem za raziskave narečnega besedja. Čeprav je bil SLA v skladu z izročilom francoske jezikovnogeografske smeri Julesa Gilliérona

zasnovan kot »geografsko urejen« slovar natančno fonetično popisanega besedja (Ramovš 1934), glavni namen zbiranja podatkov pa je bil zagotoviti dovolj gradiva za podroben jezikovnozgodovinski opis razvoja glasovja, je zaradi dolgotrajnega zbiranja gradiva, različno usposobljenih terenskih zapisovalcev, rabe različnih vrst fonetične transkripcije brez vmesne standardizacije ali posodabljanja posameznih zapisov to v sedanji obliki neuporabno za obdelavo fonetičnih podatkov. V dosedanjih leksičnih zvezkih atlasa so bili zato podatki objavljeni v surovi obliki, pospremljeni le s seznamom transkripcijskih znakov, skupaj z dosedanjimi opisi njihove fonetične vrednosti in osnutki primerjalnih transkripcijskih tabel (Kenda-Jež 2011, 2016, 2023).

V prispevku je opisan postopek načrtovane normalizacije in standardizacije slovenske narečne fonetične transkripcije ter njene uskladitve z mednarodno fonetično transkripcijo (IPA), ki bo potekal v okviru projekta MEZZANINE (Temeljne raziskave za razvoj govornih virov in tehnologij za slovenščino, J7-4642).

Predvidene so naslednje stopnje obravnave:

(1) Priprava gradiva SLA:
    (a) Znaki za zapis samoglasnikov: Izpis vzorčnih leksemov v preglednice samoglasniških odrazov na podlagi enajstih sklopov samoglasniških vprašanj iz gramatičnega dela vprašalnice (V700–V704; V720(a)–V720a; V721(a)–V721a; V727–V730b; V733A–V734a; V735–V737b; V742(a)–V744(c); V745(a)–747(c); V748–V750(b); V765(a)–V765(č); V765a(a)–V765(b)).
    (b) Znaki za zapis soglasnikov: Izpis vzorčnih leksemov oz. odrazov soglasnikov po gramatičnem delu vprašalnice SLA, vnos podatkov v podatkovno bazo za kartiranje (SlovarRed oz. DIAtlas), izdelava izoglosnih kart.
(2) Pregled po posameznih narečjih. Določitev arealov zapisov z visoko stopnjo variabilnosti, primerjava s sodobnimi dialektološkimi opisi in obstoječim zvočnim gradivom, kratke eksperimentalnofonetične ankete, predlogi za poenotenje.
(3) Sinteza. Izdelava skupnega predloga standardizirane transkripcije na dialektoloških delavnicah. Razrešitev odprtih vprašanj pri načelih

oblikovanja novih transkripcijskih znakov in odprava dosedanjih dogovornih tipov zapisa (npr. način zapisovanja dvoglasnikov).

Normalizirani podatki bodo omogočili prostorski prikaz razširjenosti slovenskih nestandardnih narečnih fonemov in njihovo nadaljnjo (statistično?) obdelavo.

Ključne besede: geolingvistika, slovenska narečja, govorjeni jezik, fonetika, fonetična transkripcija

Old Data for New Beginnings

The development of German regional dialectology and its project of research on regional dialects/languages (REDE, https://www.regionalsprache.de/) at the turn of the millennium has shown that older linguistic atlas data can be an important starting point for the study of contemporary spatial variability of spoken language. Only a detailed knowledge of the linguistic characteristics of the dialect base allows proper definition of linguistic variables, which are dependent on other factors (such as social status, education, gender, age, informal vs. formal context, etc.). However, while in the German linguistic area the older dialect data can be evaluated by comparing national and regional data collections, which were created at different time periods with different methodological approaches and contain both transcribed and audio material, the Slovenian situation is characterised by the fragmentary nature or absence of spoken language data, which to some extent has also made it difficult to plan the balance and spatial structuring of the Slovenian speech corpora GOS and GOS 2.0, which started to emerge in the millennium.

The dialect data collection of the Slovenian Linguistic Atlas (SLA, 1946-), which is currently the only source of systematically collected dialect data from the entire Slovenian linguistic area, covered by a dense network of 404 (now 417) research points, is in its present form suitable mainly for research of dialect vocabulary. Although the SLA was conceived as a "geographically ordered" dictionary of precisely phonetically transcribed vocabulary (Ramovš 1934), in the tradition of the French linguistic geography of Jules Gilliéron, and the main purpose of the data

collection was to provide sufficient material for a detailed analysis of historical phonetic development, it is rendered inappropriate for the processing of phonetic data in its present form due to the long-lasting collection of the material, differently trained fieldworkers, and the use of different types of phonetic transcription without intermediate standardisation or updating of individual records. In the previous lexical volumes of the Atlas, the data was therefore published in its raw form, accompanied only by a list of symbols used to transcribe it with previous descriptions of their phonetic values and drafts of comparative transcription tables (Kenda-Jež 2011, 2016, 2023).

The paper describes the process of the planned normalisation and standardisation of the Slovenian dialect phonetic transcription and its alignment with the International Phonetic Transcription (IPA), which will be carried out within the MEZZANINE project (Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian language, J7-4642).

We foresee the following stages:

(1) Preparation of the SLA material:
    (a) Vowel symbols: Extraction of sample lexemes into vowel reflexes tables based on eleven sets of vowel questions from the grammar part of the questionnaire (V700-V704; V720(a)-V720a; V721(a)-V721a; V727-V730b; V733A-V734a; V735-V737b; V742(a)-V744(c); V745(a)-747(c); V748-V750(b); V765(a)-V765(č); V765a(a)-V765(b)).
    (b) Consonant symbols: extraction of sample lexemes and consonant reflexes, respectively, according to the grammatical part of the SLA questionnaire; data entry into the mapping database (SlovarRed or DIAtlas); production of isogloss maps.
(2) Dialect-by-dialect review: Identification of areas of high variability of transcripts; comparison with contemporary dialectological descriptions and existing audio material; short experimental-phonetic surveys; suggestions for standardisation.
(3) Synthesis: Development of a common proposal for a standardised transcription through dialectological workshops. Resolving outstanding issues in the principles of the creation of new transcriptional characters and

elimination of previous conventional notation types (e.g. the manner of transcribing diphthongs).

(4) The normalised data will allow spatial representation of the distribution of Slovenian non-standard dialect phonemes and their further (statistical?) processing.

Keywords: geolinguistics, Slovenian dialects, spoken language, phonetics, phonetic transcription

Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: Zbornik povzetkov
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Metodički instrumentarij kao poticaj za govorenje u nastavi hrvatskoga jezika u primarnom obrazovanju

Martina Kolar Billege
Sveučilište u Zagrebu, Učiteljski fakultet, Zagreb, Hrvatska
martina.kolar@ufzg.hr

Slušanje i govorenje primarne su jezične (komunikacijske) djelatnosti koje je učenik, prije uključivanja u formalni školski sustav, na svom materinskom jeziku usvojio na razini koja mu omogućuje razumijevanje govora, govorenje i razgovaranje. Ta razina razvijenosti govorenoga modaliteta jezika omogućuje praćenje nastave i učenje. Govorenje se kao jezična djelatnost u nastavi može promatrati iz više aspekata. Govorenje je sredstvo poučavanja i učenja. Razvoj govorenoga modaliteta omogućuje postupni prijenos i u pisani modalitet jezika u početnom opismenjavanju.

Na kraju primarnoga obrazovanja u Republici Hrvatskoj od učenika se prema *Kurikulumu nastavnoga predmeta Hrvatski jezik za osnovne škole i gimnazije* (NN, 10/2019) očekuje da prepoznaje različite svrhe govorenja: osobnu i javnu, da primjenjuje različite govorne činove: zahtjev, ispriku, zahvalu i poziv; razgovara radi razmjene informacija; opisuje u skladu s jednostavnom strukturom; pripovijeda kronološki nižući događaje te razgovijetno govori i točno intonira rečenice.

Da bi se navedeni ishodi ostvarili te da bi se postupno stjecala komunikacijska jezična kompetencija, potrebno je uspostaviti razlikovni odnos govorenja kao postupka poučavanja i odrediti sadržaj govorenja kao sredstva učenja. Govorenje možemo promatrati kao jezičnu djelatnost u međudjelovanju i kao metodu poučavanja, ali i kao mjerljivi ishod učenja. U tom je kontekstu potrebno istražiti predviđene sadržaje poučavanja govorenja koji omogućuju dosezanje predviđenih ishoda učenja.

U radu ćemo istražiti metodički instrumentarij u udžbenicima za Hrvatski jezik te ustanoviti frekvenciju i vrstu poticaja za sadržaj govorenja u kontekstu dosezanja ishoda predviđenih kurikulom. Analizirat će se udžbenici od 1. do 4. razreda kako bi se odredilo jesu li sadržaji interpolirani u skladu s načelom postupnosti i vertikalno-spiralnoga programiranja.

Ključne riječi: metodika hrvatskoga jezika, govorenje, vrednovanje

Methodological Teaching Instrumentation As A Speaking Stimulus In Croatian Language Classes In Primary Education

Listening and speaking are the primary language (communication) activities that the student, prior to enrolling in a formal education system, has adopted in his mother tongue at a level that has allowed him/her to comprehend speech, speaking and conversation. This level of the development of the spoken modality of the language enables the monitoring of teaching and learning. Speaking as a language activity in class can be viewed from several angles. Speaking is a means of teaching and learning. In the teaching of early literacy, the development of the spoken modality facilitates a progressive transfer to the written modality of the language.

The Croatian Language Curriculum for Elementary Schools and High Schools (NN, 10/2019) states that students should be able to know when to speak for personal or public purposes, use a variety of speech acts, such as asking for something or apologizing, thanking someone or inviting them, converse in order to exchange information, describe using a simple structure, narrate events chronologically, and speak clearly and intonate sentences correctly by the end of primary education in the Republic of Croatia.

To achieve the aforementioned outcomes and gradually develop communicative language competence, it is necessary to establish a distinctive relationship of speaking as a teaching process and determining the content of speaking as a means of learning. Speaking can be seen as an interactive language activity and as a teaching method, but also as a quantifiable learning outcome. In this context, it is necessary to look into the intended contents of speaking lessons that enable the achievement of the desired learning outcomes.

In this paper, we will investigate the teaching methodology instrumentation in Croatian language textbooks and identify the frequency and type of speaking-related stimuli in the context of achieving the outcomes targeted by the curriculum. The material of textbooks from grades 1 through 4 will be examined to determine whether it has been interpolated in accordance with the principles of gradualism and vertical and spiral programming.

Keywords: evaluation, speaking, teaching methodology of the Croatian language

Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: Zbornik povzetkov
M. Krajnc Ivič (ur.)

Univerzitetna založba
Univerze v Mariboru

Анализ спонтанной устной речи как способ исследования стратификационной вариативности языковых кодов на польско-белорусском пограничье

Katarzyna Konczewska
Польской академии наук, Институт польского языка, Краков, Польша
katarzyna.konczewska@ijp.pan.pl

Спонтанная речь является реальным показателем уровня языковой компетенции говорящего, поскольку реализуется в вариабельных коммуникативных условиях. Как преобладающая в реальной речи, она является своеобразным лингвистическим феноменом и уникальным исследовательским материалом.

Лингвистические переменные (Labov 1966: 4–22) являются ключевой концепцией в социолингвистических исследованиях (Labov 1972), поскольку коррелируют с речевой осведомленностью говорящего.

В своем выступлении мы сосредоточим внимание на анализе спонтанной речи как способе исследования стратификационной вариативности языковых кодов в ареале со сложной социолингвистической ситуацией.

Данные для исследования были собраны нами в ходе полевых экспедиций, проведенных в 2015 – 2020 годах в малоисследованном полиэтническом, поликультурном, полилингвальном субареале по обе стороны польско-белорусской границы. Его особенностью является сохранившееся до наших дней деление на шляхетские (мелкодворянские) околицы и крестьянские деревни, возникшее в XVI в. Культурная самоидентичность субареала сформировалась под влиянием трех монотеистических религий: христианства, ашкеназийского иудаизма, ислама. К XVII в. данный субареал сформировался как многонациональный (литвины, русины, поляки, евреи, татары), многоконфессиональный (православие, католицизм, ислам, иудаизм) и многосословный (бояре, крестьяне, шляхта). В настоящее время в границах исследуемого субареала проживают представители православной и католической конфессии, преимущественно поляки и белорусы по национальности, являющиеся потомками крестьян и малоземельной шляхты.

Современную социолингвистическую ситуацию можно определить как экзоглоссную, несбалансированную, четырехкомпонентную (Польша/Беларусь; околица/деревня; шляхта/крестьяне; православные/католики). Важной составляющей его национальной и сословной идентичности является язык. Выбор языка в ситуации билингвизма и форма его существования в ситуации диглоссии зависят от сферы (хозяйственная деятельность, религия) и среды (шляхта/крестьяне, католическая деревня/православная) функционирования языка.

Результаты наших исследований отмечают выраженную корреляцию между современной социолингвистической ситуацией и историческими процессами освоения данного субареала, а также особенностями ее сословной составляющей.

Изучаемое нами пограничье является зоной активных языковых инфильтрационных процессов внутри различных славянских языковых подгрупп (польский, белорусский, русский). Процесс субстратного заимствования нередко проходил в условиях дву-, а иногда и трехязычия, поэтому субстратно-адстратные отношения требуют чрезвычайно осторожного подхода (Rembiszewska, Siatkowski 2018). Мы предлагаем учитывать прежде всего социолингвистические переменные, важным элементом которых является стратификационная вариативность.

На основании анализа собранного в ходе экспедиций корпуса устной монологической речи мы обратим внимание на локальность как фактор социально-когнитивной значимости (Jensen 2016); выделим выступающие в данном субареале языковые коды; сосредоточим внимание на языковых проявлениях социальной стратификации; выделим основные индикаторы и маркеры; определим лексические субстандарты и укажем на перцептивные особенности (Preston 1989) восприятия своего языка носителями отдельных кодов.

Предлагаемый нами подход является продуктивным методом изучения языковых контактов (Trudgill 1986) в пограничных ареалах со сложной исторической и современной социолингвистической ситуацией.

Ключевые слова: спонтанная речь, устная речь, языковые контакты, стратификационная вариативность, пограничье

Analysis Of Spontaneous Spoken Language As A Method To Study The Stratification Variability Of Language Codes In The Polish-Belarusian Borderland

Spontaneous speech is a true indicator of the speaker's level of linguistic competence, as it occurs under variable communicative conditions. As a prevalent feature of real speech, it is a specific linguistic phenomenon representing unique research material.

