

KAKO IZMERITI PERFORMANSE RAČUNARSKIH MODELA ZA KONVERZIJU GOVORA U TEKST?

Evaluacija računarskih modela za konverziju govora u tekst je postala izuzetno važna u kontekstu najnovijih tehnoloških pomaka, koji su doveli do realne upotrebljivosti ovih modela i jake tržišne konkurencije. Ovaj rad pokazuje zašto je objektivna evaluacija izazovan problem, ali i to da ovaj problem nije nerešiv ako se merenju performansi pristupi iz novog ugla. Umesto striktno objektivne evaluacije u odnosu na jedno dato rešenje, naš predlog je fleksibilna evaluacija na varijabilnom skupu podataka za testiranje modela. Predstavljamo, takođe, i jedan primer varijabilnog skupa podataka, koji se satoji od uzoraka transkribovanog govora ukupnog trajanja oko 15 sati.

Ključne reči: konverzija govora u tekst, korpus razgovornog jezika, računarski modeli, evaluacija, ortografska transkripcija

1. UVOD

Mnogo toga se promenilo u poslednjih četrdesetak godina, koliko je prošlo od pionirskog istraživanja prof. Polovine na temu leksičko-semantičke kohezije u *razgovornom jeziku* (Polovina 1986). U vreme ove studije, razgovorni jezik se morao prvo zabeležiti diktafonom (ili sličnim namenskim uređajem), zatim uzorkovati, pa ručno transkribovati. Svaki od ovih koraka je zahtevao mukotrpan i dugotrajan rad pre nego što bi se uopšte i prešlo na samu analizu. O pomoći računara smo tada mogli samo da maštamo. Iako se uveliko radilo na razvijanju sistema za automatsko prepoznavanje govora, rešenja nije bilo nigde na vidiku (Jelinek 2009). Ogromna varijabilnost zvučnog signala je bila nepremostiva prepreka za uspešno modelovanje, a sposobnost ljudskog uma da zvučni signal pretoči u pisani tekst je i dalje bila misterija. Zato su uzorci transkribovanog govora bili dragoceni i teško dostupni podaci za potrebe lingvističkog istraživanja.

Nakon decenija razvoja i ulaganja, računarska tehnologija je konačno dostigla takav nivo da je automatska transkripcija postala relativno lako dostupna čak i za srpski. Njena upotrebljivost danas daleko prevazilazi potrebe lingvističkih istraživanja. Medijske kuće, na primer, bi želele da konvertuju u tekst svoje arhive kako bi ih lakše pretraživale, razne firme bi želele da automatski prave zapisnike sa sastanaka, lekari bi želeli da dokumentuju razgovore sa pacijentima i

diktiraju izveštaje. Primene su zaista raznovrsne, a sa velikom potražnjom, javlja se i raznovrsna ponuda. Trenutno su u opticaju različita rešenja, takozvani *modeli*, i pitanje svih pitanja je: koje rešenje izabrati? Ispostavlja se da je objektivna evaluacija performansi modela iznenađujuće komplikovana.

Tema ovog rada je upravo problem evaluacije računarskih modela za konverziju govora u tekst. Cilj je da pokažemo zašto je ovaj problem izazovan, ali i to da nije nerešiv ako se revidiraju neka ustaljena uverenja u vezi sa merenjem performansi. Umesto striktno objektivne evaluacije u odnosu na jedno dato rešenje, naš predlog je fleksibilna evaluacija na varijabilnom skupu podataka za testiranje modela. Predstavljamo, takođe, i jedan primer varijabilnog skupa podataka, koji se satoji od uzoraka transkribovanog govora ukupnog trajanja oko 15 sati. Osim primarne namene usmerene ka evaluaciji računarskih modela, ovi podaci bi mogli da posluže i za lingvistička istraživanja na empirijskoj osnovi (Polovina 2015).

2. RAČUNARSKI MODELI ZA KONVERZIJU GOVORA U TEKST

Računarska obrada govora je ogromno polje istraživanja sa dugom tradicijom, tako da bi i najkraći pregled postojećih rešenja znatno prevazišao obim ovog rada. Ovde uvodimo samo najvažnije termine i koncepte koji su neophodni za bolje razumevanje problema evaluacije.

