JEZIKOVNI MODELI ZA PRIPRAVO GOVORNEGA KORPUSA: PROGRAMI ZA PREPOZNAVANJE GOVORA

TEODOR PETRIČ
Univerza v Mariboru, Filozofska fakulteta, Maribor, Slovenija
teodor.petric@um.si

DOI
https://doi.org/10.18690/um.ff.4.2024.9

ISBN
978-961-286-882-6

V preteklem desetletju, še posebej v zadnjih petih letih po uveljavljanju velikih jezikovnih modelov, ki temeljijo na arhitekturi transformerjev (pretvorbenih modelov), smo dobili vrsto programskih orodij, ki pospešujejo ustvarjanje večplastnih jezikovnih gradiv. Preizkušali smo programska orodja za prepoznavanje in pretvorbo govora v pisno obliko (tj. orodja *Razpoznavalnik, Microsoft Word Prepiši, Vosk/Kaldi* in *OpenAI Whisper*), ki so ključni za pospešeno ustvarjanje govornih korpusov. Uporabljali smo vrsto meril, ki zadevajo preprostost uporabe, časovni prihranek, morebitne stroške, zagotavljanje anonimnosti govorcev in različne vidike kakovosti pretvorbe (deleže besednih napak, število zamenjav, vstavitev in izpustov). Orodja za pretvorbo govora v pisno obliko so vidno napredovala, vendar bi si vsekakor želeli, da bi lahko izhodne formate programov prilagajali posameznim raziskovalnim potrebam, npr. vključitev diskurznih označevalcev (npr. tako imenovanih »mašil«) ali dejansko izgovorjenih skrčenih besednih oblik v zapis.

Ključne besede:
pretvorbeni modeli,
govorni korpus,
kakovost pretvorbe,
programska orodja,
raziskovalne potrebe

Univerzitetna založba Univerze v Mariboru

DOI
https://doi.org/
10.18690/um.ff.4.2024.9

ISBN
978-961-286-882-6

Keywords:
transformer models,
spoken corpora,
conversion quality,
software tools,
research needs

University of Maribor Press

Language Models for Spoken Corpus Preparation: Speech Recognition Software

Teodor Petrič
University of Maribor, Faculty of Arts, Maribor, Slovenia
teodor.petric@um.si

In the last decade, particularly in the last five years after the emergence of large language models based on transformer architectures, we have seen the development of a number of software tools that accelerate the creation of multi-layered corpora. We have tested software tools for speech recognition and conversion to written form (i.e. tools such as Razpoznavalnik, Microsoft Word Dictate, Vosk/Kaldi and OpenAI Whisper), which are crucial for accelerating the creation of spoken corpora. We have employed various criteria concerning ease of use, time-saving features, potential costs, ensuring speaker anonymity and various aspects of conversion quality (e.g. word error rates, number of substitutions, insertions and deletions). While the tools for converting speech to written form have made considerable progress, we would certainly wish for the ability to customize the output formats of these programmes to meet individual research needs, e.g. including discourse markers (such as the so-called 'fillers') or the actual spoken contracted word forms in the transcription.

1 Uvod

1.1 Pretvorbeni modeli

V preteklem desetletju, še posebej v zadnjih petih letih po uveljavljanju velikih jezikovnih modelov, ki temeljijo na arhitekturi *transformerjev* (v nadaljevanju tudi v poslovenjeni obliki: *pretvorbeni modeli*), smo dobili vrsto programskih orodij, ki pospešujejo ustvarjanje večplastnih jezikovnih gradiv. Pretvorbeni modeli imajo potencial za velike spremembe v družbi, med drugim tudi v načinu raziskovalnega dela in poučevanja.

Pretvorbeni modeli (Vaswani et al. 2017: 2–6) so vrsta arhitekture nevronskih mrež, temeljijo pa na mehanizmih pozornosti (*attention mechanism*) in samopozornosti (*self-attention mechanism*). *Mehanizem pozornosti* modelu omogoča osredinjanje na najpomembnejše dele vzorca in zanemarjanje manj pomembnih delov vzorca: npr. pri prevajanju stavka mehanizem pozornosti določi, katera beseda v izvornem jeziku je najbolj povezana z besedo v ciljnem jeziku. Mehanizem pozornosti pomaga izboljšati kakovost prevoda in razumevanje jezika.

Mehanizem *samopozornosti* je posebna vrsta mehanizma pozornosti, ki povezuje različne položaje ene same sekvence in nato izračuna vektorsko predstavitev (reprezentacijo) sekvence. Modelu omogoča, da oceni pomembnost različnih besed v stavku in dinamično prilagodi njihov vpliv na izhod. Pri razumevanju naravnih jezikov je to pomembna lastnost, saj se pomen besede lahko spremeni glede na sobesedilo znotraj stavka ali besedila. Zaradi opisanih lastnosti so pretvorbeni modeli primernejši za vzporedne računske operacije in analizo odvisnosti medsebojno oddaljenih elementov kot druge nedavno razvite programske arhitekture (npr. rekurentni ali konvolucijski modeli).

1.2 Veliki jezikovni modeli

Veliki jezikovni modeli (*Large language models*, *LLM*) so pretvorbeni modeli, ki so izurjeni na osnovi velikih količin besedilnih podatkov (npr. celotne Wikipedije ali več). Naučijo se lahko slovnice, pravopisa, besedišča in drugih znanj, ki jih potrebujejo za uporabo jezika. Prilagodimo jih lahko za opravljanje posebnih nalog, kot so povzemanje besedila, strojno prevajanje, odgovarjanje na vprašanja, pisanje besedila, iskanje podatkov in idej, za izdelovanje aplikacij in orodij za učenje in

poučevanje jezikov, za razpravljanje o različnih družbeno relevantnih temah in drugo. LLM lahko ustvari koherentno in naravno zveneče besedilo, ki ga pogosto ni mogoče razlikovati od človeškega besedila (npr. GPT-3.5/4, BERT, T5). Zaradi vseh teh lastnosti so izredno zanimivo orodje pri pripravi in ustvarjanju jezikovnih gradiv ter pri preučevanju in poučevanju naravnih jezikov.

Aktualni jezikovni modeli omogočajo različne postopke za analizo stavčnih elementov ali celotnih besedil v jezikovnih gradivih: npr.

- glasoslovje, pravopis: samodejno prepoznavanje (ASR) in zapisovanje govora (STT - speech to text),
- oblikoslovje: označevanje besednih vrst (POS),
- skladnja: odvisnostna razmerja, besedni vrstni red,
- NER: prepoznavanje poimenovanih entitet in razmerja med entitetami,
- Q&A: vprašanja uporabnika in odgovori programa o jezikovni tematiki, npr. članka ali knjige,
- povzemanje: ustvarjanje izvlečkov in povzetkov člankov ali knjig,
- koreferenčnost: ugotavljanje soodnostnosti in sopomenskosti stavčnih prvin,
- semantika: odkrivanje semantičnih sprememb skozi čas,
- pragmalingvistika: analiza čustev (sentimenta) in čustvenosti govora,
- avtomatizacija: programski agenti GPT (GPT agents) uporabljajo LLM za avtomatizacijo različnih nalog na podlagi cilja, ki ga je v naravnem jeziku določil človeški uporabnik.

V nadaljevanju sestavka se osredinjamo na preizkus in oceno programskih orodij za samodejno prepoznavanje in pretvorbo govora v pisno obliko (*Automatic Speech recognition*, ASR).

2 Preizkus orodij ASR

2.1 Motivacija preizkusa

Če želimo z govornimi viri izvesti podobne raziskave kot s pisnimi (gl. prejšnji odsek), je ključna hitra in uspešna pretvorba govora v pisno obliko. Modeli za samodejno razpoznavanje govora nam obetajo precejšen časovni prihranek pri zapisovanju govornih prispevkov.