Linguistic variables (Labov 1966: 4-22) are a key concept in sociolinguistic research (Labov 1972), as they correlate with the speaker's speech awareness.

In my presentation, I will focus on the analysis of spontaneous speech as a method to investigate the stratification variability of language codes in an area with a complex sociolinguistic situation.

I collected the data for the study during field expeditions carried out in 2015–2020 in a little-studied multi-ethnic, multicultural, polylingual sub-area on both sides of the Polish-Belarusian border. Its specific characteristic is the division preserved to this day between gentry (petty nobility) settlements and peasant villages, which emerged in the 16th century. The cultural identity of the sub-area was formed under the influence of three monotheistic religions: Christianity, Ashkenazi Judaism, and Islam. By the seventeenth century the sub-area formed as multi-ethnic (Lithuanians, Ruthenians, Poles, Jews, and Tatars), multi-confessional (Orthodoxy, Catholicism, Islam, and Judaism), and multi-class (boyars, peasants, and nobility). At present, the sub-area under investigation is inhabited by representatives of Orthodox and Catholic faiths, mostly Poles and Belarusians by nationality, the descendants of peasants, and the petty nobility.

The contemporary sociolinguistic situation of this sub-area can be defined as exoglossic, unbalanced, four-component (Poland/Belarus; settlement of the petty nobility/village; petty nobility/peasantry; Orthodox/Catholic). An important component of its national and class identity is language. The choice of language in the situation of bilingualism and the form it takes in the situation of diglossia depend on the sphere in which the language functions (economic activity, religion) and the environment (petty nobility/peasantry, Catholic village/Orthodox).

The findings of my research point to a pronounced correlation between the contemporary sociolinguistic situation and the historical processes of land settlement in the given sub-area, as well as the specifics of its class structure.

The borderlands I am investigating are a zone of active linguistic infiltration processes within different Slavic language groups (Polish, Belarusian, Russian). The process of substrate borrowing often took place in the context of bi- and sometimes trilingualism, so substrate-adstrate relations require a meticulous approach (Rembiszewska & Siatkowski 2018). I suggest first of all taking into account sociolinguistic variables, an important element of which is stratification variation.

Based on the corpus of spoken monological speech collected during my field research, I will pay attention to the locality as a factor of social and cognitive significance (Jensen 2016); highlight the language codes common in this sub-area; focus on language manifestations of social stratification; distinguish the main indicators and markers; identify lexical sub-standards and point to features (Preston 1989) of the perception of their language by native speakers of individual codes.

The approach I propose is a productive method for studying linguistic contacts (Trudgill 1986) in frontier areas characterized by a complex historical and contemporary sociolinguistic situation.

Keywords: spontaneous speech, spoken language, language contacts, stratification variability, borderland

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)

Univerzitetna založba Univerze v Mariboru

Predlog izdelave korpusa humorja v govoru za slovenščino

MIRA KRAJNC IVIČ[1], ŠPELA ANTLOGA[2]
[1] Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
mira.krajnc@um.si
[2] Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor, Slovenija
s.antloga@um.si

Humor je značilno človeški in družbeni pojav. Z njim se človek odziva na različne zunanje ali notranje manj prijetne ali nepričakovane spodbude. Kot tak je humor prisoten v številnih jezikovnokomunikacijskih interakcijah, izjema so npr. cerkev, pravo, vojska. Čeprav je predmet raziskav različnih znanstvenih področij, npr. filozofije, psihologije, sociologije in književnosti, je raziskovanje humorja v slovenščini z jezikoslovnega vidika še tako rekoč neraziskano področje. Veliko oviro zlasti pri raziskovanju humorja kot spontanega odziva v konverzaciji predstavljajo težja dostopnost gradiva, neobstoj primernih zbirk humornih besedil in prepoznavanje ubesedenega kot humornega. Za humorno (besedilo) namreč velja, da je njegov perlokucijski učinek, da ubesedeno naslovnik prepozna kot humorno, pri čemer je prav prepoznavanje izrazito subjektivne narave, lahko odvisen od trenutnega razpoloženja naslovnika.

V zvezi s humorjem nasploh in v slovenščini naju bo v prvem delu prispevka zanimalo, koliko je korpusov humorja, kateri jeziki imajo na voljo korpus(-e) govorjenega jezika, ki je označen kot humoren; kako obsežni so obstoječi korpusi, katero gradivo je v njih zajeto, kateri med njimi so prosto dostopni, kako je pri

izdelavi korpusa potekalo označevanje humornih segmentov, kateri elementi govora, npr. metafora, aliteracija, večpomenskost, pregovori, intonacija, so bili označevani.

V drugem delu prispevka bova na osnovi analize obstoječega stanja predstavili zasnovo za izdelavo korpusa humorja v slovenščini. Cilj je, da je ta korpus prosto dostopen, manjšega obsega, z ročno označenimi humornimi segmenti v sicer javni deloma vnaprej pripravljeni konverzaciji. Predlagava ročno segmentiranje zbranega gradiva na izjave in shemo za označevanje (1) humornih izjav, definiranih kot enote konverzacije, ki na intonacijski, skladenjsko-semantični ravni praviloma predstavljajo najmanjšo enoto konverzacije in hkrati na naslovnika učinkujejo humorno, oziroma humornih odsekov, če več humornih izjav predstavlja vsebinsko smiselno zaključen segment konverzacije); (2) semantičnih instrumentov, kot so metafora, metonimija, repeticija, homonimija, antiteza, polisemija, hiperbola in iteracija; (3) pragmatičnih instrumentov, vezanih na govorna oz. dialoška dejanja, z upoštevanjem stopnje ujemanja med označevalci, če je označevalcev več, pri vseh navedenih sklopih označevanja.

Zasebne spontane konverzacije v avtentičnem okolju ni mogoče posneti kljub vsem ustreznim dovoljenjem, saj govorci, nevajeni snemanja, ob tem niso več spontani, (šaljiv) pogovor postane prisiljen in netekoč. Kot primernejši način raziskovanja konverzacijskega humorja se je pokazal način, ki kot temeljno gradivo zajame javno dostopne posnetke humorne vsebine v javnih komunikacijskih položajih, v katerih nastopajo govorci, vajeni snemanja, in ki imajo določene značilnosti vsakdanje, zasebne, spontane konverzacije. Zaradi navedenega bodo pri zbiranju gradiva upoštevana določena merila, npr. razmerje med spontanostjo in nespontanostjo, vrsta prenosnika, vrsta stika med nastopajočim(i) in publiko, število nastopajočih.
Pri že obstoječih korpusih se je izkazalo, da so le korpusi manjšega obsega primerni za kakovostno jezikovno, tj. pragmatično, semantično in večpredstavno, analizo humorja.

Ključne besede: korpus, govorjeno besedilo, gradivo, humor, slovenščina

Spoken Slovene Corpus Of Humor: Draft Proposal

Humor is primarily a human and social phenomenon. People use it to respond to various external or less pleasant or unexpected stimuli. As such, humor is present in many language communicative interactions, the exception being e.g. church, law, and army. Although being a research subject in various scientific fields, e.g., philosophy, psychology, sociology and literature, humorous communication is an unexplored linguistics field in Slovene. A significant obstacle, especially in researching humor as a spontaneous response in conversation, is (un)availability of suitable data, the lack of relevant collections of humorous texts, and in general, defining what is perceived as humorous. For a humorous (text), it is common that its perlocutionary effect may depend on the addressee's current mood since recognizing something as humorous is highly subjective.

Firstly, we will be interested in humor corpora in other languages, also specifically in spoken language humor corpora, their data and size, availability, annotation procedures, and annotated elements, e.g., metaphor, alliteration, polysemy, proverbs, intonation etc. Secondly, we will present a draft proposal for creating a Slovene humor corpus based on state-of-the-art analysis. The goal is a freely available small-scaled corpus with hand-annotated humorous segments in public, partly pre-prepared conversation. We propose manual segmentation into utterances and procedure for annotating (1) humorous segments as being units of conversation that usually represent the smallest units of conversation on the intonation and syntactic-semantic level and, at the same time, have a humorous effect on the addressee, or humorous sections when several humorous utterances make a meaningfully completed segment of conversation; (2) semantic instruments such as metaphor, metonymy, repetition, homonymy, antithesis, polysemy, hyperbole, and iteration; (3) pragmatic instruments related to speech/dialogue acts, considering inter-annotator agreement in all of the annotation procedures if there is more than one annotator.

Recording an authentic humorous conversation is difficult despite obtaining all necessary permissions, as the speakers are no longer spontaneous when being recorded. The humorous conversation becomes forced and disfluent. A better way to explore conversational humor is through public conversations among speakers accustomed to being recorded, i.e., humorous public performances with specific characteristics of casual conversation. Therefore we will consider additional criteria for collecting video and audio material, e.g. the relationship between spontaneity and non-spontaneity; the type of transmission; the type of contact between the performer(s) and the audience; the number of performers.

Existing humor corpora showed that small-scaled corpus data is more suitable for quality, top-notch linguistic, i.e., pragmatic, semantic and multimodal, analysis of humor.

Keywords: corpus, spoken text, data, humor, Slovene

Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: Zbornik povzetkov
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Izboljšanje jezikovne obdelave transkripcij slovenskega govora

Nikola Ljubešić, Peter Rupnik, Taja Kuzman
Institut "Jožef Stefan", Ljubljana, Slovenija
nikola.ljubesic@ijs.si, peter.rupnik@ijs.si, taja.kuzman

V prispevku predstavimo izboljševanje jezikovne obdelave transkripcij slovenskega govora v korpusu GOS (Verdonik idr., 2013). Cevovod za jezikovno obdelavo slovenščine z najboljšimi rezultati, CLASSLA-Stanza (Ljubešić in Dobrovoljc, 2019), trenutno ponuja dve možnosti obdelave jezika: obdelavo standardnega jezika in obdelavo nestandardnih spletnih besedil.

Če uporabimo obstoječa modula za oblikoskladenjsko označevanje na testnih podatkih, ki temeljijo na korpusu GOS, dosežemo s standardnim modelom 76-odstotno klasifikacijsko točnost, z nestandardnim modelom pa 78-odstotno točnost. Dejstvo, da nestandardni model deluje bolje na tej vrsti besedil, že nakazuje na to, da so transkripcije govorjenega jezika bolj podobne jeziku nestandardnih spletnih besedil kot pa standardnemu jeziku. Ne glede na razliko med modeloma oba dosegata na splošno zelo nizke rezultate, saj medtem ko pri tej ravni označevanja navadno dosegamo vsaj 90-odstotno točnost, pri testni množici iz korpusa GOS modela napačno označita četrtino pojavnic. Nato smo naučili nove modele na učnih podatkih iz iste domene, natančneje iz nabora podatkov SST (Dobrovoljc in Nivre, 2016), ki je vzorčen iz korpusa GOS. Nemudoma – in tudi pričakovano – smo dosegli 86-odstotno točnost. Nato smo učne podatke iz nabora SST združili z drugimi učnimi korpusi za slovenščino, kot sta ssj500k (Dobrovoljc idr., 2017) in Janes-Tag (Fišer idr., 2018), in se še dodatno povzpeli na zelo sprejemljivo 92-

odstotno točnost, kar pomeni, da je model nepravilno označil manj kot desetino pojavnic.

Na podoben način smo raziskali izboljšanje lematizacije in odvisnostnega razčlenjevanja prek uporabe podatkov iz iste domene (SST) in izven domene (ssj500k in Janes-Tag) ter dosegli podobne rezultate. Pri lematizaciji ni opaziti tolikšnega izboljšanja med standardnim modelom (97-odstotna točnost) in na novo naučenim, izboljšanim modelom (98-odstotna točnost). Eden od razlogov za to je, da so ciljne leme v standardnih besedilih in transkripcijah govora enake. Za razliko od lematizacije pa z novimi modeli občutno izboljšamo odvisnostno razčlenjevanje. Medtem ko standardni model doseže 56-odstotno označeno povezanost, kar pomeni, da skoraj polovico besed napačno označi, smo z učenjem modela izključno na podatkih iz iste domene (zbirka SST) izboljšali rezultat na 66 odstotkov. Kot pri spodnjih dveh ravneh označevanja smo tudi pri tej ravni nadaljevali z dodajanjem drugih učnih korpusov za slovenščino in izboljšali označeno povezanost na 72 odstotkov, kar je dober rezultat za tako zahtevno in premalo raziskano nalogo, kot je odvisnostno razčlenjevanje transkripcij govora.

Prispevek zaključimo s predstavitvijo nadaljnjih načrtov za izboljšanje jezikovne obdelave transkripcij slovenskega govora. V nadaljevanju pričakujemo dvakratno povečanje ročno označenih učnih podatkov, kar bo nedvomno še nadaljnje izboljšalo rezultate obdelave. Poleg tega nameravamo raziskati, ali si lahko pri jezikovni obdelavi pomagamo tudi z govorom. Iz govora bomo pridobili uporabne informacije tako, da ga bomo prikazali prek predstavitev, pridobljenih s transformerji za govor, kot je XLS-R (Babu idr., 2021). Poleg tega pa nameravamo obogatiti jezikovno označevanje tudi s podatki o netekočnosti. S temi podatki ne bomo samo ponudili dodatnih informacij o jeziku, ampak bo to morda pozitivno vplivalo tudi na celotno jezikovno obdelavo.

Ključne besede: jezikovna obdelava, transkripcije govorjenega jezika, oblikoskladenjsko označevanje, lematizacija, odvisnostno razčlenjevanje

Improving Linguistic Processing of Slovenian Speech Transcripts

We present the ongoing efforts in improving the linguistic processing the transcripts of Slovenian speech in the GOS corpus (Verdonik et al. 2013). The state-of-the-art linguistic processing pipeline for Slovenian, CLASSLA-Stanza (Ljubešić and Dobrovoljc, 2019), currently supports a mode for processing standard language, and another mode for processing non-standard written communication as it is found on the Internet.

When applying the existing modules for part-of-speech tagging to our GOS-based test set, we achieve accuracy of 76 percent with the standard model and 78 percent with the non-standard model. The non-standard model performing better already signals that there is higher similarity of spoken language transcripts to non-standard written Internet language than it is to standard language. Regardless of the difference, the results can be considered very low overall, with one out of four tokens being incorrectly annotated on the annotation layer where we are used to observe 90+ percent accuracy. Once we build new models on the in-domain training data from the SST dataset (Dobrovoljc and Nivre, 2016), which is a sample of the GOS corpus, we directly, and very much expectedly, climb to an accuracy of 86 percent. Combining the SST training data with other Slovenian training data like ssj500k (Dobrovoljc et al. 2017) and Janes-Tag (Fišer et al. 2018), we further climb to a very acceptable accuracy of 92 percent, now less than one in ten tokens being incorrectly annotated.