Konverzija govora u tekst se odvija u nekoliko koraka. Zvučni talas se prvo podeli na vrlo kratke segmente zvane *prozori* iz kojih se zatim izdvoje i zabeleže najrelevantnija fizička svojstva zvuka. Zabeležene vrednosti predstavljaju numeričku reprezentaciju datog prozora – svaki prozor je predstavljen nizom brojeva, to jest, postaje vektor u višedimenzionalnom prostoru. U sledećem koraku se trenira klasifikator koji svakom prozoru pridružuje odgovarajuću fonemu. U tom smislu, svaka fonema je jedna klasa koju klasifikator predviđa na osnovu svojstava zabeleženih u prozoru. Obično se nekoliko uzastopnih prozora pridružuje istoj fonemi. Ovo mapiranje se naziva *akustički model* i uči se na velikom broju primera poravnatog zvuka i teksta kao na Slici 1. Pridružene foneme se zatim konvertuju u slovne znakove, to jest tekst.

Zbog ogromne varijabilnosti zvučnog signala, vrednosti zabeležene u prozorima nisu dovoljne da bi se odgovarajuća fonema i, dalje, slovni znak nedvosmisleno pridružili. Zato se akustičkom modelu dodaje *jezički model*. Zadatak jezičkog modela je da proceni verovatnoću svake reči dobijene konverzijom imajući u vidu prethodne reči. Drugim rečima, jezički model “ispravlja” nisku slovnih znakova koja izađe iz akustičkog modela tako što nisku koja ne odgovaraju nijednoj reči u jeziku zameni najverovatnijom rečju koja može da se pridruži datoj sekvenci prozora u datom kontekstu.

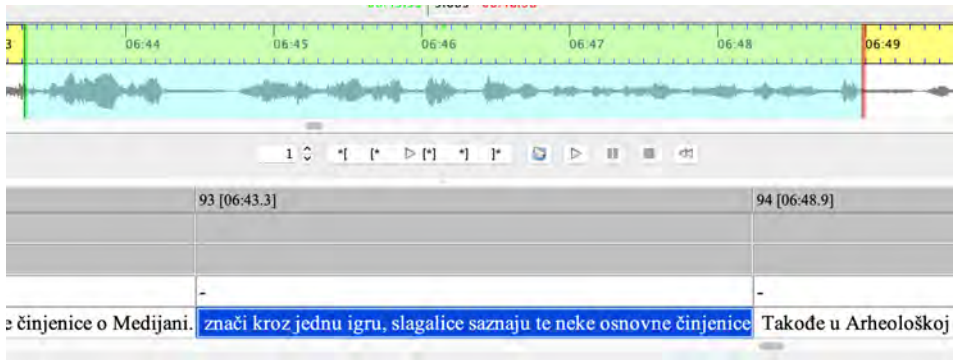
Tehnike za obuku akustičkog i jezičkog modela se brzo menjaju prateći razvoj tehnologije (Jelinek 2009). Tokom dugog vremenskog perioda, sistemi za konverziju govora u tekst su se sastojali od niza programa, gde bi svaki program bio zadužen za jedan korak u procesu. Kaldi (Povey et al. 2011) je vrlo popularan sistem ovog tipa u slobodnoj upotrebi, koji se i danas koristi u praksi, mada se smatra zastarelim u odnosu na novija rešenja. Kaldi koristi neuronske mreže za neke korake, tipično za akustičke modele, ali su klasifikatori i jezički modeli drugačijeg tipa (statistički). Kaldijevi *recepti* su konkretne primene sistema na određeni set podataka na određenom jeziku. Jedan takav recept je izrađen i za srpski (Popović et al. 2015). Veliki napreci u razvoju tehnologije za obuku neuronskih mreža od 2011. nadalje omogućili su da se svi koraci objedine u jednu ogromnu neuronsku mrežu. Iako se konceptualno zadržavaju iste komponente, njihov sled u sklopu jedne velike mreže postaje apstraktniji i fleksibilniji.

Tek je pojava *velikih prenosivih neuronskih mreža* 2019. omogućila iskorak u performansama koje vidimo danas. Korišćenjem ove tehnologije modele je moguće obučavati na ogromnim količinama audio materijala na različitim jezicima. Takođe, moguće je delimično obučavanje i bez poravnatog teksta, dakle samo na audio podacima, kao u slučaju modela XLS-R (Babu et al. 2021), ali se ipak najbolji rezultati dobijaju obukom na paralelnim podacima. To je upravo pristup primenjen u izradi modela Whisper (Radford et al. 2023), koji u poslednje vreme daje zaista impresivne rezultate.