Modeli ASR lahko zajamejo tako globalne kot lokalne značilnosti govornih virov, predhodna priprava na velikih količinah neoznačenih govornih podatkov pa jim omogoča usvajanje splošnih akustičnih vzorcev in znanj.

Prilagodimo jih lahko že na osnovi majhne količine označenih podatkov za določeno nalogo (*fine-tuning*), kot je npr. prepisovanje telefonskih klicev. Cilj, ki smo si ga zastavili, tj. ustvarjanje multimodalnega govornega korpusa, je kompleksnejši. Prav nam bi prišla programska oprema za pospeševanje zamudnega zapisovanja in urejanja govornih prispevkov.

2.2 Izbor orodij ASR

Programskih orodij za samodejno prepoznavanje govora je kar nekaj, večinoma pa imajo določene omejitve ali pa imamo določene zadržke do njihovih zahtev (npr. previsoki stroški, posnetek mora biti kratek, anonimnost ni zagotovljena, med jeziki ni slovenščine …).[1]

Preizkusili smo nekaj izmed razpoložljivih orodij za pretvorbo govora v pisno obliko, ki imajo razmeroma preprost vmesnik za delo z njimi:[2]

- *Razpoznavalnik*, različica *e2e*,
- *Razpoznavalnik*, različica *Kaldi*,
- *Microsoft Word* z vstavkom *Prepiši* (*Transcribe*),
- *Vosk*/*Kaldi*, vtičnik v programu *Subtitle Edit*,
- *Whisper* in različica *Whisperx*.

Razpoznavalnik (tj. splošni razpoznavalnik *e2e*) je program za prepoznavanje govora v slovenščini. Razvil in trži ga konzorcij, ki ga sestavljata Zemanta in Univerza v Ljubljani (Lebar Bajec et al. 2022). Gre za model ASR, ki temelji na tehnologiji NVIDIA NeMo, ki je okvir za razvoj modelov AI za konverzacijske aplikacije. Model je bil treniran na naboru podatkov *Artur* (Verdonik et al. 2023), ki vsebuje 630 ur transkribiranega govora v standardni slovenščini. Program je dostopen na

[1] Na spletni strani paperswithcode (https://paperswithcode.com/datasets?task=speech-recognition) so tudi povezave do podatkovnih nizov za učenje modelov. (31. 8. 2023).
[2] Zelo obetaven projekt je *FAIRSEQ* podjetja *Meta (prej: Facebook)*, ki pa je trenutno dostopen predvsem programerjem, zato ga v našem izboru programskih orodij ni.

spletni strani slovenscina.eu[3], kjer lahko uporabniki naložijo zvočne posnetke in dobijo njihovo prepisano besedilo. Program lahko prepoznava govor do dolžine 300 sekund. Repozitorij programa je na GitHubu[4], kjer je mogoče najti več informacij o modelu in njegovem delovanju.

Kaldi (Povey et al. 2011), program za prepoznavanje govora v slovenščini, je dostopen na spletni strani slovenscina.eu[5], kjer lahko uporabniki izberejo med dvema modeloma ASR: splošnim razpoznavalnikom *Kaldi* in razpoznavalnikom *e2e*. Splošni razpoznavalnik *Kaldi* je model ASR, ki temelji na odprtokodnem orodju *Kaldi*, ki je namenjeno raziskovalcem s področja prepoznavanja govora. Program lahko prepoznava govor do dolžine 300 sekund. Več informacij o projektu *Kaldi* na spletnih straneh *Kaldi*[6] in na *GitHubu*[7].

Microsoft Office 365 in *Word Prepiši (Transcribe)*: program za prepoznavanje govora je del storitve Microsoft Office 365, ki jo je razvil in trži Microsoft. Gre za funkcijo *Word Prepiši* (*Transcribe*), ki omogoča pretvorbo govora v besedilo. Program uporablja model za prepoznavanje govora, ki temelji na storitvi Azure Cognitive Services AI, ki je platforma za razvoj modelov umetne inteligence za konverzacijske aplikacije. Model je bil naučen na velikih količinah transkribiranega govora v več kot 100 jezikih. Program lahko prepoznava govor iz različnih virov, kot so mikrofoni, zvočne datoteke ali shramba blob. Spletna različica programa lahko prepoznava govor do dolžine 300 sekund, omejitev za nameščeno različico programa pa je velikost naložene zvočne datoteke (manj kot 300 MB). Več informacij o programu za prepoznavanje govora lahko je na spletnih straneh podjetja *Microsoft* (npr. *Microsoft* support[8] in *Azure*[9]).

Subtitle Edit je brezplačen in priljubljen urejevalnik podnapisov za video posnetke, ki v različici 3.6.13 podpira uporabo modelov ASR (*Vosk*/*Kaldi* in *Whisper*) kot vtičnikov.

[3] https://slovenscina.eu/razpoznavalnik (31. 8. 2023).
[4] https://github.com/clarinsi/Slovene_ASR_e2e (31. 8. 2023).
[5] Prav tam.
[6] https://www.kaldi-asr.org/doc/about.html (31. 8. 2023).
[7] https://github.com/kaldi-asr (31. 8. 2023).
[8] https://support.microsoft.com/en-us/office/transcribe-your-recordings-7fc2efec-245e-45f0-b053-2a97531ecf57 (31. 8. 2023).
[9] https://azure.microsoft.com/en-us/products/ai-services/speech-to-text (31. 8. 2023).

Vtičnik *Vosk/Kaldi* v programu *Subtitle Edit*: program za prepoznavanje govora je vtičnik *Vosk/Kaldi*. Vtičnik *Vosk/Kaldi* je razvil in trži *Alpha Cephei*, ki je podjetje, specializirano za rešitve na področju prepoznavanja govora. Vtičnik *Vosk/Kaldi* uporablja model ASR, ki temelji na odprtokodnem orodju *Kaldi*, ki je namenjeno raziskovalcem s področja prepoznavanja govora. Model je bil treniran na različnih naborih podatkov, odvisno od podprtega jezika. Vtičnik *Vosk/Kaldi* lahko prepoznava govor v več kot 30 jezikih in variantah, med njimi slovenščine še ni. Program lahko prepoznava govor iz različnih virov, kot so mikrofoni, zvočne datoteke ali video posnetki. Program nima časovnih omejitev za dolžino posnetkov za pretvorbo. Več informacij o programu za prepoznavanje govora in jezikovnih modelih lahko je na *Alphacephei*[10] in na *GitHubu*[11].

Whisper je razvil in trži *OpenAI*, raziskovalna organizacija, ki se ukvarja z umetno inteligenco. *Whisper* je brezplačen model ASR, ki temelji na tehnologiji transformerjev, ki so vrsta nevronskih mrež, ki uporabljajo mehanizem samopozornosti za učenje iz sekvenc podatkov. Model je bil treniran na velikih količinah transkribiranega govora v več kot 50 jezikih in variantah. Program lahko prepoznava govor v angleščini, francoščini, nemščini, kitajščini, španščini, slovenščini in drugih jezikih. Sistem za prepoznavanje govora je bil naučen na 680.000 urah večjezičnega in večopravilnega gradiva, zbranega v medmrežju. Po mnenju podjetja vodi uporaba tako velikega in raznolikega nabora podatkov do izboljšane robustnosti pri prepoznavanju različnih naglasov, ozadnega šuma in tehničnega jezika. Poleg tega omogoča prepisovanje v več jezikih, pa tudi prevajanje iz teh jezikov v angleščino. Program lahko prepoznava govor iz različnih virov, kot so mikrofoni, zvočne datoteke ali video posnetki. Omejitve glede dolžine posnetka program nima. Modele *Whisper* podjetja *OpenAI* lahko uporabljamo preko različnih vmesnikov ali programov: kot vtičnik programa *Subtitle Edit* ali v *ukazni vrstici* ali v skriptu računalniškega jezika *Python* na lokalnem računalniku ali v oblačnih storitvah kot npr. *Google Colab*. Whisper podpira tudi uporabo grafičnega pospeševalnika, kar omogoča bistveno hitrejše prepoznavanje govora. Več informacij o programu za prepoznavanje govora lahko je na spletnih straneh *Whisper*[12] in na *GitHubu*.[13]

[10] https://alphacephei.com/vosk/ (31. 8. 2023).
[11] https://github.com/alphacep/vosk-api (31. 8. 2023).
[12] https://openai.com/research/whisper (31. 8. 2023).
[13] https://github.com/openai/whisper (31. 8. 2023).