In a similar manner, we investigate the room for improvement on the lemmatisation and the syntactic dependency parsing levels through a mixture of in-domain (SST) and out-of-domain data (ssj500k and Janes-Tag), with somewhat similar results. For lemmatization there is not as much of improvement to be observed between the standard model (97 percent accuracy) and the newly trained, improved model (98 percent accuracy), inter alia, due to the lemmatisation process aiming at the same lemma targets in standard texts as is in transcripts of spoken data. On the other hand, the syntactic dependency parsing is drastically improved through the newly

trained models. While the standard model achieves labeled attachment score of 56 percent accuracy, making almost every other word badly parsed, training a model only on the SST in-domain training data improves the parsing accuracy to 66 percent. Following the same recipe as for the two lower annotation layers, adding other available Slovenian training data, improves labeled attachment score accuracy to 72 percent, which is a reasonable accuracy for such a hard and rather underresearched task as dependency parsing of spoken data transcripts is.

We close our contribution with goals for further improvement in linguistic annotation of spoken transcripts of Slovenian. The manually annotated training data are expected to double in size in the following period, which will surely bring additional improvements to our results. Furthermore, we plan on exploring the usefulness of the spoken signal on the task at hand. We will exploit the spoken signal through representations obtained from speech transformer models such as XLS-R (Babu et al. 2021). Finally, we plan to enrich the linguistic annotation with information on disfluencies, which might, beside being informative by itself, have a positive impact on the overall linguistic processing as well.

Keywords: linguistic processing, transcripts of spoken language, part-of-speech tagging, lemmatisation, syntactic dependency parsing

Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: Zbornik povzetkov
M. Krajnc Ivič (ur.)

Univerzitetna založba Univerze v Mariboru

Fonološka zmožnost bosansko govorečih priseljenk in priseljencev

Jana Lovrec Srša, Gjoko Nikolovski
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
janalovrecsrsa@gmail.com, gjoko.nikolovski@um.si

Od vseh jezikovnih ravnin, ki se obravnavajo pri poučevanju določenega jezika kot drugega in tujega jezika, je najmanj zastopana fonetično-fonološka raven. Razlog za to je dejstvo, da je sporazumevanje možno tudi takrat, ko izgovorjava ni povsem pravilna. Uporaba komunikacijskega pristopa pri poučevanju tujih jezikov tudi vpliva na »zanemarjanje« fonetično-fonološke ravni, saj se tako učenci kot tudi učitelji zavedajo, da je sporazumevanje kljub nepravilni/neustrezni izgovorjavi možno.

Prispevek obravnava fonološko zmožnost neslovensko govorečih priseljenk in priseljencev iz Bosne in Hercegovine, ki živijo v Mariboru in se za uspešno integracijo v slovensko okolje učijo slovenščino kot drugi in tuji jezik (SDTJ). Za potrebe prispevka je analiziran korpus njihovih govorjenih besedil, v katerih so obravnavane in definirane glasoslovne težave, s katerimi se soočajo pri učenju SDTJ. Analiza temelji na posnetkih njihovih govorjenih besedil ter registrira sledeče težave: izgovor polglasnika (*danes*), izgovor vzglasnega *v-* pred (ne)zvenečim soglasnikom (*včasih*), izgovor izglasnega *-v* v položaju za samoglasnikom ali *r* (*domov, vrv*), izgovor predloga *v*, izgovor morfemskega *-ol-* za nekdanji zvočniški glas v položaju pred soglasnikom (*jabolko*), izgovor izglasnega *-l* v položaju za samoglasnikom (*je delal*),

izgovor *-l-* v položaju za samoglasnikom in pred soglasnikom (*gledalci*), težave z naglasnim mestom, težave s kakovostjo samoglasnikov idr.

Ključne besede: glasoslovne težave, izgovor, slovenščina kot drugi in tuji jezik, bosanske priseljenke in priseljenci, jezikovne interference, jezikovna integracija

Phonological Competence of Bosnian-speaking Immigrants

Of all the linguistic levels addressed in the teaching of one language as a second and foreign language, the phonetic-phonological level is the least represented. This is because communication is possible even when the pronunciation is not completely correct. The use of a communicative approach in foreign language teaching also affects the "neglect" of the phonetic-phonological level, as both students and teachers are aware that communication is possible despite incorrect/inadequate pronunciation.

The paper deals with the phonological ability of non-Slovenian-speaking immigrants from Bosnia and Herzegovina who live in Maribor and learn Slovene as a second and foreign language (SDTJ) to successfully integrate into the Slovenian environment. For this paper, a corpus of their spoken texts is analysed, in which the phonetic difficulties they face in learning SDTJ are discussed and defined. The analysis is based on recordings of their spoken texts and registers the following problems: pronunciation of the semivowel (*danes*), pronunciation of the *v-* before the (un)voiced consonant (*včasih*), pronunciation of the final *-v* after the vowel or *r* (*domov, vrv*), pronunciation of the preposition *v*, pronunciation of the morphemic *-ol-* for the formerly voiced consonant in front of the consonant (*jabolko*), pronunciation of the final *-l* after the vowel (*je delal*), pronunciation of *-l-* in the position after the vowel and before the consonant (*gledalci*), accent placement problems, vowel quality problems, etc.

Keywords: phonological problems, pronunciation, Slovene as a second and foreign language, Bosnian immigrants, language interference, language integration

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Govor u filmu kao predložak jezičnostilske analize – mogućnosti automatske transkripcije govornih vrednota

IVA NAZALEVIĆ ČUČEVIĆ, DAVOR NIKOLIĆ
Sveučilište u Zagrebu, Filozofski fakultet, Zagreb, Hrvatska
inazalev@ffzg.hr, dnikoli@ffzg.hr

U izlaganju će biti riječi o govoru u filmu kao predlošku jezičnostilske analize, pri čemu će se posebna pozornost posvetiti razmatranju mogućnosti doprinosa računalnih programa za automatsku transkripciju filma te analizi rezultata automatske transkripcije govornih vrednota filma *H-8*... Nikole Tanhofera (1958). Riječ je o filmu koji je 2020. proglašen najboljim hrvatskim igranim filmom svih vremena (v. Hina 24. XII. 2020, Pavičić 2020, Čegir 2020). Tematski i žanrovski intrigira filmsku povijest i kritiku i novijega vremena (npr. Škrabalo 1998, Peterlić 2005, Šakić 2007a, 2007b, 2007c, Gilić 2008, Pavičić 2017, Pavičić 2020, Čegir 2020), a predmet je i jezikoslovnih proučavanja (npr. Nazalević Čučević 2021). U okviru filmskih 1950-ih svrstava se u filmove suvremene teme, svakodnevice i realizma. Likovi sudjeluju u svakodnevnome događaju – putovanju. Realističnost se svakodnevnoga događaja pojačava prisutnošću Naratora, sveznajućega nad-lika (Peterlić 2005). U okviru prethodnih jezikoslovnih analiza njegovu se govoru pristupilo kao onome koji pripada informativnome žanru novinarsko-

publicističkoga stila dvočlane strukture – prvi pripada izvještaju o prometnoj nesreći te je stilski blizak govoru filmskoga žurnala, a drugi komentaru. Da bi se otkrilo na koji način i s kojim ciljem stilski aktiviraju jezik predmetnih podžanrova, analizirale su se njegove jezične i stilske osobitosti, među kojima i enalaga u konstrukciji rečenica. Pretpostavilo da će u govoru izvještaja, objektivnoga žanra, prevladavati iskazi sa strukturom jednostavne rečenice, dok će u govoru komentara, subjektivnoga oblika, prevladati složene strukture, i to hipotaksa. Pretpostavke su se ispitivale na temelju tekstualnoga zapisa nastala slušanjem i bilježenjem. Utvrdilo se da je u govoru izvještaja jednaka zastupljenost jednostavnih i složenih konstrukcija, pri čemu hipotaktičke strukture trostruko više dominiraju nad parataktičkima. Gotovo ravnomjerna zastupljenost jednostavnih i složenih struktura utvrdila se i u govoru komentara, u kojemu je hipotaktičkih struktura četiri puta više od parataktičkih (usp. Nazalević Čučević 2021). Za potrebe ovoga izlaganja iste će se pripovjedačke dionice analizirati na temelju računalnoga programa za automatsku transkripciju teksta. Zadanim parametrima, u prvome redu onima koji se odnose na stanku, usporedit će se rezultati prve (slušanje i zapisivanje) i druge metode (računalni program). Uz to što će biti zanimljivo promotriti što je u odnosu na ljudsko uho utvrdilo računalo, cilj je rada predočiti rezultate koji bi mogli biti korisni proučavateljima govora na filmu jer bi se uputilo na mogućnosti alata pri analizi stilskih vrednota te bi se predočili konkretni rezultati za konkretne korisnike, npr. osobe oštećena sluha.

Ključne riječi: govor u filmu, jezičnostilska analiza, automatska transkripcija, govorne vrednote, stanka.

Speech In Film As A Subject Of Linguistic And Stylistic Analysis – Possibilities Of Automatic Transcription Of Spoken Language Features

In this research the speech in the film will be submitted to linguistic and stylistic analysis, where special attention will be paid to considering the possibility of the contribution of computer programs for the automatic transcription and to the analysis of the results of the automatic transcription of the spoken language features of the Nikola Tanhofer's film *H-8...* (1958). It is a film that was declared the best Croatian movie of all time in 2020 (see Hina 24. XII. 2020, Pavičić 2020, Čegir 2020). Thematically and genre-wise, it continues to intrigue film historians and critics (e.g. Škrabalo 1998, Peterlić 2005, Šakić 2007a, 2007b, 2007c, Gilić 2008, Pavičić 2017, Pavičić 2020, Čegir 2020), as well as linguists (e.g. Nazalevic Čučević 2021). Within the framework of the 1950s, it is classified as a movie with contemporary themes, a movie of everyday life and realism. The characters take part in an everyday event - a journey. The realism of everyday events is enhanced by the presence of the Narrator, an omniscient super-character (Peterlić 2005). In the framework of previous linguistic analyses, the Narrator's speech was approached as one that belongs to the informative genre of journalistic style with a two-part structure - the first belongs to a report on a traffic accident and is stylistically close to the speech of a film newsreel, and the second to a commentary. In order to find out in what way and with what goal they stylistically activate the language of the subject subgenres, its linguistic and stylistic peculiarities were analysed, including enallagies in the construction of sentences. It was assumed that in the speech of the reports (objective genre) statements with the structure of a simple sentence will prevail, while in the speech of commentary (subjective genre) complex structures will prevail, namely hypotaxis. Assumptions were examined on the basis of the text record created by listening and writing. It was found that in the speech of the report there is an equal representation

of simple and complex constructions, whereby hypotactic structures dominate three times more than paratactic ones. An almost equal representation of simple and complex structures was also found in the speech of commentary, in which there are four times more hypotactic structures than paratactic ones (cf. Nazalević Čučević 2021). For the purposes of this presentation, the same narrative sections will be analysed on the basis of a computer program for automatic text transcription. The results of the first method (listening and writing) and the second method (computer assisted transcription) will be compared with the given parameters, primarily those related to the pause. In addition to the fact that it will be interesting to observe what the computer has determined in relation to the human ear, the aim of the paper is to present results that could be useful to researchers of speech in film, as it would refer to the possibilities of the tool in the analysis of the spoken language features, and would present specific results for specific users, e.g. persons with impaired hearing.

Keywords: speech in film, linguistic and stylistic analysis, automatic transcription, spoken language features, pause

Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: Zbornik povzetkov
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Jezikovni modeli v jezikoslovni analizi

Teodor Petrič
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
teodor.petrici@um.si

V prispevku preučujemo možnosti za uporabo nekaj izmed novejših razpoložljivih jezikovnih modelov in sorodnih tehnoloških pripomočkov za pretvorbo govornega vira v pisno obliko, pripravo lastnega večplastnega jezikovnega gradiva in pretežno nenadzorovano jezikoslovno analizo gradiva. Izhajamo iz perspektive raziskovalca, ki ima malo ali nekaj programerskega znanja in dela na zmogljivejšem prenosnem računalniku. Pri preučevanju možnosti se nam poraja več vprašanj: npr. katere materialne potrebe so povezane z uporabo velikih jezikovnih modelov, koliko programerskega znanja potrebuje raziskovalec za prilagajanje modela lastnim potrebam, kolikšen prihranek časa prinaša uporaba jezikovnih tehnologij v primerjavi z lastnoročnim in lastnoumskim delom, katera opravila lahko raziskovalec uspešneje rešuje s tehnološkimi pripomočki, kako lahko raziskovalec objavi gradivo oz. prispeva svoje gradivo za dopolnjevanje velikega govornega korpusa? Da bi našli zadovoljive odgovore na gornja vprašanja, smo preizkušali več vrst že naučenih (angl. pretrained) jezikovnih modelov, ki so razpoložljivi na spletnih portalih slovenscina.eu, clarin.si, huggingface.co in github.com. Za reševanje nalog smo si pomagali s programerskimi forumi, ChatGPT in z BingChat. Naloge, ki smo jih zastavljali orodjem, so samodejni prepis govorjenega besedila (dialoga in monologa) v različne pisne oblike (npr. v obliki podnapisov ali kot tabelo) za izvoz v druga specializirana programska orodja za klasifikacijo jezikovnih znakov (npr. oblikoskladenjsko analizo, prepoznavanje sentimenta in čustvenosti govora) ali za

druge naloge (npr. povzemanje besedila). Izbranih je bilo več vrst jezikovnih vzorcev: dnevno informativna oddaja, športni prenos, pogovorna oddaja, dvo- ali večjezična oddaja in otroški govor. V prispevku primerjamo pridobljene rezultate z izidi uporabe ustreznih jezikovnih modelov za angleščino ali nemščino.