U vreme sistema Kaldi procenjivalo se da je potrebno oko 2 000 sati transkribovanog audio materijala da bi se obučio iole upotrebljiv model. Poređenja radi, XLS-R je obučen na 500 000 sati (doduše netranskribovanog) audio materijala, dok je Whisper obučen na 680 000 sati transkribovanog audio materijala. Oba ova modela su višejezična, pa tako uključuju i srpski. Iako nije moguće tačno utvrditi poreklo i sastav srpskih podataka, performanse su na tom nivou da se može očekivati široka primena u bliskoj budućnosti.

3. ZAŠTO JE EVALUACIJE IZLAZA MODELA PROBLEM

Da bismo izmerili performanse modela za konverziju govora u tekst, izlaz modela se poredi sa segmentom teksta koji zaista odgovara datom segmentu govora. Segment sa kojim poredimo izlaz modela naziva se *referenca* i smatra se jedinim tačnim rešenjem tako da se svako odstupanje od reference računa kao greška.



Slika 1: Prikaz zvučnog talasa poravnatog sa odgovarajućim segmentom teksta u specijalizovanom programu za ručnu transkripciju EXMARaLDA.

Mera koja se tradicionalno koristi u evaluaciji sistema za konverziju govora u tekst je *stopa greške na nivou reči*, na engleskom *word error rate* (WER). U poslednje vreme sve više se koriste i mere poput *stope greške na nivou karaktera* (engl. *character error rate*, CER) i chrF (Popović 2015), ali ćemo se mi ovde zbog jednostavnosti fokusirati samo na WER, pošto je to najčešća mera i dovoljna za naše izlaganje.

Merom WER se izražava broj odstupanja izlaza modela od reference u odnosu na dužinu referentnog segmenta, kao što je pokazano u formuli (1). Brojilac u formuli čine oznake za broj dodatih reči (insertions, I), broj zamena (substitutions, S) i broj izostavljenih reči (deletions, D). Ovo su tri tipa odstupanja od reference. Imenilac predstavlja dužinu referentnog segmenta merenu u broju reči (N).

$$WER = \frac{I+S+D}{N} \cdot 100 \quad (1)$$

Vrednosti I, S i D se dobijaju primenom čuvenog algoritma Levenštajnova razdaljina (Levenshtein 1965),¹¹ koji garantovano pronalazi minimalnu udaljenost između dve sekvence, u našem slučaju dve niske reči: izlaza i reference.

Primer na Slici 1 pokazuje referencu (ručnu transkripciju). Za ilustraciju mere WER upoređićemo sa ovom referencom (R) jedan prilagođeni primer izlaza modela (M) i izbrojati odstupanja (O) :

¹¹ Pojednostavljeno objašnjenje se može naći u udžbeniku (Jurafsky, Martin 2024).

M	znači	i	kroz	jednu	igru	slagalice	sa	znaju	-	neke	osnovne	činjenice
R	znači	-	kroz	jednu	igru	slagalice	-	saznaju	te	neke	osnovne	činjenice
O	-	I	-	-	-	-	I	S	D	-	-	-

$$WER = \frac{2+1+1}{12} \cdot 100 = \frac{1}{3} \cdot 100 \approx 33\%$$

Vidimo da je ova mera izuzetno stroga, tako da mala odstupanja, kao što su ne prepoznavanje granica između reči ili izostavljanje kratkih reči u navedenom primeru, dovode do dosta visoke stope greške. Vrednost ove mere može da izađe i na preko 100%, što se dešava kada je izlaz modela duži od reference i uglavnom pogrešan. Ovo ne bi bio problem kada bi referenca zaista bila jedino tačno rešenje. Činjenica da je mera vrlo osetljiva na male razlike ne bi bila problematična sama po sebi kada bi referenca zaista bila jedino tačno rešenje. To, međutim, nikada nije tako.

U evaluaciji konverzije govora u tekst često se zanemaruje ili zapostavlja činjenica da gotovo svaki segment govora može *tačno* da se transkribuje na različite načine, što zavisi od toga u kojoj meri se teži doslovnosti transkripcije.