2.3 Merila ocenjevanja

Uporaba orodja ASR naj jezikoslovcu čimbolj olajša pretvorbo govorjenih virov in njihovo implementacijo v multimodalni korpus, ki ga želi sestaviti s programsko opremo (npr. s programom *Elan* ali drugimi korpusnimi orodji). Merila, po katerih smo ocenjevali dosežke programskih orodij:

- cenovno ugodna ali celo brezplačna uporaba,
- preprost uporabniški vmesnik,
- primerna hitrost (pretvorbe, shrambe, nalaganje in povezanih opravil),
- primerne strojne zahteve za širši krog uporabnikov,
- zagotovljeno varstvo osebnih podatkov govorcev na posnetkih,
- uporabnost za več jezikov (poleg slovenščine),
- prilagodljivost govornim virom slabše kakovosti (brez dodatnega programskega izboljševanja zvočnega posnetka),
- prilagodljivost neknjižnemu govoru,
- različne vhodne oblike (oblike zvočnih in video datotek),
- izpis pretvorjenega gradiva v različne oblike,
- časovni žigi govornih prispevkov (timestamps),
- sposobnost razlikovanja govorcev (speaker diarization, separation, identification),
- natančnost prepoznavanja govorjenih besed (knjižna in neknjižna izreka), kar preverjamo z metriko WER (word error recognition) in preizkusi urejanja ustvarjenega pisnega gradiva,
- samodejno postavljanje ločil,
- uporabnost za dalj časa trajajoče zvočno ali filmsko gradivo,
- programska razširljivost ali prilagodljivost orodja.

2.4 Zvočno in filmsko gradivo

Zvočno gradivo v preizkusih za ocenjevanje uporabnosti programa za zgoraj predstavljen namen je bilo raznoliko sestavljeno iz:

- slovenska zborna izreka (Šeruga-Prek et al. 2004, 5 minut),
- vsakdanji pogovor (smalltalk) v slovenščini (*Sosedska neljubezen*, 2 min.),

- pogovor z otrokom v slovenščini (5 in 90 min.),
- pogovori z otrokom v nemščini (90 in 120 min.),
- razprava v nemški televizijski oddaji *13 Fragen* (tema: Instagram, 35 min.),
- Fantom iz opere v nemščini (pesem, 4 min.),
- Mešanje jezikov - angleščine in španščine (nadaljevanka *Kraljica juga*, > 5 min.).

Večinoma je bilo vsako gradivo preizkušeno z vsakim orodjem dvakrat, izjemoma celo več kot dvakrat ali samo enkrat.

2.5 Preizkus programskih orodij

WER (*Word error rate*, delež besednih napak) in CER (*Character error rate*, delež črkovnih napak) sta kazalnika uspešnosti sistema za samodejno prepoznavanje govora. Merita razliko med referenčnim besedilom in samodejno prepoznanim besedilom s štetjem števila popravkov (zamenjav, izpustov ali vstavitev), ki so potrebni za pretvorbo enega besedila v drugo. WER deluje na besedni ravni, CER pa na črkovni ravni. V primerjavi je treba upoštevati vrsto in dolžino besedilnih zaporedij. Če so kratka in sestavljena iz določenih zaporedij (kot so telefonske številke, številke socialnega zavarovanja itd.), je CER po navadi ustreznejši od WER, če pa so dolga in sestavljena iz povedi (kot npr. knjige, časopisi itd.), je WER po navadi uporabnejši kot CER. Praviloma je WER tri- do štirikrat višji od CER in je premo sorazmeren z njim. Če se izboljšuje CER, se bo izboljšal tudi WER. Vendar ta zveza ni vedno linearna ali dosledna, zlasti pri velikem številu vstavitev.[14]

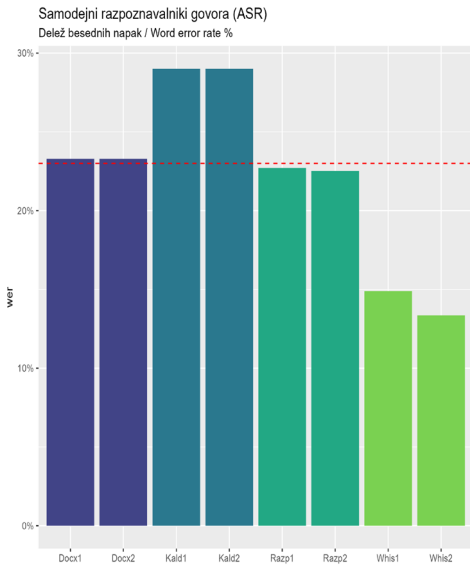
Pri ocenjevanju kakovosti pretvorbe govora smo uporabljali le količino *WER*, ki je po navadi primeren za primerjavo napak na besedilni ravni. Za izračun vrednosti WER smo uporabljali knjižnico *wersim* (Proksch, Wratil, and Wäckerle 2018) v računalniškem jeziku R (R Core Team 2023). Sestavili smo programski skript za izračun količin, sestavo tabel in izpis grafikonov.

[14] Evaluate OCR Output Quality with Character Error Rate (CER) and Word. Towards Science (https://towardsdatascience.com/evaluating-ocr-output-quality-with-character-error-rate-cer-and-word-error-rate-wer-853175297510), (31. 8. 2023).

2.5.1 Slovenska knjižna izreka

Najprej bomo predstavili izide preizkusa orodij z zvočnim posnetkom, ki predstavlja vzorec slovenske zborne izreke (Šeruga-Prek et al. 2004). Izbrali smo peti posnetek na zgoščenki, v katerem slišimo *Ajdo Kalan* izgovarjati posamezne besede, besedne zveze ali kratke povedi o različnih, vsebinsko nepovezanih temah. Zvočni vzorec je bil posnet v studiu RTV Slovenije in zato tudi akustično zelo kakovosten. Za preizkus smo posnetek omejili na prvih pet minut.

Preizkušali smo *Razpoznavalnik* (oba modela: *e2e* in *Kaldi*), OpenAI *Whisper* in Microsoft Word *Prepiši (Transcribe)*. Izide prikazujeta preglednica 1 in diagram 1.[15] Deleži besednih napak (*wer*) so pri vseh orodjih nizki. Najnižje vrednosti je dosegel *Whisper*, čeprav slovenščina ni med jeziki, ki jih najbolje obvlada. Razmeroma tesno mu sledita *Razpoznavalnik* (model *e2e*) in *Word Prepiši*. Najslabše se je odrezal model *Kaldi*.