Ključne besede: jezikovni modeli, jezikovno gradivo, ustno sporazumevanje, klasifikacija besedilnih prvin, transkripcija

Language Models In Linguistic Analysis

The paper explores the possibilities of using some of the more recently available linguistic models and related technological tools to convert spoken sources into a written form, produce one's multi-layered linguistic material, and the largely unsupervised linguistic analysis of the material. We start from the perspective of a researcher who has little or some programming skills and works on a powerful laptop. In considering the possibilities, several questions arise: e.g. what material needs are associated with the use of large language models, how much programming knowledge does the researcher need to adapt the model to their own needs, how much time is saved by using linguistic technologies compared to handwritten and self-taught work, which tasks can the researcher solve more successfully with technological aids, how can the researcher publish the material, or how can the researcher contribute their material to complement a large speech corpus? To answer the above questions satisfactorily, we tested several types of pre-trained language models available on slovenscina.eu, clarin.si, huggingface.co, and github.com. To solve the tasks, we used programming forums, ChatGPT and BingChat. The tasks we asked the tools to do are to automatically transcribe spoken text (dialogue and monologue) into different written forms (e.g. as subtitles or as a table) for export to other specialized token classification tools (e.g. morphosyntactic analysis, sentiment, and emotional speech recognition). Several types of language samples were selected: a daily news program, a sports broadcast, a talk show, a bilingual or multilingual

program, and children's speech. This paper compares the results obtained with the outcomes of using the corresponding language models for English or German.

Keywords: language models, corpus, oral communication, token classification, transcription

Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: Zbornik povzetkov
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Poslušati med vrsticami: parlamentarni govor in njegovi zapisi

Ina Poteko,[1] Marko Stabej,[1] Kaja Jošt[2]
[1]Univerza v Ljubljani, Filozofska fakulteta, Ljubljana, Slovenija
ina.poteko@ff.uni-lj.si, marko.stabej@ff.uni-lj.si
[2] Državni zbor Republike Slovenije, Ljubljana, Slovenija
kaja.jost@dz-rs.si

Parlamentarni govor je eden redkih govornih virov v slovenskem prostoru, ki se stalno dopolnjuje in je prosto dostopen javnosti ter posledično tudi raziskovalcem in raziskovalkam, pri čemer pa se je treba zavedati problematike zvrstnosti tovrstnega govorjenega diskurza – govor v parlamentarnih razpravah je večinoma bran ali govorjen na podlagi predhodne priprave, saj poteka v okviru dnevnega reda in pravil poslovnika, pri čemer pa povsem spontano oglašanje ni mogoče.

Poleg zvočnih oziroma video posnetkov so na voljo tudi zapisi sej, za katere v Državnem zboru Republike Slovenije skrbita Oddelek operaterski servis, ki »opravlja naloge pisanja, urejanja in objave dobesednih zapisov sej Državnega zbora in njegovih delovnih teles ter drugih dogodkov« (*Službe Državnega zbora*), in Dokumentacijsko-knjižnični oddelek, ki med drugim »opravlja redakcijo in ureja zapise sej Državnega zbora, lektorira zahtevnejša besedila Državnega zbora ter svetuje na jezikovnem področju« (*Službe Državnega zbora*). Proces zapisovanja sej je v celoti digitaliziran šele od leta 2010, na spletni strani pa je neposredno po seji najprej

objavljen neverificiran zapis, ki ga pozneje nadomesti urejena in verificirana (uradna) verzija.

Na podlagi zapisov parlamentarnega govora so nastali tudi korpusa ParlaMint in siParl ter orodje Parlameter. Vendar pa se pri zapisih poraja vprašanje, kakšno gradivo sploh imamo pred sabo. Javnosti sta dostopna le podatka iz Poslovnika DZ, da se »[o] delu na seji državnega zbora vodijo dobesedni zapisi (zvočni zapis in njegov prepis)« ter iz publikacij sejnih zapisov državnega zbora, ki v uvodu navajajo sledeče:

> Simultano ob zvočnem zajemanju nastaja besedilo, ki je na spletu dostopno s približno polurnim zamikom. V uredništvu sejnih zapisov se ob poslušanju zvočnega posnetka preveri avtentičnost zapisanega, besedilo pa se uredi v skladu s strokovnimi merili prenosa govorjene besede v zapisano. Takšno preverjeno in jezikovno urejeno besedilo na spletnem naslovu zamenja prvi prepis. (*Sejni zapisi Državnega zbora*)

Iz informacij, ki so javnosti prosto dostopne na spletu, tako ni mogoče natančno ugotoviti, kako zapisi nastajajo, v kolikšni meri so prekrivni z izvorno povedanim in do katerih sprememb pride v procesu. Za splošno javnost je dikcija iz poslovnika, da se vodijo dobesedni zapisi, zavajajoča, saj to v večji meri velja le za verzijo, ki je objavljena neposredno po seji, ne pa tudi za verificirano (uradno) objavo sejnih zapisov.

V prispevku na podlagi delovnih izkušenj iz obeh oddelkov, ki v državnem zboru skrbita za zapis seje, analize zapisov ter internega gradiva državnega zbora ugotavljamo, kakšne so zapisovalne in uredniške prakse zapisovanja sej v Državnem zboru Republike Slovenije v obdobju 2010–2022. Primerjalno predstavljamo tudi zapisovalne prakse drugih parlamentov (na primer Anglije in Finske). Prispevek se poleg tega ukvarja tudi s problematiko prenosa govorjene besede v pisno obliko ter prevprašuje posamezne uredniške in lektorske posege v zapisih parlamentarnega govora.

Ključne besede: parlamentarni govor, Državni zbor Republike Slovenije, sejni zapisi, govorni vir, prenos govora v pisno obliko

Listening Between The Lines: The Parliamentary Speech And Its Transcripts

The parliamentary speech is one of the few sources of speech in Slovenia that is constantly updated and freely available to the public and thus to researchers. However, it is important to bear in mind the problematic nature of this type of spoken discourse – speeches in parliamentary debates are mostly read or spoken on the basis of prior preparation, as they take place within the framework of the agenda and the rules of procedure, and it is not possible for them to be completely spontaneous.

In addition to audio and video recordings, the National Assembly of the Republic of Slovenia also provides transcripts of its sessions, which are managed by the Operator Service, which is "responsible for writing, editing and processing the verbatim records of sessions of the National Assembly and meetings of its working bodies and other events" (Services of the National Assembly), and by the Documentation and Library Section, which, among other things, "provides language editing of the verbatim records of National Assembly sessions and official correspondence, and offers linguistic advice" (Services of the National Assembly). The process of transcription has only been fully digitised since 2010, with an unedited transcript being published on the website immediately after the session, which is later replaced by an edited and verified (official) version.

The ParlaMint and siParl corpora and the Parlameter tool were also created from the published transcripts of the parliamentary speeches. However, the transcripts raise the question of what kind of material we have in front of us in the first place. The only information available to the public comes from the Rules of Procedure of the National Assembly, which state that "verbatim records (audio recordings and transcription) shall be kept of the proceedings of the National Assembly", and from the publications of the National Assembly's sessions, which state in their introduction as follows:

> Simultaneously with the audio capture, the text is produced and made available online with a delay of about half an hour. The editorial office checks the transcript for authenticity by listening to the audio recordings, and the text is edited according to professional criteria for the transfer of the spoken word into the written word. This verified and linguistically edited text replaces the first transcript on the web address. (*Sejni zapisi Državnega zbora,* translated by I. P.)

The information that is freely available to the public on the internet makes it impossible to know exactly how the transcripts are created, to what extent they overlap with what was originally said, and what changes are made in the process. For the general public, the wording in the Rules of Procedure that verbatim records are to be kept is misleading, as this largely applies only to the version that is published immediately after the session, and not to the verified (official) publication of the session transcripts.

In this paper, we present our experience from the two departments responsible for the transcripts of sessions in the National Assembly, as well as the analysis of the transcripts and the internal materials of the National Assembly, with which we identify the transcribing and editing practices of the transcriptions of sessions in the National Assembly of the Republic of Slovenia in the period 2010–2022. Practices of other parliaments (e.g., England and Finland) are also presented. In addition, the paper also addresses the issue of transferring the spoken word into written form and examines individual editorial and proofreading interventions in the transcripts of parliamentary speech.

Keywords: parliamentary speech, National Assembly of the Republic of Slovenia, parliamentary sessions, speech source, transfer of speech into written form

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Govor in govorna komunikacija v učnih načrtih za osnovno šolo in gimnazijo ter v katalogih znanj

SIMONA PULKO, MELITA ZEMLJAK JONTES
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
simona.pulko@um.si, melita.zemljak@um.si

Govor je človekova dejavnost, ki je pomembna predvsem zaradi svoje sporočevalne oz. komunikacijske funkcije jezika. Deluje na izrazni ravni s pomočjo slušnega prenosnika, zato je izrednega pomena, da se otrok od rojstva uri v komunikaciji, pri tem pa mu okolica nudi govorno podporo in zgled. Govor tako v razvoju posameznika opravlja individualno vlogo, pri komunikaciji z okolico pa družbeno vlogo. Eden temeljnih namenov jezikovnega pouka je razvijanje sporazumevalne zmožnosti v slovenskem (knjižnem) jeziku, tj. praktično in ustvarjalno obvladovanje vseh sporazumevalnih dejavnosti (tudi govora) in jezikovnosistemskih osnov (*Učni načrt za slovenščino* 2018).

Prispevek bo predstavil analizo poučevanja in učenja govora (splošnega govornega izobraževanja) pri predmetu slovenščina. Razvijanje sporazumevalne (tudi govorne) zmožnosti zasledimo kot enega od splošnih ciljev v vseh učnih načrtih na vseh stopnjah izobraževanja v učnih načrtih za pouk slovenščine. Primerjalno bosta analizirana učni načrt za slovenščino za osnovno šolo iz leta 2011 in učni načrt za slovenščino za osnovno šolo iz leta 2018; primerjalno analizirani bodo tudi učni načrt za slovenščino v gimnazijah (splošnih, klasičnih in strokovnih) iz leta 2008 in katalogi znanj za slovenščino. V analizi bodo predstavljeni pregled, prisotnost

poučevanja govora in strategije ter pristopi za poučevanje govora v različnih učnih načrtih; ovrednoteni bosta doslednost in kontinuiteta pri razvijanju govora ter posledično govorne zmožnosti.

Predmet raziskovanja bosta kvalitativna in kvantitativna prisotnost govora in razvijanja govora ter sporazumevalne zmožnosti kot temeljnih ciljev v vseh izbranih učnih načrtih in katalogih znanj, kolikšen je poudarek na razvijanju pravorečne zmožnosti in na razvijanju zapisovalne zmožnosti govora v različnih načinih transkribiranja, upoštevajoč ciljno skupino, tj. učeče se v osnovni šoli oz. v različnih vrstah srednjih šol.

Ključne besede: poučevanje govora, učni načrt, slovenščina, osnovna šola, gimnazija, katalog znanja, transkribiranje govora

Speech and Speech Communication in Curricula for Elementary School, Grammar Schools and in Catalogs of Knowledge

Speech is human activity important mainly for its communicative language function. It operates on the expressive level with the help of an acoustic medium, thus it is important for a child to excercise communication from birth, while the environment provides him/her verbal support and example. Speech plays an individual role in one's development, and in communication with the environment a social role. One of the fundamental purposes of learning language at school is the development of communication skills in Slovene (literary) language, i.e. practical and creative mastery of all communication activities (including speaking) and language system basics (*Učni načrt za slovenščino* 2018).

The paper will present an analysis of teaching and learning the speech (general speech education) at lessons of Slovene language. Developing communication ability (including speech ability) is one of the general goals of all the curricula at every level of education in the curricula for Slovene language. Comparative curricula analysis includes the curriculum for Slovene in elementary school (2011), the curriculum for Slovene in elementary school (2018); the curriculum for Slovene in grammar schools (general, classical and professional) from 2008 and catalogs of knowledge for Slovene language. The analysis will present an overview, the presence of speech, teaching strategies and approaches to teaching speech in different curricula; consistency and continuity in developing speech and consequently the speaking ability will be evaluated.

The research will focus on the qualitative and quantitative presence of speech, speech development and communication ability as fundamental goals in all the selected curricula and catalogs of knowledge, the amount of emphasis on standard pronunciation and development of orthographic ability through different types of transcribing, taking into account the target group, i.e. pupils in primary school or in different types of grammar schools.

Keywords: teaching the speech, curriculum, Slovene, primary school, grammar school, catalog of knowledge, speech transcribing

Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: Zbornik povzetkov
M. Krajnc Ivič (ur.)

Univerzitetna založba Univerze v Mariboru

Komentarji novic Regionalobala.si med govorjenim in pisnim diskurzom

Maša Rolih
Univerza na Primorskem, Znanstveno-raziskovalno središče, Koper, Slovenija
masa.rolih@gmail.com

Prispevek obravnava prvine govorjenega jezika v komentarjih novic Regionalobala.si. Zanima nas, v kolikšni meri so komentarji zapisani na način spontanega govora in katere značilnosti govorjenega diskurza kažejo. Sredstva govorjenega diskurza bodo analizirana s pomočjo v zadnjem mesecu objavljenih zapisanih komentarjev novic portala Regionalobala.si.

Značilnosti zapisov komentarjev spletnih novic lahko razložimo primerjalno s spontanim govorom oz. spontanim govorjenim diskurzom, v zvezi z zapisi v preučevanih okolij uporabljamo predvsem pojme spletni diskurz, spontani diskurz, spletna komunikacija in spontana komunikacija. Spletni diskurz se uporablja kot realizacija spletne komunikacije, spontani diskurz pa je realizacija spontane komunikacije in je ključna značilnost spletnega diskurza. Za zapise v okviru spletnih komunikacijskih okolij ne moremo uporabiti izraza spontani govor oz. spontani govorjeni diskurz, saj gre v bistvu za zapise, primerjalno s spontanim govorom pa jih razlagamo zato, ker vsebujejo ključne značilnosti spontanega govorjenega diskurza in so v bistvu njegova reprodukcija. Razlog za to je uporabniška, socialno- in funkcijskozvrstna heterogenost spletnega okolja, v katerem uporabniki zavzemajo različne identitetne položaje, vloge in nagovarjajo druge uporabnike, ki so z njimi v

enakovrednem in/ali neenakovrednem družbenem razmerju. Tako v okviru spletnih komunikacijskih okolij, kot so tudi komentarji novic, najdemo značilnosti različnih socialnih in funkcijskih zvrsti.

Za spletno okolje je značilna prenosniška heterogenost, saj splet omogoča posredovanje tako pisne kot tudi govorne komunikacije. Spletna komunikacijska okolja, kot so tudi komentarji dnevnih novic, se po leksiki, semantiki in skladnji približujejo govorjenemu jeziku oz. spontani komunikaciji. Pogosto vsebujejo značilnost pogovornih jezikovnih zvrsti (pokrajinski pogovorni jezik, sleng, narečja), vsebujejo veliko referenčnih izrazov (npr. prvoosebnih in drugoosebnih zaimkov), eliptičnih stavkov (tj. nezapolnjenih stavčnih vzorcev, prekinitev, premorov, prekrivajočega govora ipd.) deiktov, besedilnih aktualizatorjev in diskurznih označevalcev.

Ključne besede: govorjeni jezik, spletni diskurz, leksika, semantika, skladnja

Regionalobala.Si News Comments Between Spoken And Written Discourse

The paper deals with the elements of spoken language in the comments of Regionalobala.si news. We are interested in the extent to which the comments are written in the manner of spontaneous speech and which characteristics of spoken discourse they show. The means of spoken discourse will be analysed with the observation of the written comments on the Regionalobala.si news portal published in the last month.

The characteristics of online news comment records can be explained by comparison to spontaneous speech or spontaneous spoken discourse. In relation to writings in the studied environments, we mainly use the terms online discourse, spontaneous discourse, online communication and spontaneous communication. Online

discourse is used as a realization of online communication, and spontaneous discourse is a realization of spontaneous communication and is a key feature of online discourse. We cannot use the term spontaneous speech or spontaneous spoken discourse, since it is essentially written, but compared to spontaneous speech the writings contain the key features of spontaneous spoken discourse and are mostly its reproduction. The reason for this is the user, social and functional heterogeneity of the online environment, in which users occupy different identity positions, roles and address other users who are in an equal and/or unequal social relationship with them. Thus, in the context of online communication environments, such as news comments, we find characteristics of different social and functional genres.

The online environment is characterized by media heterogeneity, as the web enables the transmission of both written and spoken communication. Online communication environments, such as daily news comments, approach spoken language in terms of lexis, semantics and syntax of spontaneous communication. They often contain characteristics of colloquial language genres (provincial colloquial language, slang, dialects), contain many referential expressions (e.g. first-person and second-person pronouns), elliptical sentences (i.e. unfilled sentence patterns, interruptions, pauses, overlapping speech, etc.), deicts, textual actualizers and discourse markers.

Keywords: spoken language, online discourse, lexicon, semantics, syntax

Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: Zbornik povzetkov
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Pomen lahkega branja in zvočne podpore za uporabnike aplikacije Digi-Scena

Pika Škerlj
Center Korak, Kranj, Slovenija,
pika.skerlj@center-korak.si

DIGI-SCENA je projekt, ki ranljivim skupinam omogoča aktivno vključevanje v digitalno družbo. Primarni ciljni skupini projekta DIGI-SCENA so ljudje z intelektualno oviro in pridobljeno možgansko poškodbo. Glavni nosilec projekta je Zavod RISA, center za splošno, funkcionalno in kulturno opismenjevanje, projektni partner je Center KORAK, Kranj, ki izvaja rehabilitacijo za osebe po pridobljeni možganski poškodbi. Digitalna pismenost je ena izmed ključnih vseživljenjskih kompetenc.

Aplikacija DIGI-SCENA, ki je dostopna na naslovu digiscena.si, bo omogočala druženje, klepet in svetovanje v varnem digitalnem okolju. DIGI-SCENA se od večine internetnih klepetalnic razlikuje, ker ne zahteva registracije in vpisa z elektronsko pošto in drugimi osebnimi podatki. Delovala bo na platformi ZOOM in bo aktivna ob vnaprej določenih terminih. Uporabnikom aplikacije bo omogočila priložnost, da razvijajo svoje digitalne veščine v skladu z individualnimi zmožnostmi in predhodnimi znanji. Aplikacija je opremljena z navodili v t. i. lahkem branju in zvočno podporo.

Lahko branje je namenjeno predvsem ljudem, ki zaradi različnih oviranosti trajno potrebujejo lahko berljive informacije in publikacije, ter ljudem, ki imajo slabše razvito veščino branja ali slabo poznajo jezik. Priprava besedil v lahkem branju poteka s testno skupino uporabnikov. Uporabniki poleg jezikovnih prilagoditev potrebujejo še druge prilagoditve, ki so odvisne od kognitivnih in drugih ovir. Testna skupina uporabnikov je sodelovala tudi pri oblikovanju strani in razporeditvi funkcij na strani DIGI-SCENA.

Uporabniki si pri uporabi aplikacije lahko pomagajo z različnimi možnostmi prilagoditev: z večanjem, manjšanjem črk, uporabo visokega kontrasta ali uporabo zvočne podpore. Zvočna podpora bo uporabnikom, ki težje berejo, poenostavila ali omogočila samostojno uporabo aplikacije. Omogočila bo lažjo navigacijo uporabnikom, ki težje berejo. Zvočna podpora je prepoznavna po piktogramu doprsne figure človeka s pogovornim oblakom. Vse funkcije in informacije na DIGI-SCENI so opremljene z zvočno podporo v slovenskem jeziku. Govorno in pisno sporazumevanje na vseh področjih javnega življenja v Sloveniji poteka v slovenščini. Za dolgoročno ohranitev jezika in vključujočo družbo je pomembna tudi njegova uporaba na internetu. Gumb za pridružitev klepetu je zaradi omejitev platforme ZOOM v jezikih, ki jih platforma podpira.

V vedno bolj digitalizirani družbi je pridobivanje digitalnih veščin ključnega pomena za socialno vključevanje. V okviru projekta izvajamo promocijsko-pripravljalne delavnice v različnih organizacijah statističnih regij ter predavanja/delavnice za zainteresirano javnost. Uporabnikom in podpornim osebam je všeč, da je uporaba aplikacije enostavna, ima zvočno podporo v slovenščini ter nima reklamnih pojavnih oken.

Pri razvoju aplikacije so bili upoštevani štirje gradniki informacijske pismenosti: komunikacija, reševanje problemov, uporaba informacijske tehnologije in načini razmišljanja. Poudarjen je bil pomen digitalne pismenosti in dostopnost informacij z vidikov izobraževanja/usposabljanja posameznika in prilagoditvah informacij.

Projekt Digitalna aktivacija za socialno vključenost in enakopravnost (DIGI-SCENA) je bil izbran na podlagi Javnega razpisa za digitalno preobrazbo nevladnih in prostovoljskih organizacij ter povečanje vključenosti njihovih uporabnikov v informacijsko družbo 2021–2023. Projekt sofinancira Ministrstvo za javno upravo

iz sklada za NVO. Projekt bo zaključen z medijsko konferenco. Po izteku projekta je načrtovano vsaj triletno vzdrževanje aplikacije. Končni cilj projekta, ki je v Sloveniji zaenkrat edinstven, je krepitev moči uporabnikov za socialno vključenost, za čim bolj samostojno informirano odločanje in dostop do pomoči – podpore, ko to potrebujejo. Projekt ima potencial za zmanjševanje digitalnega razkoraka v Sloveniji.

Ključne besede: digitalizacija, klepet, vključevanje, ranljive skupine, zvočna podpora

Significance Of Easy-To-Read Language And Audio Support For Users Of The DIGI-SCENA Application

DIGI-SCENA is a project that aims to enable socially disadvantaged groups to actively participate in digital society. The project is primarily aimed at people with intellectual disabilities and those who acquired traumatic brain injuries. The project is implemented by RISA Institute, a center for general, functional and cultural literacy. The project partner is Center KORAK, Kranj, a rehabilitation center for people with acquired brain injuries. Digital literacy is one of key competences for the entire life.

The application is accessible on digiscena.si, and allows socializing, chatting and consulting in a secure and moderated digital environment. DIGI-SCENA does not require registration and input of email and other personal data. It uses the ZOOM platform and will be active at set times. The app is equipped with easy-to-read instructions and audio support in Slovenian.

Two groups that use easy-to-read language are people that need easy-to-read information and publications all the time, and people with poorly developed reading skills or low language proficiency. The creation of easy-to-read content involved a test group of users that also participated in the design of the site and the arrangement of features on the DIGI-SCENA.

Users can customize the app with a variety of modifications: by increasing or decreasing the size of the letters, by using high contrast, or by using audio support. Audio support is intended to facilitate or enable users who have difficulty reading. Audio support can be identified by the pictogram of a human figure above the waist with a conversation cloud. All functions and information on DIGI-SCENA are provided with audio support in Slovenian language, the official language of the Republic of Slovenia. Spoken and written communication in Slovenian is used in all areas of public life in Slovenia. Digitization is the key to the long-term preservation of the language. Due to current available languages on ZOOM platform, "join button" is in currently supported languages only.

Acquiring digital skills is crucial for social integration in an increasingly digitalized society. For this reason, in addition to developing the app, free educational workshops are organized for future users. Users and their supporters like that the DIGI-SCENA is easy to use, provides audio support in Slovenian, and does not contain advertising pop-ups.

Four building blocks of information literacy were considered in the development of the app: communication, problem solving, use of information and communication technology, and thinking skills. The importance of digital literacy and information accessibility was prioritized. Users will have the opportunity to overcome social isolation and receive support in Slovenian.

The project "Digital Activation for Social Inclusion and Equality" (DIGI-SCENA) was selected in the public tender for the digital transformation of non-governmental and voluntary organizations and the greater involvement of their users in the information society 2021–2023. The project is co-financed by the Ministry of Public Administration from the Fund for Non-Governmental Organizations. The project will be concluded with a media conference. After the end of the project, app

maintenance is planned for at least three years. For the time being, the project is unique and has the potential to reduce the digital divide in Slovenia.

Keywords: audio support, chat, digitalization, inclusion, vulnerable groups

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Standardi transkribiranja narečnega korpusa GOKO

KLARA ŠUMENJAK
Univerza na Primorskem, Fakulteta za humanistične študije, Koper, Slovenija
klara.sumenjak@fhs.upr.si

Korpus GOKO (Govorni korpus Koprive na Krasu), ki je dostopen na spletni strani http://jt.upr.si/GOKO/, je leta 2013 nastal kot eden izmed ciljev doktorske disertacije *Opis govora Koprive na Krasu na osnovi dialektološkega korpusa*. Gre za realizacijo modela dialektološkega korpusa, ki bi z ustrezno nadgraditvijo in prilagoditvijo lahko služil kot izhodišče za gradnjo referenčnega dialektološkega korpusa.

Načela gradnje korpusa GOKO lahko v grobem razdelimo na tri glavne teme: 1) izhodišča za terensko delo, 2) pravna podlaga za gradnjo korpusa in 3) izhodišča za gradnjo korpusa. V prispevku bo predstavljena zadnja tema, ki obsega velikost korpusa, demografsko vzorčenje, posnetke in kriterije za njihov izbor, označevanje korpusa in transkripcijo, ki bo osrednja tema prispevka.

Najpomembnejši kriterij pri izdelavi načel transkribiranja govora, v našem primeru koprivskega krajevnega govora kraškega narečja, sta namen uporabe gradiva in naslovnik – od njiju je odvisna natančnost transkripcije. Več slovenskih avtorjev, predvsem slovstvenih folkloristov/etnologov in dialektologov (prim. Ivančič Kutin, 2011; Karničar, 2008; Kenda Jež, 2011; Klinar idr., 2012; Smole, 1994; Stanonik, 2001; Škofic, 2006; Zemljak Jontes idr., 2002) se je ukvarjalo z vprašanjem

transkripcije narečnih besedil, žal pa še niso našli povsem enotne rešitve, kako zapisati besedilo.

Zavedati se moramo, da imata dialektolog in zapisovalec oz. zbiratelj slovstvene folklore različne raziskovalne cilje – dialektologu je namreč pomembna predvsem izrazna plat besedila, izkazana z natančno fonetično transkripcijo, medtem ko je etnologu bolj pomembna vsebina besedila in vidi v narečnih potezah zapisanega le njegovo dodano vrednost.

Različni pa so tudi naslovniki: dialektološko gradivo z natančno fonetično transkripcijo je namenjeno predvsem dialektologom in jezikoslovcem, etnološko (s prilagojeno transkripcijo) pa tudi širši javnosti.

V slovenski dialektologiji se uporablja t. i. »nova nacionalna« fonetična transkripcija, ki sledi osnovnim načelom transkripcije Slovanskega lingvističnega atlasa (OLA) (Kenda Jež, 2011, 80–81), »tj. dogovorjen sistem znakov, ki zaznamuje kakovost in kolikost glasov, vključno z naglaševanjem (jakostnim/dinamičnim ali tonemskim)« (Smole, 1994, 150). Taka transkripcija je zelo zahtevna tudi za dialektologa, saj pri prekodiranju od njega zahteva veliko znanja, zbranosti in potrpljenja. »Je pa edini način, ki ohranja največ prvin, s katerimi je pripovedovalec oblikoval pripoved« (Ivančič Kutin, 2011, 61).

Korpus GOKO je namenjen tako dialektologom kot tudi širši javnosti, zato je zbrano gradivo zapisano v treh različicah, opremljenih z zvočnimi posnetki, in sicer a) v fonetičnem zapisu, ki upošteva vse glasoslovne značilnosti govora Koprive na Krasu – ta oblika je namenjena jezikoslovni, zlasti dialektološki analizi zapisanega govora; b) v poenostavljenem narečnem zapisu, kjer so ohranjene temeljne glasoslovne značilnosti krajevnega govora (zapisani so npr. mesto naglasa, diftongi, polglasnik ...) in c) v poknjiženi različici, kjer je vsaka posamezna beseda zamenjana s svojo knjižno ustreznico, na ravni besedne zveze in stavka pa se ohranjajo posebnosti govorjenega jezika. Taka različica je nujna za iskanje po korpusu, saj bo uporabnik najverjetneje iskal po knjižni besedi, npr. *zgodba*, ne pa po ortografskem zapisu besede, npr. *štorja*.

Glavno vprašanje pri poenostavljeni različici zapisa je, do kolikšne mere poenostaviti besedilo, da je različica še vedno reprezentativna za zapisani narečni govor, zato je pomembno pri zapisu ohraniti večino glasoslovnih posebnosti krajevnega govora, vendar pa jih zapisati z znaki, ki jih lahko berejo tudi laični uporabniki (torej s črkami knjižne abecede, ki jim lahko dodamo znak za polglasnik ipd.).
Tudi po desetih letih od začetka gradnje korpusa GOKO ostaja še precej odprtih vprašanj o načinu in vrsti transkripcije narečnih besedil, zato bodo v članku natančneje predstavljeni zlasti standardi in kriteriji transkribiranja v tem korpusu in možnosti izboljšave (predvsem s transkripcijo IPA).