Slika 1: Slovenska knjižna izreka (deleži besednih napak, WER)
Vir: lasten

[15] Uporabljene kratice: *wer* = delež besednih napak v odstotkih (wer = 100*(ins + del + sub)/word.ref); *sub* = število besednih zamenjav; *ins* = število vstavljenih besed, *del* = število zbrisanih besed; *word.ref* = število pojavnic v referenčnem besedilu; *words.hyp* = število pojavnic v samodejno prepoznanem besedilu; rdeča črtkana črta v diagramu je *mediana*; *Docx* = Microsoft Word Prepiši, *Razp* = Razpoznavalnik, model e2e; *Kald* = Razpoznavalnik, model Kaldi; *Whis* = Openai Whisper; številka za imenom orodja je številka preizkusa

Preglednica 1 prikazuje podrobnejšo sliko in nam približuje vzorce programskega vedenja v procesu prepoznavanja govora, kar bo moč videti tudi v rezultatih sledečih preizkusov. Oba modela *razpoznavalnika* (*e2e* in *Kaldi*) izkazujeta majhno število vstavljenih izrazov (vrednosti *ins* so nizke), *Whisper* pa vstavlja več izrazov kot ostala orodja, teži torej k vstavljanju sobesedilno verjetnih besed, če izgovorjene besede ne prepozna. *Word Prepiši* podobno kot *Razpoznavalnik* (model *e2e*) malokdaj vstavlja izraze, zelo rad pa jih izpusti ali v določenih primerih zamenja.[16]

Tabela 1: Slovenska knjižna izreka (deleži besednih napak, WER)

ASR	wer	sub	ins	del	words.ref	words.hyp
Kald1	29	86	6	58	524	470
Kald2	29	86	6	58	524	470
Razp1	23	64	1	54	524	471
Razp2	23	64	1	53	524	472
Whis1	15	47	10	18	524	516
Whis2	13	47	4	19	524	509
Docx1	23	36	1	85	524	440
Docx2	23	36	1	85	524	440

Vir: lasten

Po pregledu podrobnih rezultatov se je pri prepoznavanju vzorca slovenske zborne izreke najbolje obnesel *Whisper* (najnižji delež besednih napak, malo izpuščenih besed), sledi *Razpoznavalnik* (model *e2e*). *Word Prepiši* pa je v preizkusu prepoznanih besed pristal na tretjem mestu, saj je izpustil več besed kot *Razpoznavalnik* (*e2e*). Četrti model (*Kaldi*) pa zaostaja z razmeroma velikim številom zamenjav in izpustov. Tudi z ozirom na ostala zgoraj navedena merila bomo prihranili največ časa z orodjem *Whisper*, saj omogoča pretvorbo dolgih posnetkov, uporabo grafičnega procesorja za pospeševanje prepoznavanja govora in več izhodnih oblik (npr. tudi podnapise, ki jih zlahka uvozimo v *Elan* za ustvarjanje večsteznega jezikovnega gradiva), idr. Razlika med *Razpoznavalnikom* in *Wordom Prepiši* pa se zmanjšuje v korist Microsoftovemu orodju zaradi (opcionalnega) razlikovanja govorcev in izpisa v različnih izhodnih oblikah (s časovnimi žigi ali brez). Oboje lahko prihrani čas pri ustvarjanju jezikovnega gradiva.

Orodja za prepoznavanje govora spremljajo programi za *vstavljanje ločil*. Ustreznosti vstavljenih ločil nismo preverjali, vendar smo dobili vtis, da *Whisperjeva* programska sestavina za vstavljanje ločil ne zaostaja za *Razpoznavalnikom* (*e2e*). *Microsoftov* program

[16] Microsoft Word Prepiši po navadi zamenja vulgarne izraze z zvezdicami, kar se v evidenci šteje kot zamenjava.

in *Kaldi* sta bila v tem pogledu pogosteje manj dosledna kot *Whisper* ali *Razpoznavalnik* (*e2e*).

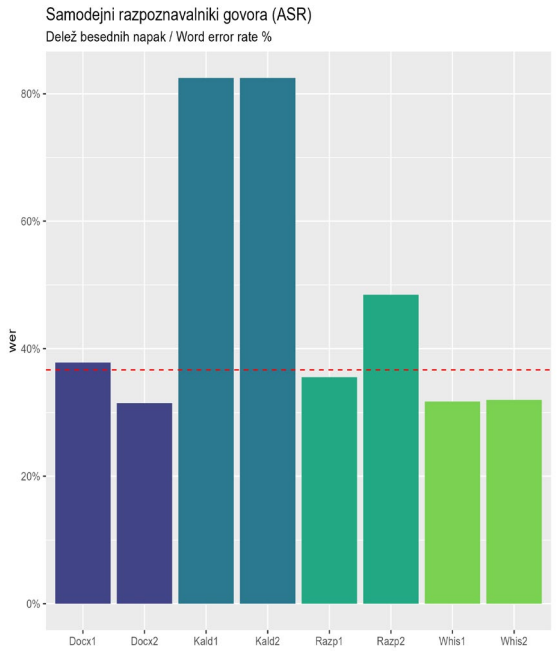
Precejšen delež *Whisperjevih* napak je šlo na račun neupoštevanih pravopisnih pravil. Program je torej pravilno "slišal", ampak izbral napačen grafem. V zvočnem vzorcu je bilo več besed, kjer je bil izgovorjen nezveneči pripornik [s], zapisati pa ga je bilo treba z grafemom <z> (npr. v besedah *razprava, razplet, privez*). *Whisper* je take besede pogosto zapisal z grafemom <s>, kar ustreza nezvenečemu priporniku [s] (npr. kot v hrvaški besedi *rasprava*). Podobno se je *Whisper* tudi pri dvoumnih besedah (npr. *poseg* ali *posek*, *obseg* ali *obsek*) odločil za grafem <k> in ne za <g>. Nekatere zamenjave se nanašajo na oblike, kjer pišemo grafem <l>, izgovarjamo pa [w] (kot npr. v glagolskih oblikah *preletel, obvestil*). Pri glagolskih oblikah je *Whisper* glagole nekajkrat spregal s tematskim samoglasnikom /a/ namesto z /e/ (npr. *vlečajo* namesto *vlečejo*, *preletal* namesto *preletel*, *privedal* namesto *privedel*, *pripomogal* namesto *pripomogel*).

Občasno je prišlo tudi do težave, ali vstaviti črko <j> ali ne (npr. *pomenjska razlika* namesto *pomenska*, *dalnovod* namesto *daljnovod*). Te besedne oblike pričajo o tem, da je program sicer akustično prepoznal besedne oblike, da pa pravopisno ni dovolj podkovan ali da meša pravopisne oblike v slovenščini s tistimi v podobnih slovanskih jezikih.

2.5.2 Slovenski pogovorni jezik

Dvominutni odlomek iz slovenskega filma *Sosedska neljubezen*[17] je vzorec, ki se od prejšnjega razlikuje po več lastnostih: ne govori samo ena oseba, temveč tri (mati, oče, hči najstnica), osebe ne govorijo v zbornem jeziku in zvočna kakovost zaostaja za tistim v prejšnjem posnetku, saj se pogovor odvija v naravnem prostoru v hiši. V ozadju pogovora ni izrazitih ali motečih šumov. Izidi o deležu besednih napak (*wer*) so podobni tistim za zborni jezik: *Whisper* ima nekoliko nižji delež besednih napak kot *Razpoznavalnik* (*e2e*) in *Word Prepiši*, močneje pa zaostaja *Kaldi* (slika 2).

[17] DD Studio produkcija, 2012 (https://www.youtube.com/watch?app=desktop&v=aRot1XH3RDE&embeds_referring_euri=https%3A%2F%2Fmichalekskolky.cz%2F&feature=emb_imp_woyt) (31. 8. 2023)



Slika 2: Slovenski pogovorni jezik (deleži besednih napak, WER)
Vir: lasten

Podrobnejši izidi v preglednici 2 kažejo podobne težnje kot pri razpoznavanju vzorca zborne izreke: *Whisper* dodaja besede, tako da je število vseh besed celo večje kot v referenčnem besedilu. Druga orodja (*Razpoznavalnik e2e*, *Word Prepiši* in *Kaldi*) pogosteje izpuščajo besede.