Ključne besede: narečni korpus, GOKO, transkribiranje, kraško narečje, Kopriva na Krasu

GOKO Dialect Corpus Transcription Standards

The GOKO corpus (Speech Corpus of Kopriva in the Karst), which is available on the website http://jt.upr.si/GOKO/, was created in 2013 as one of the aims of the doctoral thesis *Dialectological description of the speech of Kopriva*. It is the realisation of a model of a dialectological corpus, which, with appropriate upgrading and adaptation, could serve as a starting point for the compilation of a reference dialectological corpus.

The principles for the compilation of the GOKO corpus can be divided into three main themes: 1) the starting points for the field study, 2) the legal basis for the compilation of the corpus, and 3) the starting points for the compilation of the corpus. The last topic will be presented in this paper, and includes the size of the corpus, demographic sampling, imagery and selection criteria, corpus tagging and transcription, which will be the focus of this paper.

The most important criteria in developing of principles for transcribing speech, in our case the Kopriva local speech of the Karst dialect, are the purpose of the material and the addressee - the accuracy of the transcription depends on them. Several Slovene authors, mainly Slovene folklorists/ethnologists and dialectologists (cf. Ivančič Kutin, 2011; Karničar, 2008; Kenda Jež, 2011; Klinar et al., 2012; Smole, 1994; Stanonik, 2001; Škofic, 2006; Zemljak Jontes et al., 2002), have worked on the issue of transcribing dialect texts, but unfortunately, they have not yet found a completely unified solution for how to transcribe the text.

We should be aware that the dialectologist and the recorder or collector of vernacular folklore have different research objectives - for the dialectologist, it is primarily concerned with the expressive aspect of the text, as demonstrated by an accurate phonetic transcription, whereas for the ethnologist is more concerned with the text itself and sees only the added value of the dialectal features of what is written down.

The recipients are also different: dialectological material with precise phonetic transcription is mainly intended for dialectologists and linguists, while ethnological material (with adapted transcription) is also intended for the wider audience.

In Slovenian dialectology the so-called "new national" phonetic transcription is used, which follows the basic principles of the Slavic Linguistic Atlas (OLA) transcription (Kenda Jež, 2011, 80-81), "i.e. such a transcription is also very demanding for the dialectologist, as it requires a great deal of knowledge, concentration and patience when transcribing. However, it is the only way to preserve most of the elements with which the narrator has shaped the story" (Ivančič Kutin, 2011, 61).

The GOKO corpus is intended both for dialectologists and the general public, so the collected material is transcribed in three versions, accompanied by sound recordings, namely a) in the phonetic transcription, which takes into account all the phonetic features of the local speech - this form is intended for linguistic, especially dialectological analysis of the recorded speech; b) in the simplified transcription, which preserves the basic phonetic features of the local speech (e.g., place of accent, diphthongs, semivowels) c) in the standard version, where each individual word is replaced by its standard lexical equivalent, while preserving the peculiarities of the spoken language at the level of phrases and sentences. Such a version is necessary

for searching the corpus, as the user is likely to be looking for a standard word, e.g. *zgodba*, rather than for the orthographic spelling of a dialectal word, e.g. *štorja*.

The main problem with the simplified version is how much to simplify the text so that the version is still representative of the transcribed dialect, so it is important to preserve most of the phonetic features of the local speech, but to write them in characters that can be read by laypeople (i.e. letters of the standard alphabet, to which a semi-vowel can be added, etc.).

Even ten years after the start of the GOKO corpus, there are still many open questions about the way in which dialect texts are transcribed, so this article will look at the standards and criteria for transcription in this corpus and the possibilities for improvement (especially in IPA transcription).

Keywords: dialect corpus, GOKO, transcription standards, Karst dialect, Kopriva in the Karst

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Prednosti in slabosti dvotirnega zapisovanja govora v slovenskih govornih virih

DARINKA VERDONIK,[1] MITJA TROJAR,[2] ANDREJA BIZJAK[1]
[1] Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor, Slovenija, darinka.verdonik@um.si, andreja.bizjak1@um.si
[2] ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, Ljubljana, Slovenija mitja.trojar@zrc-sazu.si

Časovno in finančno najzahtevnejši korak izdelave govornih virov je transkribiranje posnetkov. To delo poleg zapisa govora običajno vključuje tudi popis podatkov o govorcih in posnetih govornih dogodkih, segmentacijo govora na osnovne enote – izjave, označevanje menjavanja govorcev, označevanje akustičnega ozadja (npr. prisotnost šuma ali glasbe) in akustičnih dogodkov (nenadni zvoki od zunaj ali nastali z govorili, kot je kašljanje, glasni vdihi ipd.) ter osnovnih prozodičnih značilnosti (smeh, premori ipd.).

Zaradi časovne in finančne zahtevnosti transkribiranja se ob izdelavi govornih virov vedno iščejo načini, kako izvedbo čim bolj ekonomizirati. En način je, da dobimo posnetke, ki že imajo zapis govora (na primer iz parlamenta ali iz medijskih hiš). Toda ti zapisi navadno niso ustrezno natančni za raziskovalne namene in jih je treba dodatno urejati. Drug način je uporaba tehnologije avtomatskega razpoznavanja govora za izdelavo transkripcij. Tudi ta rešitev ni popolna, nekaj napak bo vedno prisotnih. Poleg tega je vprašljivo, koliko bo primerna za zahtevnejše oblike govora, kot so govor v šumnem okolju ali slaba kvaliteta posnetka. Prav tako bi bilo treba

razviti dodatna orodja za prepoznavo govorcev in označevanje menjavanja govorcev.

Ročna izvedba zapisa govora bo tako še nekaj časa neizogiben korak izdelave govornih virov višje kvalitete. Na podlagi priporočil EAGLES (Gibbon et al. 1997) in praks v govornih korpusih lahko ločimo več ravni zapisa govora, od ortografskega prek fonetičnega, pripravljenega avtomatsko z algoritmi grafemsko-fonemske pretvorbe, do natančnega fonemskega zapisa v fonetični abecedi. Pri ortografskih zapisih lahko nadalje ločujemo:

- standardni ortografski zapis, to je povsem standardiziran zapis, v katerem izgovorjene besede in besedne oblike zapišemo z najbližjimi standardnimi besedami in besednimi oblikami,
- razširjen ortografski zapis z dodatnimi pravili in/ali seznami za zapisovanje posebnosti izgovorjenih besed in besednih oblik.

V slovenskih govornih virih je uveljavljena praksa zapisovanja na oba navedena načina, prvi način je poimenovan standardizirani, drugi pogovorni zapis. To pomeni, da je govor zapisan dvakrat. Ta praksa je bila vzpostavljena s korpusom GOS (Verdonik et al. 2013) ter delno prilagojena in posodobljena ob izdelavi govorne baze ARTUR (http://hdl.handle.net/11356/1772). Podobno prakso najdmo še pri drugih, zlasti slovanskih jezikih, na primer češkem, slovaškem, hrvaškem.

Potem ko je bilo na ta način v govorni bazi ARTUR zapisanih več kot 300 ur javnega, nejavnega in parlamentarnega govora, je smiselno vprašanje, ali je ta dvotirni sistem zapisovanja govora potreben tudi vnaprej: kaj so njegove prednosti in kaj slabosti glede na dodaten zahtevani trud? V prispevku bomo predstavili razloge za dvotirni način zapisovanja; pregledali osrednje značilnosti priporočil za pogovorni in standardizirani zapis na podlagi izkušenj pri transkribiranju baze ARTUR; predstavili izkušnje, kje se poraja največ problemov pri zapisovanju in kje je pričakovati dodatne zaplete, če bi se izvajal samo standardizirani zapis; ter analizirali čas, potreben za izvedbo dodatnega nivoja zapisovanja.

Ključne besede: transkribiranje, standardizirani zapis, ortografska transkripcija, pogovorni zapis, fonetična transkripcija

Advantages and Disadvantages of Two-level Speech Transcription in the Slovenian Speech Resources

The most time-consuming and costly step in the production of speech resources is transcription of speech. In addition to transcribing what was said it includes listing information about speakers and recorded speech events, segmenting speech into basic units - utterances, annotating speaker turns, acoustic background (e.g. the presence of noise or music), acoustic events (sudden sounds from outside or produced by speaker, such as coughing, loud breaths, etc.) and basic prosodic features (laughter, pauses, etc.).

This is time- and cost-consuming process therefore it is important to make the transcription process as economical as possible. One way is to obtain recordings that already have transcriptions (for example, from parliament or from media). However, such recordings are usually not sufficiently accurate for research purposes and need further editing. Another way is to use automatic speech recognition technology to produce transcriptions. This solution is also not perfect, some errors will always be present. Moreover, the quality of automatic transcriptions is questionable for non-standard speech, simultaneous speech, speech in noisy environments or at poor recording quality. Additionally, tools for speaker recognition and turn-taking annotation should be developed.

Manual transcription will thus continue to be an inevitable step in the production of high quality speech resources, as it is the only way to ensure accurate information at all levels of transcription. Based on EAGLES recommendations (Gibbon et al. 1997) and practices in speech corpora, we can distinguish several levels of speech transcription, from orthographic through phonetic, prepared automatically by grapheme-to-phoneme conversion algorithms, to precise phonemic transcription in phonetic alphabet. Orthographic transcription can be divided into:

1. standard orthographic transcription, i.e. transcription with standard lexicon where non-standard words and word-forms are written with the nearest standard equivalents,

2. an extended orthographic transcription with additional rules and/or lists for writing down the words and words-forms which are not part of the standard lexicon.

In the Slovenian speech resources, both levels of orthographic transcription are used, the former is termed standardized transcription, the latter the pronunciation-based transcription. This means that all speech is transcribed twice. This practice was established with the GOS corpus (Verdonik et al. 2013) and updated while creating the ARTUR speech database (http://hdl.handle.net/11356/1772). Similar practices can be found in other languages, particularly Slavic – Czech, Slovak, Croatian.

In the ARTUR speech database, 300 hours of public, non-public and parliamentary speech were transcribed using this two-level transcription system. At the end of this work, the central question is whether such two-level transcription system should remain in the future: what are its advantages and disadvantages, considering the additional effort required? In this paper, we will present the reasons for the two-level transcription; we will sum the most recent recommendations for both levels of transcription that emerge from the ARTUR speech database; we will point-out our experiences where are the most problematic issues in both transcription modes and what problems should be expected if only standardised transcription was used; and analyse the time required to implement the additional level of transcription.

Keywords: transcription, standardised transcription, orthographic transcription, pronunciation-based transcription, phonetic transcription

Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: Zbornik povzetkov
M. Krajnc Ivič (ur.)
Univerzitetna založba Univerze v Mariboru

Multimodalna kohezija v literarnem branju

Branislava Vičar, Katja Plemenitaš
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
branislava.vicar@um.si, katja.plemenitas@um.si

Prispevek analizira kompleksno součinkovanje različnih semiotskih kodov pri literarnem branju. Literarno branje lahko opredelimo kot multimodalno besedilno vrsto, ki jo opredeljuje javno glasno branje literarnih besedil. Navadno poteka pred občinstvom, lahko pa je tudi medijsko posredovano, in sicer kot spletni javni dogodek ali kot posnetek brez neposredne prisotnosti občinstva. Besedila literarnega branja so torej glasno brana različica zapisanih literarnih besedil, torej so po svoji naravi monološka. Medosebni odnos, ki se ustvarja med literarnim branjem, je odnos med interpretko_om in občinstvom. Kadar glasno branje izvaja sam avtor_ica literarnega besedila, se za to besedilno vrsto pojavlja izraz avtorsko branje (Podbevšek 2021).

Z raziskavo preučujeva, kako se dejanja na višji ravni realizirajo skozi interakcijo branja ter slišnih in vidnih semiotskih virov, kot so geste, pogled, obrazna mimika, rokovanje s predmeti, vizualne projekcije, vokalizacija, fizični prostor in glasba. Multimodalna kohezija je pri tem opredeljena kot »povezava med načini, ki se materializirajo v strukturi multimodalnega besedila« (Stöckl in Bateman 2022). Da bi opisali, kako se pri literarnem branju različni semiotski kodi medsebojno povezujejo, sva v raziskavo vključili koncepta modalne kompleksnosti, tj. medsebojnega delovanja različnih kodov, in modalne intenzivnosti, tj. relativne pomembnosti določenega semiotskega koda (Messner 2022). Za osvetlitev multimodalnih praks v

literarnem branju sva v uporabili kombinacijo multimodalne interakcijske analize (Norris 2019) in multimodalne registrske analize (Martin in Rose 2007). Korpus za analizo vsebuje video posnetke branja literarnih del v slovenščini. Z analizo si bova prizadevali pokazati, »kako lahko različne kombinacije semiotskih kodov součinkujejo, da tvorijo koherentne komunikacijske artefakte« (Bateman 2014). Ugotovitve raziskave osvetljujejo način, kako so semiotski kodi medsebojno povezani in kako se medsebojno dopolnjujejo v multimodalni celoti.

Ključne besede: multimodalni diskurz, multimodalna interakcija, kohezija, čezmodalna kohezivna vez, literarno branje

Multimodal Cohesion in Literary Reading

The paper analyzes the complex interplay of semiotic modes in the setting of literary readings. Literary reading can be defined as a multimodal text type that is characterized by the public reading of literary texts. It usually takes place in front of an audience, but it can also be transmitted by media as an online public event or as a recording without the direct presence of an audience. Texts of literary reading are spoken versions of written literary texts, and as such monologic in nature. The interpersonal relationship that is created during literary readings is the relationship between the performer and the audience. When reading aloud is performed by the author of a literary text, the term authorial reading can be applied to this text type (Podbevšek 2021).

The study deals with the way in which specific higher-level actions are realized through the interaction of reading and audible and visible semiotic resources, such as gesture, gaze, facial expression, handling of objects, visual projections, vocalizing, physical space, and music. In this study, we understand multimodal cohesion as "links between modes that materialize in the structure of a multimodal text" (Stöckl and Bateman 2022). To describe how different modes are interlinked in literary reading, we include the concepts of modal complexity, i.e., the interplay of different

modes, and modal intensity, i.e., the relative importance of specific modes (Messner 2022). To explore multimodal practices in literary readings, this study adopts a mixed-methods approach by combining multimodal interaction analysis (Norris 2019) and multimodal register analysis (Martin and Rose 2007). We will work with a corpus consisting of video data from different readings of literature in Slovene. We will illustrate "how diverse combinations of semiotic modes can work together to form coherent communicative artifacts" (Bateman 2014). The research findings shed light on the interaction of modes in literary readings and their contribution to the multimodal text as whole.