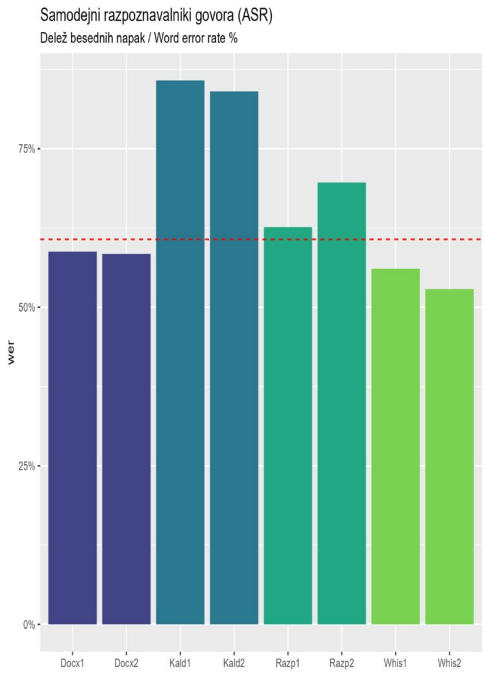
Tabela 2: Slovenski pogovorni jezik (deleži besednih napak, WER)

ASR	wer	sub	ins	del	words.ref	words.hyp
Kald1	82	18	0	307	394	87
Kald2	82	18	0	307	394	87
Razp1	36	73	35	25	394	405
Razp2	48	88	6	91	394	308
Whis1	32	61	35	19	394	414
Whis2	32	62	35	19	394	414
Docx1	38	58	5	82	394	314
Docx2	31	44	8	66	394	330

Vir: lasten

2.5.3 Pogovor s slovenskim otrokom

Glede na to, da sem pred leti sestavil manjši govorni korpus o otroškem govoru in o izsledkih poročal v več znanstvenih prispevkih (npr. Petrič 2016, 2021), me je še posebej zanimalo, kako uspešni so novi modeli za pretvorbo govora v težavnem naravnem pogovornem okolju. Izbrani petminutni zvočni vzorec je v dveh pogledih še težji za razpoznavanje govora: posnetki so slabše kakovosti kot filmski dialog ali posnetek zborne izreke, govorci v ozadju se ne slišijo tako dobro kot sogovornik v vlogi snemalca blizu mikrofona in eden izmed sogovornikov je trileten otrok. Dobili smo vtis, da je otroški glas orodjem glasovno manj domač kot odrasli glasovi. Med učnimi vzorci orodij za razpoznavanje govora najbrž ni takih (ali le malo takih) z otroškimi glasovi. Rezultati kažejo, da je mediana deleža besednih napak (WER) pri vseh preučevanih orodjih precej slabša, z vrednostjo preko 60 %. Po deležu besednih napak rahlo prednjači *Whisper* pred *Wordom Prepiši* in *Razpoznavalnikom* (*e2e*), *Kaldi* pa tudi tu močno zaostaja (slika 3).



Slika 3: Pogovor s slovenskim otrokom (deleži besednih napak, WER)
Vir: lasten

Vedenjski vzorec orodij je podoben zgoraj opisanemu (tabela 3): *Whisper* pogosteje dodaja besede kot ostala orodja (čeprav ne tako močno kot v prejšnjih vzorcih), slednja orodja pa pogosteje izpuščajo besedno gradivo (*Razpoznavalnik e2e* in *Word Prepiši* sta izpustila več kot 300 besed, *Kaldi* celo več kot 600 od 808 v referenčnem besedilu). Vsa orodja (razen *Kaldi*) so v pogovornem gradivu pogosto zamenjevala besede.

Tabela 3: Pogovor s slovenskim otrokom (deleži besednih napak, WER)

comparison	wer	sub	ins	del	words.ref	words.hyp
Kald1	86	50	0	643	808	165
Kald2	84	58	0	621	808	187
Razp1	63	184	8	307	808	502
Razp2	70	177	7	372	808	436
Whis1	56	225	24	181	808	629
Whis2	53	165	8	246	808	563
Docx1	59	152	3	317	808	491
Docx2	58	156	11	297	808	515

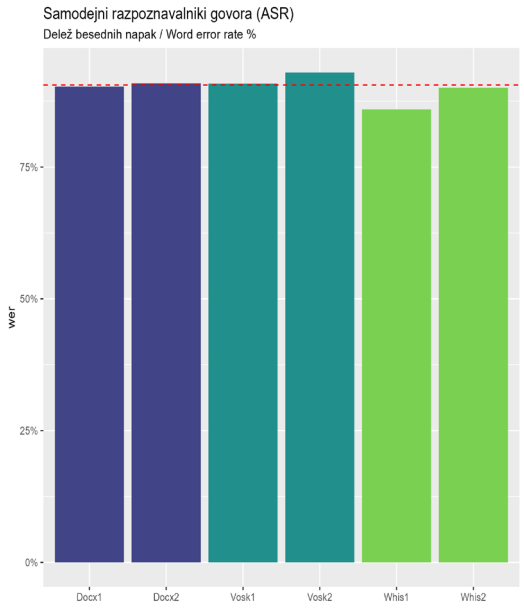
Vir: lasten

Razpoznavanje vsakdanjega spontanega govora, ki je bil posnet v naravnih prostorih, je delal vsem orodjem velike preglavice. Dodatna težava je bilo sodelovanje otroka v pogovoru. Pri zapisovanju vsakdanjih pogovorov v slabših akustičnih razmerah, še posebej, če sodelujejo majhni otroci, žal ne bi prihranili veliko časa. Najbrž primanjkuje tovrstnih zvočnih vzorcev za učenje programskih orodij, kar še posebej velja za slovenščino.

2.5.4 Pogovor z nemškima otrokoma

Preizkus s podobnima zvočnima vzorcema v nemščini, tj. pogovor z nemškima otrokoma *Emely* in *Falko*[18], je prav tako prinesel dokaj slabe izide. Posnetka sta trajala dobrih 60 minut oz. dobrih 90 minut. Občasni hkratni govor dveh ali več oseb, otroški glasovi, neenakomerna glasnost človeških glasov in slabša zvočna kakovost v snemalnem prostoru (igralnici) so botrovali visokim deležem besednih napak (WER > 75 %) pri vseh uporabljenih orodjih, čeprav spet rahlo manj z orodjem *Whisper* (slika 4).

[18] Childes, https://childes.talkbank.org/access/German/Szagun.html (31. 8. 2023).



Slika 4: Pogovor z nemškima otrokoma (deleži besednih napak, WER)
Vir: lasten

Poglavitna značilnost razpoznavanja zvočnih posnetkov z otrokoma je bilo izpuščanje neprepoznanega besednega gradiva (tabela 4). V krajšem pogovoru (ok. 60 minut) je najmanj besed izpuščal *Whisper*, v daljšem pogovoru (ok. 90 minut) pa ga je v tem oziru s tesnim izidom premagal *Word Prepiši*. Podobno razmerje je vidno tudi pri številu zamenjav. Model *Vosk* za nemščino se je glede na število izpustov in zamenjav slabše odrezal kot *Whisper* in *Word Prepiši*. Vendar se to pri deležih napačnih besed skorajda ne pozna.

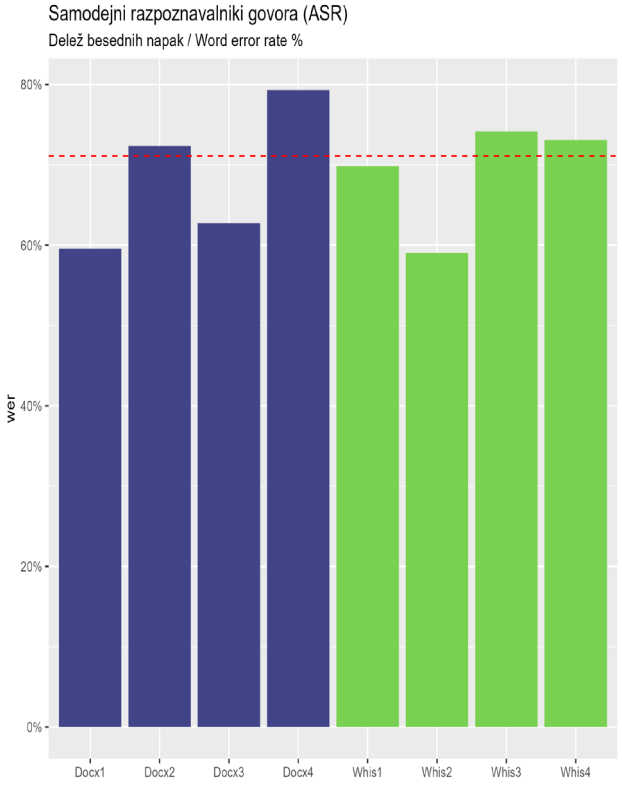
Tabela 4: Pogovor z nemškima otrokoma (deleži besednih napak, WER)