Key-words: multimodal discourse, multimodal interaction, cohesion, cross-modal cohesive tie, literary reading

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)

Univerzitetna založba Univerze v Mariboru

Tvorba korpusů mluveného jazyka

MILOSLAV VONDRÁČEK
Slezská univerzita v Opavě, Opavě, Česká republika
miloslav.vondracek@fpf.slu.cz

Příspěvek shrnuje zkušenosti s obstaráváním zvukových záznamů neoficiálních komunikačních situací a s jejich přepisem pro účely Českého národního korpusu – pro tvorbu korpusů mluvené češtiny. Spolu se studenty jsme pořídili zvukový záznam více než 220 dialogických situací a jejich jednoúrovňový převod do písemné formy. Při té příležitosti jsme museli řešit řadu praktických problémů. Ty vedly k formulaci podstatných teoretických otázek.

Za dobu naší práce na korpusových podkladech (od r. 2005 cca do r. 2012) jsme například dospěli k rezignaci na signalizaci začátku a konce věty a souvětí, jak jsou obvyklé v psané formě řeči. Bylo třeba vypořádat se se zápisem hezitačních zvuků komunikačně relevantních (souhlasné, váhavé, odmítavé aj. *hmm, emm, eee*...) a s foneticko-fonologickými deformacemi slov procházejících slovnědruhovou transpozicí, zejm. útvarů směřujících mezi partikule a interjekce (*člověče, čoveče, čoeče, čéče*). Při tom všem měla být zajištěna uživatelská zpracovatelnost takto přepsaného jazykového materiálu, tj. jeho automatické strojové rozpoznání a opatření metajazykovými daty, stejně jako následná dohledatelnost možných výrazových forem.

Další okruh poznatků se týká percepčních a kognitivních limitů zpracovatelů zvukového záznamu. Zjišťujeme, že editoři sond píší to, co předpokládají, že slyší, slyšeno být může nebo má, a že více či méně podléhají tendenci zdůrazňovat nepravidelnosti mluvené řeči proti jevům pravidelným, nebo naopak inovace

přehlížejí a prosazují ustálené formy psaného jazyka. Totéž v podstatné míře platilo pro hranice výpovědí, dokud byly zaznamenávány.

Idealizace mluvené řeči při převodu do psané formy (systemizace parole) spočívá v pravidelném členění komunikátu na relativně symetrické, nepříliš rozsáhlé výpovědi složené z relativně izolovaných jednotek roviny lexikální, ukončené koncovým interpunkčním znaménkem – bez ohledu na (obtížně identifikovatelný) koncový předěl vyjadřovaný prostředky prozodickými. Co však odkrýváme, je relativita jednotek mluveného jazyka.

Editoři sond věnují, podle mé zkušenosti, zápisu mluvené řeči potřebnou pozornost. V kolísání grafických forem, které zvolí, se projevují vágní hranice jednotek řeči, komplikované limity vnímání reprodukovaného mluveného projevu. Každý uživatel korpusů mluvené komunikace bude s těmito omezeními muset počítat. Psané korpusy mohou být dobrým korektivem spektra volených výrazových variant. Pokud budou cílem korpusového výzkumu jednotky nižších rovin, přepis může být přinejmenším vodítkem pro zevrubné zkoumání zvukového záznamu. Tendence omezit variabilitu forem v první úrovni přepisu se prosazuje v nové metodologii. Naopak, ani variabilita užitých grafických forem není na závadu. Naznačuje intuitivní postřeh editora o variantní funkci jazykové jednotky. Pokud se stanou studenti jako editoři sond díky této práci vnímavější k jazyku, naplní se i původně nezamýšlený smysl existence korpusů. Výsledkem našich postřehů je metodologie tvorby korpusu, od té doby neustále zdokonalovaná. Příspěvek přináší přehled základních otázek a snaží se poskytnout teoretické odpovědi i metodiku řešení.

Klíčová slova: korpusy mluveného jazyka, relativita jednotek řeči, pravidla přepisu, zvukový záznam neoficiálních komunikačních situací

Creation of Spoken Language Corpora

The paper summarizes my experience with the acquisition of audio recordings of unofficial communication situations and with their transcription for the purposes of the Czech National Corpus: for the creation of corpora of spoken Czech. With the

students we made audio recordings of more than 220 dialogue situations and we processed their one-level conversion into written form. We had to solve a number of practical problems on that occasion. These difficulties led us to important theoretical questions.

We have resigned ourselves to noting the beginning and end of sentences and clauses as they are usual in the written form of speech, for example. It was necessary to deal with the notation of communicatively relevant hesitant sounds (agreeable, indefinite, negative, etc. *hmm, emm, eee...*) and with phonetic-phonological deformations of words, which are the result of transposition, especially formations directed between particles and interjections (*člověče, čoveče, čoeče, čéče*). In all of this, the user processability of the transcribed language material was to be ensured, i. e. automatic machine recognition and provision of the material with meta-linguistic data, as well as the subsequent traceability of possible expression forms.

Another area of knowledge concerns the perceptual and cognitive limits of audio recording processors. We find that editors write what they think they hear, what can or should be heard. They are more or less subject to the tendency to emphasize the irregularities of spoken language against regular phenomena. Or, on the contrary, they overlook innovations and promote established forms of written language. The same was essentially true of sentence boundaries as long as they were recorded.

During the conversion of spoken language into a written form, parole is systematized and idealized. But what we uncover, however, is the relativity of the units of spoken language. In my experience, the editors of the probes pay the necessary attention to the recording of the spoken speech. Vague boundaries of speech units, complicated limits of perception of reproduced speech are reflected in the fluctuations of the graphic forms they choose. Every corpora user of spoken communication will have to reckon with these limitations. Written corpora can be a good corrective to the spectrum of chosen expression variants. The transcript can at least be a guide to a thorough examination of the audio recording. The tendency to limit the variability of the forms in the first level of transcription is enforced in the new methodology. On the contrary, the variability of graphic forms is not a problem either. It indicates the intuitive perception of the editor about the variant function of the language unit. If students become more receptive to language as a result of this work as editors of probes, the originally unintended meaning of the

existence of corpora will also be fulfilled. The result of our observations is the corpus creation methodology, which has been continuously improved since then. The contribution provides an overview of the basic questions and tries to provide theoretical answers as well as a solution methodology.

Keywords: spoken language corpora, relativity of speech units, transcription rules, audio recording of unofficial communication situations

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)

Univerzitetna založba Univerze v Mariboru

Uporaba mikrofenomenološkega intervjuja pri raziskovanju igralčevega govora

MARTIN VRTAČNIK
Akademija za gledališče, radio, film in televizijo, Ljubljana, Slovenija
martin.vrtacnik@agrft.uni-lj.si

Število raziskav govora se je v prvih dvajsetih letih tretjega tisočletja povečalo, k čemur so pripomogle spremenjene družbene okoliščine ter bolj izpopolnjene in dostopnejše tehnologije za raziskovanje govora. Nastaja tudi vse več raziskav umetniškega govora, tudi odrskega govora, kar je najverjetneje posledica organiziranja kolokvijev oziroma znanstvenih simpozijev o umetniškem govoru na Akademiji za gledališče, radio, film in televizijo. Raziskovalci govora uporabljajo različne metodologije in interdisciplinarno povezujejo jezikoslovje in druge vede. Po načelu interdisciplinarnosti se na področju govora povezujeta teatrologija in slovenistika.

Pionir gledališkega lektorstva Oton Župančič je zapisal, da je jezik čustvo in misel, ter poudarjal pomen igralčeve telesnosti in duševnosti. Njegov naslednik Mirko Mahnič je pisal o duhovni substanci glasu, Stanko Škerlj o psihičnem substratu govora, Jože Tiran pa v razpravljanju o igralčevem ustvarjalnem procesu uporablja pojme podzavest, delovanje živčevja in možganov ter vzbujanje čustev. Vse to potrjuje dejstvo, da je govor rezultat kognitivnih procesov, zato je tudi pri raziskovanju odrskega govora treba izhajati iz naravoslovja oziroma biologije. Na

pomen bioloških procesov pri celostni obravnavi igralčevega govora opozarja eden pomembnejših premišljevalcev o umetniškem govoru Kristijan Muck.

Izkušnje gledaliških ustvarjalcev iz ZDA kažejo, da sta kognitivna znanost in teatrologija povezani že desetletja. V osemdesetih letih 20. stoletja je izsledke raziskav kognitivne znanosti pri poučevanju igre in režije začel uporabljati John Emigh. Približno dvajset let kasneje sta kognitivni obrat v teatrologiji utemeljila urednika Bruce McConachie in F. Elizabeth Hart v zborniku *Performance and Cognition: Theatre Studies and the Cognitive Turn* (2006), natančneje pa je kognitivno nevroznanost s področjem gledališke igre povezal Rick Kemp v delu *Embodied Acting: What Neuroscience Tells Us About Performance* (2012). Gledališki ustvarjalci za razliko od nevroznanstvenikov v raziskavah ne uporabljajo tehnologije, kot je funkcionalna magnetna resonanca, pozitronska emisijska in računalniška tomografija, sledenje očesnim gibom in merjenje avtonomnih odzivov kože, pač pa uporabljajo drugačne metode, med drugimi opis svojega izkustva. Tako so nekateri slovenski dramski igralci opisali oblikovanje lastnega odrskega govora.

Subjektivne izkušnje so bile do nedavnega izključene iz znanstvenega raziskovanja. Ker se le-te odvijajo tudi pod pragom zavesti, je njihovo opisovanje zapleteno, zato si znanstveniki prizadevajo razviti stroge metode za njihovo natančno proučevanje. Nevrobiolog Francisco J. Varela je poudaril, da človeškega uma ne moremo raziskovati le z nevroznanstveno tehnologijo, zato je oblikoval nevrofenomenološki program, ki ga razvija utemeljiteljica mikrofenomenologije Claire Petitmengin. Oblikovala je mikrofenomenološki intervju, ki nam omogoča, da intervjuvanca pripravimo do tega, da se zave svoje subjektivne izkušnje in jo opiše.

Metoda intervjuja je v teatrologiji pogosta. Eva Pori je z metodo rekonstrukcijskega intervjuja pridobivala informacije o oblikovanju igralskega govornega in telesnega izraza na Odru 57, vprašanje pa je, ali bi bila v teatrologiji uporabna tudi metoda mikrofenomenološkega intervjuja. Dramske igralce bi vprašali, kako posvojijo posamezno izjavo v dramskem besedilu, da izberejo ustrezno intonacijo, pomen in smisel te izjave.

Čeprav se mikrofenomenološki intervju, ki je bil zasnovan za raziskovanje epilepsije, intuicije in meditacije, uporablja tudi na področju umetnosti, pa raziskav odrskega govora s to metodo v literaturi ni. Predpostavljamo, da bi z mikrofenomenološkim

intervjujem lahko dodatno raziskali najmanj raziskano področje odrskega govora – ustvarjanje igralčevega glasovnega sloja.

Ključne besede: odrski govor, gledališki lektor, kognitivni obrat, nevroestetika, mikrofenomenološki intervju

The Use of Microphenomenological Interview in the Research of Actor's Speech

Speech research increased during the first twenty years of the third millennium, which was aided by changed social circumstances and more sophisticated and accessible speech research technologies. There are also more and more research on artistic speech, including stage speech, which is most likely the result of the organisation of conventions or scientific symposiums on artistic speech at the Academy of Theatre, Radio, Film and Television. Speech researchers employ different methodologies, linking linguistics and other sciences in an interdisciplinary way. According to the principle of interdisciplinarity, theatre studies and Slovene studies are connected in the field of speech.

The pioneer of theatre language consulting, Oton Župančič, wrote that language is emotion and thought, and emphasized the importance of the actor's physicality and mentality. His successor, Mirko Mahnič, wrote about the spiritual substance of voice, Stanko Škerlj about the psychic substrate of speech, while Jože Tiran, in discussion of the actor's creative process, uses the concepts of the subconscious, the functioning of the nervous system and brain, and the arousal of emotions. All this confirms the fact that speech is the result of cognitive processes, which is why, even when researching stage speech, it is necessary to start from natural science or biology. Kristijan Muck, one of the most crucial thinkers on artistic speech, stresses the importance of biological processes in the holistic treatment of an actor's speech.

The experience of theatre creators from the USA demonstrates that cognitive science and theatre studies have been connected for decades. In the 1980s, John Emigh began to use the findings of cognitive science research in teaching acting and directing. About twenty years later, editors Bruce McConachie and F. Elizabeth Hart established the cognitive turn in theatre studies in the book *Performance and Cognition: Theatre Studies and the Cognitive Turn* (2006), whilst Rick Kemp connected cognitive science with the field of theatre play more precisely in the work *Embodied Acting: What Neuroscience Tells Us About Performance* (2012). Unlike neuroscientists, theatre creators do not use such technology as functional magnetic resonance, positron emission and computed tomography, tracking eye movements and measuring autonomic skin responses in their research, but instead use different methods, including the description of one's own experience. This is how some Slovene drama actors described the formation of their own stage speech.

Until recently, subjective experiences were excluded from scientific research. Since these occur below the threshold of consciousness, it is complicated to describe them, so scientists strive to develop rigorous methods to study them more precisely. Neurobiologist Francisco J. Varela pointed out that human mind cannot be studied only with neuroscientific technology, which is the reason why he designed a neurophenomenological program, which is being developed by the founder of microphenomenology, Claire Petitmengin. She designed a microphenomenological interview, which enables us to make the interviewee aware of their own subjective experience and describe it.

This interviewing method is common in theatre studies. Moreover, with the reconstructive interviewing method, Eva Pori obtained information about the formation of actor's speech and body expression on Oder 57 (Stage 57), but the question is whether the microphenomenological interviewing method would be useful also in theatre studies. Drama actors would be asked how they adopt a particular statement in a dramatic text to select the appropriate intonation, meaning and sense of that statement.

Although the microphenomenological interview, which was devised to examine epilepsy, intuition and meditation, is used also in the field of art, there are no studies of stage speech employing the method in question in the existing literature. We presume that with the help of microphenomenological interview we could further

explore the least researched area of stage speech – the creation of the actors' vocal layer.

Keywords: stage speech, language consultant, cognitive turn, neuroaesthetics, microphenomenological interview

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV
M. Krajnc Ivič (ur.)

Univerzitetna založba Univerze v Mariboru

Standardizacija prekmurske transkripcije samoglasnikov: študija primera

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

Eden izmed ključnih ciljev sodobnih raziskav spontane govorjene rabe jezika je pregled stanja in opredelitev potreb po govornih podatkih ter pripadajoči raziskovalni infrastrukturi. Pri tem je pomemben tudi socialnozvrstni vidik, pri čemer predstavljajo narečne skupine različne raziskovalne izzive.

Sodelujoče skupine bodo v projektu *Mezzanine* v raziskavo vključile različne narečne skupine. Prispevek se bo osredinjal na pilotno obravnavo panonske narečne skupine, natančneje prekmurskega narečja, s ciljem, da se preveri primernost v projektu predvidenega postopka standardizacije slovenske narečne transkripcije samoglasnikov. V analizo bosta zajeta (a) gradivska tabela samoglasniških odrazov za prekmursko narečje, pripravljeno po jezikovnogeografskem gradivu gramatičnega dela vprašalnice za *Slovenski lingvistični atlas*, in (b) arhivsko zvočno gradivo, kolikor je ohranjeno v zvočnem arhivu Dialektološke sekcije ZRC SAZU in v zvočnem arhivu Filozofske fakultete Univerze v Mariboru.

Za prekmursko narečje bodo na izbranem gradivu analizirana vprašanja (a) obstoja oz. ohranitve kolikostnih nasprotij, (b) kakovosti sestavin dvoglasnikov, (c) kakovosti dolgega naglašenega *a*-ja, (č) kakovosti reduciranega nenaglašenega *i*-ja (v

primerjavi z visokimi centraliziranimi samoglasniki v drugih narečjih). Predvidene stopnje obdelave podatkov predstavljajo: 1. analiza narečnega gradiva (pretežno iz 2. polovice 20. st.) in ugotavljanje transkripcijskih razhajanj oz. problematičnih točk zapisa; 2. primerjava z načinom zapisa v dialektološki literaturi in gradivu, ki je dostopno po posameznih regionalnih središčih; 3. akustična analiza dostopnih zvočnih posnetkov, po potrebi terensko delo za pridobivanje novega zvočnega gradiva, eksperimentalnofonetična analiza; 4. sinteza ugotovljenega stanja. Metodologija dela in izbrani rezultati obdelave podatkov bodo predstavljeni na konferenci.

Ključne besede: spontana govorjena raba jezika, prekmursko narečje, kolikost samoglasnikov, kakovost samoglasnikov, eksperimentalnofonetična analiza

Standardization of Prekmurian Vowel Transcription: a Case Study

One of the key goals of modern spontaneous spoken language research use is to review the situation and define the needs for speech data and the associated research infrastructure. Its social language varieties aspect is also an important point of view, with different dialectal groups presenting different research challenges.

Different Mezzanine project participating groups will research different dialectal groups. The paper will focus on the pilot study of the Pannonian dialectal group, more specifically the Prekmurje dialect, with the aim of verifying the suitability of the standardization process of the Slovene dialectal vowel transcription vowels envisaged in the project. The analysis will include (a) a table of vowel reflections for the Prekmurje dialect, based on the linguistic-geographical material of the grammatical part of the questionnaire for the *Slovene Linguistic Atlas*, and (b) archival audio material, as far as it is preserved in the audio archive of the Dialectological Section of the ZRC SAZU and in the audio archive of the Faculty of Arts of the University of Maribor.

For the Prekmurje dialect, the questions of (a) existence or preservation of quantitative contrasts, (b) quality of diphthong components, (c) quality of long stressed *a*, (d) quality of reduced unstressed *i* (in comparison to high centralized vowels in other dialects) will be envestigated. The anticipated stages of data processing are: 1. analysis of dialectal material (mainly from the second half of the 20th century) and identification of transcription discrepancies or problematic transcription segments; 2. comparison of the principles of transcribing, accessible in dialectological literature and material gathered for regional centers; 3. acoustic analysis of accessible audio recordings, if necessary fieldwork to acquire new audio material, experimental phonetic analysis; 4. synthesis of the established situation. The methodology and selected data processing results will be presented at the conference.

Keywords: spontaneous spoken language use, the Prekmurje dialect, vowel quantity, vowel quality, experimental phonetic analysis

Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: Zbornik povzetkov
M. Krajnc Ivič (ur.)

Univerzitetna založba Univerze v Mariboru

Raziskovanje govorjenega umetniškega jezika

Nina Žavbi
Akademija za gledališče, radio, film in televizijo Univerze v Ljubljani, Ljubljana, Slovenija
nina.zavbi@agrft.uni-lj.si

Raziskovanje govora je v primerjavi z raziskovanjem jezika precej zapostavljeno, še bolj pa je zapostavljeno raziskovanje umetniškega govora. Govor je zvočna realizacija jezika, je konkretna ubeseditev v določenih besedilnih in zunajbesedilnih okoliščinah, poleg zvočne pa ga spremlja tudi vidna komponenta. Šele z opazovanjem obeh lahko v polnosti razumemo in analiziramo vse pomembne govorne prvine.

Govorjeni umetniški jezik je z vidika raziskovanja zanimiv, saj se pogosto naslanja na zapisano predlogo, na umetniško besedilo, ki je nato govorno interpretirano pred publiko; pogosto ga ustvarjalno oblikujejo govorni profesionalci (npr. igralci). To drži tudi za odrski govor – ki je govorna izvedba dramskega besedila (predvsem v dramskem gledališču, v nekaterih drugih oblikah odrske umetnosti pa se lahko tudi zelo približa spontanemu, zasebnemu itd.). Gre za govor, ki skuša dajati vtis sprotne tvorjenosti, naravnosti, spontanosti, kljub temu da to ni. Osebe, ki besedilo izgovarjajo v gledališki uprizoritvi, govorijo tako, kot da besedilo ubesedujejo prvič, torej spontano, v resnici pa je govor vnaprej pripravljen po besedilni predlogi, vsi govorni vidiki (zvrstnost; besedje; pravorečni elementi – naglasi, izgovor glasov; prozodična sredstva – hitrost, glasnost govora, intonacija, premori, register itd.; tudi vidna nebesedna govorna sredstva – mimika, geste) so v resnici do popolnosti premišljeni, ozaveščeni in izvedbeno fiksirani.

V proučevanju odrskega govora so se v zgodovini posluževali različnih metod, predvsem slušnozaznavne analize (Podbevšek, Sušec Michieli itd.). V sodobnosti se s slušnozaznavno kombinira akustična (fonetična) analiza z različnimi računalniškimi programi (npr. Praat). Model raziskovanja (Žavbi, 2017) slovenskega odrskega govora je bil pripravljen po zgledu raziskav odrskega govora nekaterih hrvaških raziskovalcev (Škarić, Varošanec-Škarić, Vrban Zrinski) in slovenskega medijskega govora (Tivadar, Huber).

Podrobnejše znanstvene analize govorjenega umetniškega jezika so bile (v zgodovini) mogoče komaj takrat, ko so obstajale tehnične možnosti ponovnega poslušanja govora – posnetki. Pred tem se je izhajalo le iz ogleda in poslušanja posamezne predstave, torej se je govor popisovalo bolj celostno, tudi po spominu ter na podlagi kritiških zapisov. Avdioposnetki so omogočili večkratno poslušanje, zato tudi slušno analizo, ki je bila zelo poglobljena. Videoposnetki so omogočili hkratno proučevanje vidnih spremljevalcev govora. Vendar pa za sodobne načine raziskovanja ni dovolj imeti posnetke predstav. Za akustično analizo, ki v raziskovanje prinese večjo mero objektivnosti, torej preverjanje slušnih zaznav, potrebujemo primerne računalniške programe, npr. Praat, ki je uporaben za večjo skupino ljudi, saj je prosto dostopen in dokaj enostaven. Za kvalitetno proučevanje je po pridobljenih meritvah ključna poglobljena interpretacija rezultatov oziroma postavljanje pridobljenih podatkov v kontekst (tako kontekst dramskega besedila kot tudi celotne uprizoritve ter upoštevanje zunanjih dejavnikov, npr. velikosti dvorane, družbenega konteksta, umetniškega konteksta oz. obdobja, dramskega avtorja itd.).

V prispevku bom poskušala prikazati razvoj raziskovanja umetniškega (odrskega) govora na Slovenskem. Osredotočila se bom na sodobni model raziskovanja, pri katerem kombiniramo slušnozaznavno in akustično analizo. Opredelila bom prednosti takšnega načina in predstavila strategijo raziskovanja, ki odrski govor proučuje v razmerju do dramskega besedila ter kot enega od uprizoritvenih dejavnikov. Interdisciplinarni način raziskovalne rezultate kontekstualizira in tako osmišlja proučevanje umetniškega fenomena odrskega govora, ki povezuje znanost in umetnost.

Ključne besede: jezik, govor, umetniški govor, akustična analiza, Praat

Researching Artistic Speech

The research of speech, compared to that of language, is particularly overlooked; even more unexplored is the research of artistic speech. Speech is the sound realisation of language, the concrete verbalisation in specific textual and extratextual settings. It comprises sound and visual components. Only through observing both can we completely understand and analyse all the essential speech elements.

From a research perspective, spoken artistic language is unique because it often relies on the written form, on the artistic text, which, often after being creatively shaped by speech professionals, is publicly spoken and interpreted. The same holds true for stage speech, usually the spoken realisation of a dramatic text (in different forms of stage arts, it can also be spontaneous, personal, etc.). It represents speech that attempts to give an impression of being created on the spot, natural and spontaneous, even though it is not. Actors in a theatre performance speak as if uttering the text for the first time. In reality, their speech is prepared in advance according to textual material. All speech perspectives (genre; vocabulary; orthoepic elements – accent, pronunciation; prosodic features – tempo, volume, intonation, pauses, register, etc.; also, the visual non-verbal speech elements – facial expressions, gestures) are entirely deliberate, conscious and fixed.

In the past, the study of stage speech used various methods, especially auditory-recognition analysis (Podbevšek, Sušec Michieli). Recently, auditory-recognition analysis has been combined with computer acoustic (phonetic) analysis (e.g., Praat). A model for researching (Žavbi, 2017) Slovenian stage speech was prepared based on stage speech research by Croatian researchers (Škarić, Varošanec-Škarić, Vrban Zrinski) and Slovenian media speech research (Tivadar, Huber).

Historically, detailed scientific analyses became possible only with the arrival of technical capabilities for repeated listening – recordings. Before then, analysis was only possible by viewing and listening to individual performances, thus, speech was described more holistically, even from memory or based on reviews. With repeated listening made possible by audio recordings, listening analysis became very detailed. Video recordings allowed the simultaneous study of the visual accompaniments of

speech. However, contemporary study methods demand more than accessible performance recordings. For an acoustic analysis that brings greater objectivity to research and confirms auditory perceptions, we need suitable computer programmes, such as Praat, which is accessible because it is free and easy to use. For a quality analysis, we need a thorough complex interpretation of the results or positioning of the gathered data in context (that of the drama text and entire staging as well as that of external factors, such as the auditorium size, the social context, the playwright's artistic context or time, etc.).

The paper aims to show the development of researching artistic (stage) speech in Slovenia. It focuses on the contemporary model of research, combining auditory-recognition and acoustic analysis. It defines the method's advantages and presents a research strategy that studies stage speech in relation to the dramatic text and as one of the staging factors. The interdisciplinary research results contextualise and give meaning to studying the artistic phenomena of stage speech, which combines science and art.

Keywords: language, speech, artistic speech, acoustic analysis, Praat

INFRASTRUKTURA ZA RAZISKAVE GOVORA V HUMANISTIKI IN JEZIKOVNIH TEHNOLOGIJAH: ZBORNIK POVZETKOV

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

Zbornik povzetkov s 6. mednarodne znanstvene konference Slavistični znanstveni premisleki prinaša povzetke uveljavljenih domačih in tujih raziskovalcev in raziskovalk govora in govorjenega diskurza s področij humanistike, družboslovja in jezikovne tehnologije. Zbrani povzetki prinašajo pregled aktualnega stanja in perspektiv uporabe govornih virov, predstavljene in obravnavane so tematike, vezane na različne vrste govorjenega diskurza (npr. politični, narečni, gledališki), na kontekst govorjenega diskurza (npr. sociolingvistika, lingvistika variacij), na pomen in vlogo večpredstavnosti v govorni komunikaciji (npr. pri literarnem branju, v političnem diskurzu). Predstavljene bodo še tematike, vezane na snemanje govora in govorne komunikacije, standarde transkribiranja in vrste korpusnega označevanja, z jezikovno tehnološkega vidita pa še tematike o orodjih za analizo govora in govorne komunikacije. Zbrane povzetke povezuje ugotovitev, da razumevanje zakonitosti govorjenega jezika pomembno dopolnjuje jezikovna spoznanja sploh.

Ključne besede:
govorjeni jezik,
govorjeni diskurz,
jezikovne tehnologije,
humanistika,
družboslovje

Univerzitetna založba Univerze v Mariboru

DOI https://doi.org/10.18690/um.ff.5.2023
ISBN 978-961-286-735-5

INFRASTRUCTURE FOR SPEECH RESEARCH IN THE HUMANITIES AND LANGUAGE TECHNOLOGIES: BOOK OF ABSTRACTS

MIRA KRAJNC IVIČ (ED.)
University of Maribor, Faculty of Arts, Maribor, Slovenia
mira.krajnc@um.si

The Book of Abstracts of the 6th International Scientific Conference Slavic Scientific Reflections brings together abstracts from renowned national and international researchers in the fields of speech and spoken discourse from the humanities, social sciences and language technologies. The collected abstracts provide an overview of the current state of the art and perspectives on the use of spoken resources, present and discuss topics related to different types of spoken discourse (e.g. political, dialect, theatrical), the context of spoken discourse (e.g. sociolinguistics, linguistics of variation), the meaning and role of multimedia in spoken communication (e.g. sociolinguistics, linguistics of variation), the role of multimedia in spoken discourse (e.g. literary reading and in political discourse). Topics related to speech and speech communication recording, transcription standards and types of corpus annotation will also be presented, as well as, from a linguistic-technological point of view, topics on speech and speech communication analysis tools. The collected abstracts are connected by the finding that understanding the laws of spoken language significantly complements linguistic knowledge in general.

Keywords:
spoken language, spoken discourse, language technologies, humanities, social sciences

https://doi.org/10.18690/um.ff.5.2023 DOI
978-961-286-735-5 ISBN
University of Maribor Press

18. 5.–19. 5. 2023
Maribor, Slovenija