ASR	wer	sub	ins	del	words.ref	words.hyp
Vosk1	91	182	0	1277	1606	329
Vosk2	93	324	0	1742	2223	481
Whis1	86	331	19	1026	1606	598
Whis2	90	596	0	1406	2223	817
Docx1	90	261	0	1189	1606	417
Docx2	91	660	0	1361	2223	862

Vir: lasten

2.5.5 Petje: slovenske pesmi

Pesmi spadajo redko kdaj med učna gradiva orodij za razpoznavanje govora. Tudi za človeka predstavlja prepoznavanje besedila v opernem libretu ali pesmi v muziklu precejšen izziv. Zato lahko pričakujemo slabše izide kot pri prepoznavanju govora. V tem primeru je *Razpoznavalnik* povsem odpovedal, saj je namesto besedila izpisal samo dva vprašaja. Orodji *Whisper* in *Word Prepiši* sta sicer izkazovala višje deleže besednih napak kot pri studijskih ali filmskih posnetkih govora, vendar pa nižje (slika 5 in tabela 5) kot pri posnetkih spontanih pogovorov z otrokom v naravnih prostorih (slika 3). Izbrali smo štiri slovenske pesmi (*Nipke – Popoln lajf, Mi2 – Oda gudeki, N'toko – Seks v mestu, Trkaj – 1 mf 2*).[19]



Slika 5: Pesmi v slovenščini (deleži besednih napak, WER)
Vir: lasten

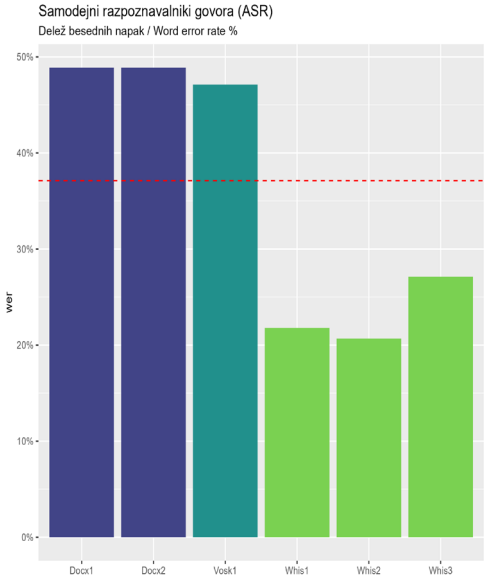
[19] Uporabljena besedila so na naslovu https://www.besedilo.si/ (31. 8. 2023).

Tabela 5: Pesmi v slovenščini (deleži besednih napak, WER)

ASR	wer	sub	ins	del	words.ref	words.hyp
Whis1	70	334	36	151	789	676
Whis2	59	277	26	21	586	586
Whis3	74	276	39	319	867	587
Whis4	73	677	184	178	1643	1545
Docx1	60	250	9	206	789	591
Docx2	72	193	1	229	586	359
Docx3	63	252	37	234	867	665
Docx4	79	543	22	718	1643	929

Vir: lasten

2.5.6 Nemški pogovorni jezik



Slika 6: Nemški pogovorni jezik (deleži besednih napak, WER)
Vir: lasten

Izbrali smo približno dvominutni odlomek *Sind deine Eltern Terroristen?* (Ali sta tvoja starša terorista?) iz nemškega filma *Zweiohrküken*.[20] Pričakovali smo dobre izide kot v odlomku iz slovenskega filma (gl. zgoraj). Filmski odlomek se odvija v baru, v ozadju slišimo klavirsko glasbo, žvenketanje kozarcev in zamolkle pogovore ljudi.

[20] Zweiohrküken je režiral Til Schweiger (2009). Odlomek: https://www.youtube.com/watch?v=yj4Bgij6AsI (31. 8. 2023).

Gre za tri kratke pogovore med moškim in žensko, kot peti govorec pa sodeluje še točaj. Kljub zvočni kulisi so človeški glasovi večinoma razločni. Osebe se pogovarjajo v nemškem pogovornem jeziku brez izrazitega regiolekta. Deleži besednih napak, ki jih izkazujeta *Word Prepiši* in *Vosk*, sta opazno večja kot *Whisperjev* delež (slika 6).

Po podrobnejših podatkih v preglednici sodeč (tabela 6), sta *Vosk* in *Word Prepiši* izpustila znatno več besed kot *Whisper*. Tudi zamenjav je v *Whisperjevem* zapisu manj. Opazen je že zgoraj večkrat opisan "vedenjski" vzorec, da *Whisper* vstavlja več besed kot druga orodja. Če *Whisper* besednega gradiva ne prepozna ali če je računalniški procesor preobremenjen, si *Whisper* izmišljuje besede, besedne zveze ali cele povedi.

Lep primer za programske "halucinacije" je prav iz tega nemškega filmskega odlomka. Mlada Lana vpraša Moritza, ali bi kaj spil z njo (*Trinkst du was?*). Mlademu Moritzu pa se je zavozlal jezik in ni sposoben odgovoriti. Namesto jeclanja je *Whisper* petkrat zapored zapisal poved, da ne želi biti v razmerju z neko osebo (*Nein, ich hab keine Lust auf eine Beziehung.*). Ta poved se v tem filmskem odlomku ne pojavlja nikjer. Zdi se, da je *Whisper* vstavil poved, ki je v tej situaciji možna alternativa, da bi tako preprečil vrzel v besedilu. Glede na obsežnost programa je možno, da je bil računalniški procesor občasno preobremenjen, tako da *Whisper* del zvočnega vzorca ni mogel analizirati. Whisper je najbrž privzeto tako nastavljen za večjo ustvarjalnost, podobno kot ChatGPT. *Whisper* ima stikalo *temperature*. Če to stikalo ali parameter nastavimo na nižjo stopnjo ali celo na ničlo, potem začne tudi *Whisper* izpuščati več besed.

Tabela 6: Nemški pogovorni jezik (deleži besednih napak, WER)

ASR	wer	sub	ins	del	words.ref	words.hyp
Vosk1	47	98	0	114	450	336
Whis1	22	31	32	24	450	458
Whis2	21	24	6	60	450	393
Whis3	27	48	32	38	450	440
Docx1	49	79	21	110	450	353
Docx2	49	71	21	118	450	345

Vir: lasten

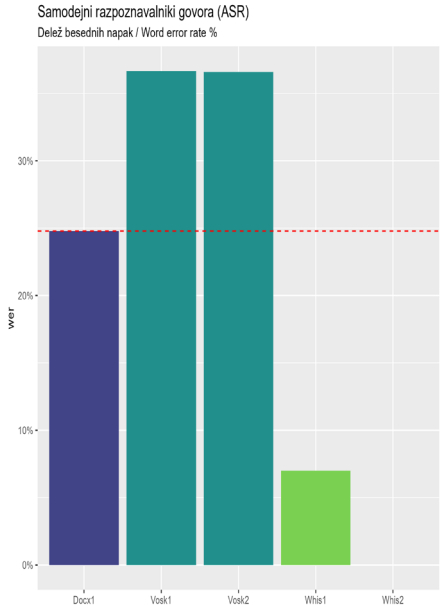
2.5.7 Nemška razprava o instagramu

Nemška pogovorna oddaja *13 Fragen* na temo *instagram*[21] je značilen primer studijskega posnetka: odlična akustična kakovost, vsi govorci so dobro slišni, govorcem se v glavnem ne zatika, ozadnega šuma je malo. Posnetek je trajal 35 minut. Referenčnega besedila v tem primeru nimamo. Zato smo se odločili, da zapise ostalih programov primerjamo z drugim *Whisperjevim* zapisom (*Whis2*). Tudi v tem primeru je opazno, da sta si privoščila *Vosk* in *Word Prepiši* veliko število izpustov (tabela 7 in slika 7).

Tabela 7: Nemška razprava o instagramu (deleži besednih napak, WER)

ASR	wer	sub	ins	del	words.ref	words.hyp
Vosk1	37	1471	174	1223	7958	6890
Vosk2	37	1466	174	1223	7958	6890
Whis1	7	222	121	210	7958	7866
Whis2	0	0	0	0	7958	7958
Docx1	25	836	407	665	7958	7697

Vir: lasten



Slika 7: Nemška razprava o instagramu (deleži besednih napak, WER)
Vir: lasten

[21] ZDF, *13 Fragen*, Filter, Fake und nackte Haut – Macht Instagram uns unglücklich?: https://www.youtube.com/watch?v=-6smnitK4zs (31. 8. 2023).

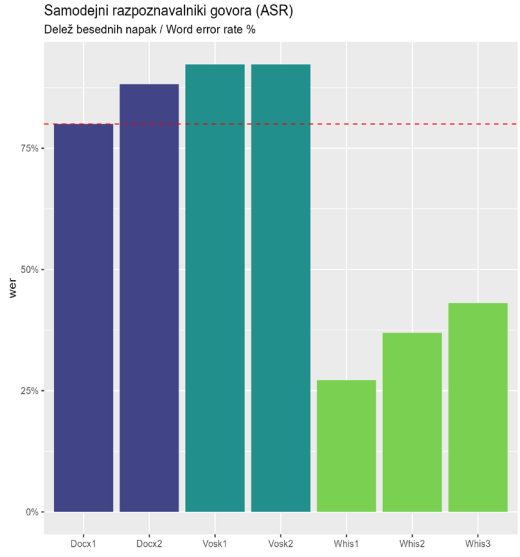
2.5.8 Petje: rock opera v nemščini

Pri zapisovanju slovenskih pesmi so se vsa orodja približno enako slabo odrezala ali celo odpovedala. Drugačno sliko vidimo na primeru nemške različice rock opere *Fantom iz opere*.[22] Orodja smo preizkušali z naslovno pesmijo opere v nemščini. *Whisper* je izkazoval bistveno nižji delež besednih napak kot *Vosk* in Word *Prepiši*. Iz tabele 8 je razvidno, da je to povezano ali z večjim število zamenjav, ali v drugih primerih z večjim številom izpustov.

Tabela 8: Petje v rock operi v nemščini (deleži besednih napak, WER)

ASR	wer	sub	ins	del	words.ref	words.hyp
Vosk1	92	35	0	145	195	50
Vosk2	92	97	0	83	195	112
Whis1	27	30	6	15	195	186
Whis2	37	36	2	31	195	167
Whis3	43	51	3	26	195	173
Docx1	80	51	6	86	195	122
Docx2	88	90	1	80	195	115

Vir: lasten



Slika 8: Petje v rock operi v nemščini (deleži besednih napak, WER)
Vir: lasten

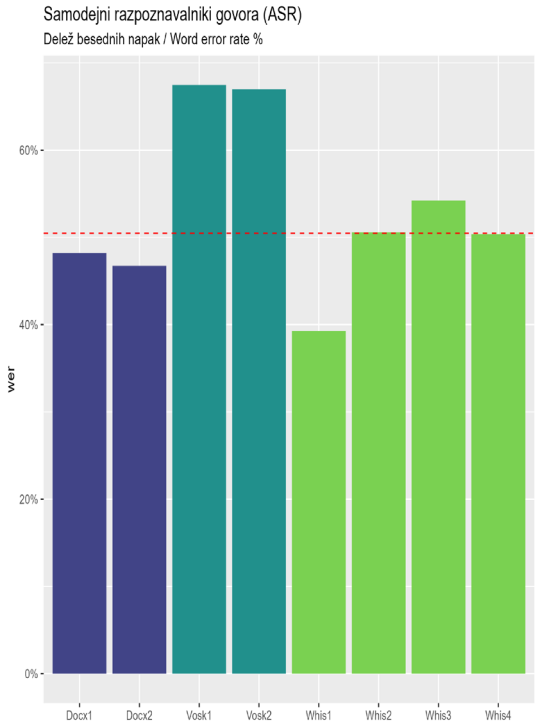
[22] https://www.youtube.com/watch?v=HJPxwneN-wk (31. 8. 2023).

2.5.9 Preklapljanje med jeziki

Tabela 9: Preklapljanje med jeziki (deleži besednih napak, WER)

ASR	wer	sub	ins	del	words.ref	words.hyp
Vosk1	67	74	0	206	415	209
Vosk2	67	73	0	205	415	210
Whis1	39	69	19	57	415	391
Whis2	51	93	17	83	415	360
Whis3	54	77	63	30	415	480
Whis4	50	86	21	84	415	362
Docx1	48	54	6	134	415	293
Docx2	47	54	6	134	415	287

Vir: lasten



Slika 9: Preklapljanje med jeziki (deleži besednih napak, WER)

Vir: lasten

Preizkušali smo, kako se orodja za prepoznavanje govora obnesejo, če osebe govorijo v različnih jezikih. Tovrstni izidi so še posebej zanimivi za raziskovalce, ki preučujejo preklapljanje med jeziki (code switching ipd.) in slenga. Glede na to, da so mnogi razpoznavalniki najuspešnejši v nekaterih svetovno bolj razširjenih jezikih

(v angleščini, španščini, francoščini, nemščini, …) smo izbrali odlomek iz nadaljevanke *Kraljica Juga* (*Queen of the South*[23]), in sicer več kot sedem minut iz desete epizode pete sezone. V tej nadaljevanki je značilno mešanje angleščine in španščine. Prevladuje sicer angleščina, španščina pa prevzame pobudo, ko gre za čustveno napete trenutke in preklinjanje. Glede deleža besednih napak *Vosk* zaostaja za *Wordom Prepiši* in *Whisperjem* (tabela 9 in slika 9). *Whisper* je v enem preizkusu prednjačil pred *Wordom Prepiši*, v treh preizkusih pa je imel podobno slabe izide kot *Word Prepiši*. Deleži napak so bili pri vseh orodjih razmeroma visoki, ker smo izbrali odlomek, med katerim se je odvijal dvoboj med mehiškima akterjema in med katerim se je poleg preklinjanja in stokanja slišalo tudi pokanje strelnega orožja in drugi šumi.

2.5.10 Povzetek preizkusov

Tabela 10: Povzetek lastnosti programskih orodij ASR

Merila	Zemanta/UL Razpoznavalnik	Microsoft Word Prepiši	Kaldi Vosk/Kaldi	OpenAI Whisper
Brezplačna uporaba	Da	Komercialna naročnina	Da	Da
Preprost vmesnik	Da	Da	Da (Subtitle Edit)	Da (Subtitle Edit)
Primerna hitrost	Da (strežnik)	Da (strežnik)	Da	Da (z GPU, pospeševalnikom)
Zmerne strojne zahteve	Da	Da	Da	Odvisno od velikosti modela
Varstvo osebnih podatkov	Neznano	Neznano	Da (lokalna uporaba)	Da (lokalna uporaba)
Uporabnost za več jezikov	Ne (samo slovenščina)	Da	Da (ni slovenščine)	Da
Prilagodljivost ob slabšem zvoku	Ne	Ne	Ne	Omejeno
Natančnost (knjižni jezik)	Visoka (knjižni jezik)	Visoka (knjižni jezik)	Zmerno visoka	Visoka (knjižni jezik)
Natančnost (neknjižni jezik)	Zmerno visoka	Zmerno visoka	Pod povprečjem	Zmerno visoka
Različne vhodne oblike	Mikrofon, zvočni vzorec	Mikrofon, zvočni in video vzorec	Zvočni vzorec in video	Zvočni vzorec in video
Izpis v različne oblike	Ne	Da	Omejeno (podnapisi)	Da
Časovni žigi	Ne	Da (z napakami)	Da	Da (z napakami)

[23] Ameriška nadaljevanka, ki sta jo razvijala M.A. Fortin in Joshua John Miller. Peta sezona je bila prvič predvajana 2021.

Merila	Zemanta/UL Razpoznavalnik	Microsoft Word Prepiši	Kaldi Vosk/Kaldi	OpenAI Whisper
Razlikovanje govorcev	Ne	Da (z napakami)	Ne	Omejeno (Python in Whisperx)
Samodejno postavljanje ločil	Da	Da	Omejeno uporabno	Da
Uporabnost za daljše gradivo	Ne (do 300 sekund)	Omejeno (do 300 MB, 60 minut)	Da	Da (tudi več vzorcev zapored)
Programska prilagodljivost	Ne	Ne	Omejeno	Da (Python in Whisperx, Google Colab z GPU)

Vir: lasten (stanje 30. 4. 2023)

Tabela 11: Srednje vrednosti orodij ASR pri zapisovanju slovenskega govorjenega besedila[24]

ASR	wer_m	sub_m	ins_m	del_m	quo_m
Docx	58	90	6	134	0.71
Kald	82	54	0	307	0.23
Razp	42	80	6	72	0.84
Whis	41	73	20	48	0.94

Vir: lasten

Tabela 12: Srednje vrednosti orodij ASR pri zapisovanju nemškega govorjenega besedila

ASR	wer_m	sub_m	ins_m	del_m	quo_m
Docx	58	90	6	134	0.71
Vosk	67	98	0	206	0.51
Whis	41	73	20	48	0.94

Vir: lasten

3 Sklep

Preizkušali smo več programskih orodij za prepoznavanje in pretvorbo govora v pisno obliko, ki naj bi skrajšala čas, potreben za ustvarjanje govornih korpusov. Uporabljali smo vrsto meril, ki zadevajo preprostost uporabe, časovni prihranek, morebitne stroške, zagotavljanje anonimnosti govorcev in različne vidike kakovosti pretvorbe. V sklepu strnemo nekaj izmed rezultatov, ki izpostavljajo prednosti ali slabosti pretvorbe slovenskega govora v pisno obliko:

[24] Kratice: wer_m, sub_m, ins_m, del_m = mediana deležev besednih napak in števil zamenjav, vstavitev in izbrisov vsakega orodja za prepoznavanje govora, quo_m = mediana količnika med številom besed prepoznanega besedila in referenčnega besedila.

- natančnost prepoznavanja govora je v slovenščini slabše kot v angleščini, španščini ali nemščini, kar je mogoče povezovati z razmeroma majhnim in premalo raznovrstnim zvočnim gradivom v slovenščini za učenje modelov ASR;
- slabše prepoznavanje v slovenščini je opaznejše, kadar gre za pogovorni, otroški jezik, zapeto besedilo ali posnetke slabše zvočne kakovosti;
- postavljanje ločil je na zadovoljivi ravni (velja za: *Razpoznavalnik e2e*, *Word Prepiši*, *Whisper*);
- časovni žigi so netočni (razen pri *Vosk/Kaldi*);
- razlikovanje govorcev je sicer programsko mogoče, vendar ne deluje prav dobro (*Word Prepiši*, *Whisper*),
- najenostavnejše je razlikovanje govorcev z *Word Prepiši*;
- največ različnih izhodnih formatov ponuja *Whisper*;
- vhodna oblika je po navadi zvočna datoteka, vendar *Whisper, Word Prepiši in Vosk/Kaldi* sprejemajo tudi video posnetke;
- enostavna in hkrati vsestransko uporabna je kombinacija programov *Subtitle Edit* in *Whisper(x)*,
- največjo hitrost pretvorbe dosežemo s programom *Whisper(x)* na grafičnem procesorju (na lokalnem prenosniku ali npr. pri *Google Colab*),
- zaporedno pretvorbo več zvočnih datotek je mogoče s programom *Whisper*;
- na internetu je več predlog za ustvarjanje računalniških skriptov (npr. v Pythonu) za pretvorbo govora s programom *Whisper*;
- v različnih preizkusih je *Whisper* v povprečju naredil najmanj besednih napak;
- največ jezikov (tudi slovenščino) poznata *Word Prepiši* in *Whisper*.

Novejša orodja za pretvorbo govora v pisno obliko so vidno napredovala. Kljub vsemu napredku bi si jezikoslovci vsekakor želeli, da bi lahko programe za samodejno prepoznavanje govora in pretvorbo v pisno obliko prilagajali posameznim raziskovalnim potrebam (npr. vključitev diskurznih označevalcev ali dejansko izgovorjenih skrčenih besednih oblik v zapis).

Literatura

Iztok, LEBAR, Marko BAJEC, Žan BAJEC, Mitja RIZVIČ, 2022: *Slovene Conformer CTC BPE E2E Automated Speech Recognition Model RSDO-DS2-ASR-E2E 2.0. http://hdl.handle.net/11356/1737* (31. 8. 2023).

Teodor PETRIČ, 2016: Dolgoročna raziskava o razvoju otroškega govora: Slovenske Samostalniške Sklanjatve. *Zbornik Prispevkov s Simpozija 2015*. Ur. Franc Marušič, Petra Mišmaš, Rok Žaucer. Nova Gorica: Založba Univerze. 91–112. *http://www.ung.si/media/storage/cms/attachments/2016/10/21/13/45/01/Zbornik-\%C5\%A0D9_okt.16_splet.pdf* (31. 8. 2023).

Teodor PETRIČ*, 2021:* Razvoj slovenskih glagolskih oblik in spregatev na primeru predšolskega otroka. *Škrabčevi Dnevi 11: Zbornik Prispevkov s Simpozija 2019*. Ur. Franc Marušič, Petra Mišmaš, Rok Žaucer. Nova Gorica: Založba Univerze. 78–101. *http://www.ung.si/media/storage/cms/attachments/2021/01/27/12/29/22/Zbornik-\%C5\%A0D11-2021-3.pdf* (31. 8. 2023).

Daniel POVEY, Ghoshal ARNAB, Gilles BOULIANNE, Lukas BURGET, Ondrej GLEMBEK, Nagendra GOEL, Mirko HANNEMANN, et al., 2011: The Kaldi Speech Recognition Toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society.* DOI: https://doi.org/10.1017/pan.2018.62 (31. 8. 2023).

Sven-Oliver PROKSCH, Christopher WRATIL, Jens WÄCKERLE, 2018: Testing the Validity of Automatic Speech Recognition for Political Text Analysis. *Political Analysis* 27/3, 339–359.

R Core Team, 2023: R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/ (31. 8. 2023).

Cvetka ŠERUGA-PREK, Emica ANTONČIČ, Ajda KALAN, Ivan LOTRIČ, 2004: *Slovenska Zborna Izreka. Aristej.*

Ashish VASWANI, Noam SHAZEER, Niki PARMAR, Jakob USZKOREIT, Llion JONES, Aidan N GOMEZ, Łukasz KAISER, Illia POLOSUKHIN, 2017: Attention Is All You Need. *Advances in Neural Information Processing Systems*. Ed. I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett. Vol. 30. Curran Associates, Inc. *https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf* (31. 8. 2023).

Darinka VERDONIK, Andreja BIZJAK, Mirjam SEPESY MAUČEC, Lucija GRIL, Simon DOBRIŠEK, Janez KRIŽAJ, Gregor STRLE et al., 2023: *ASR Database ARTUR 1.0 (Transcriptions). http://hdl.handle.net/11356/1772* (31. 8. 2023).