Obično se kao najdoslovnija uzima *fonetska transkripcija*, gde se umesto ortografskih slovnih znakova koristi međunarodni fonetski standard (International Phonetic Alphabet, IPA). Na primer, reč *znači* bi se po ovom standardu pisala /z n ʌ: tʃ i: /. Nivo nijansiranja u ovoj vrsti transkripcije ilustruje fonetski znak ʌ: umesto slovnog znaka *a*, koji pokazuje da se radi o zatvorenom dužem vokalu. Ovakva transkripcija se zaista koristila dugo vremena u razvoju i evaluaciji mašinskih modela. Jasno je da je njena izrada bila izuzetno spora i skupa: za transkribovanje govora u trajanju od jednog sata potrebno je minimum 200 sati stručnog rada, dakle za 2 000 sati govora neophodnih za treniranje modela 400 000 sati rada na transkripciji, što je 50 000 radnih dana ili otprilike 150 godina. Za evaluaciju je veći problem to što, i pored međunarodnog standarda i stručnog pristupa, nije bilo moguće proizvesti samo jednu tačnu transkripciju. Na primer, postavilo bi se pitanje da li je govornica zaista izgovorila dugo ʌ:, ili je možda ipak bilo ʌ bez dužine. To je, recimo, u srpskom posebno osetljivo pitanje jer standard propisuje obaveznu dužinu, dok se u realnom govoru može čuti čitav spektar dužina od vrlo kratkih do standardnih. U toj situaciji, neminovna je nekonzistentnost, koja, kao što smo videli u primeru gore dovodi do znatno različitih stopa greške.

Savremeni sistemi ne koriste fonetsku transkripciju ni za treniranje ni za evaluaciju. Segmenti govora se direktno uparaju sa segmentima teksta pisanog po ortografskim normama datog jezika. Ovakva transkripcija se zove *ortografska* i mnogo je dostupnija. Osim što za nju nije potreban visok nivo stručnosti

(potrebno je samo biti pismen), nije potrebno ni utvrđivati dužinu vokala i slične fenomene. Obično se procenjuje da je za ortografsko transkribovanje jednog sata govora potrebno najmanje 20 sati rada, dakle 10 puta manje nego za fonetsku transkripciju. Takođe, smatra se da je ortografska norma stabilna i konzistentna. S druge strane, ortografska norma ne beleži detalje govora koji su modelima bitni za prepoznavanje glasova. Takođe, ne beleže se ni elementi govora kao što su hezitacije, ponavljanja, ispravljanja, smeh i slično. Bez obzira na to, obrada velike količine podataka (stotine hiljada sati govora) ipak omogućava savremenim modelima dosta uspešno mapiranje. Problem evaluacije, međutim, ne rešava se ni ortografskom transkripcijom jer ortografska norma ne pokriva brojna pitanja mapiranja govora u tekst, što otvara prostor za nekonzistentan zapis čak i kad se propisana norma prati u potpunosti.

Kao ilustraciju, navodimo još tri moguće transkripcije za isti segment govora kao u prethodnom primeru, dakle ukupno četiri transkripcije koje su, svaka na svoj način, tačne:

M	znači	i	kroz	jednu	igru	slagalice	sa	znaju	-	neke	osnovne	činjenice
R1	znači	-	kroz	jednu	igru	slagalice	-	saznaju	te	neke	osnovne	činjenice
O1	-	I	-	-	-	-	I	S	D	-	-	-

M	znači	i	kroz	jednu	igru	slagalice	sa	znaju	-	neke	osnovne	činjenice
R2	znači	-	kroz	l	igru	slagalice	-	saznaju	te	neke	osnovne	činjenice
O2	-	I	-	S	-	-	I	S	D	-	-	-

M	znači	i	kroz	jednu	-	igru	slagalice	sa	znaju	-	-	neke
R3	znači	-	kroz	jednu	ovaj	igru	slagalice	-	saznaju	kažem	te	neke
O3	-	I	-	-	D	-	-	I	S	D	D	
M	osnovne	činjenice										
R3	osnovne	činjenice										
O3	-	-										

M	znači	i	kroz	jednu	-	igru	slagalice	sa	znaju	-	-	neke
R4	znači	-	kroz	l	ovaj	igru	slagalice	-	saznaju	kažem	te	neke
O4	-	I	-	S	D	-	-	I	S	D	D	
M	osnovne	činjenice										
R4	osnovne	činjenice										
O4	-	-										

Izmereni (zaokruženi) WER skorovi u ove četiri varijate su sledeći:

R1: WER = 33% , **R2:** WER = 42%, **R3:** WER = 43%, **R4:** WER = 50%.

Isti izlaz modela dobija znatno različite skorove u zavisnosti od proizvoljnih odluka u izradi reference. U R3 i R4 imamo dva govorna elementa koja nisu uključena u R1 i R2, a ne nalaze se ni u izlazu modela. Izostavljanje ovakvih elemenata nije greška, već rezultat težnje ka zapisu bližem pisanom jeziku u referenci, odnosno vrste podataka za treniranje u slučaju modela. S druge strane, prisustvo ovih elemenata je realno, pa tako ni njihovo uključivanje nije greška i neki modeli bi ih mogli prepoznati i uključiti u izlaz.

Ove četiri varijante su samo ilustracija jednog kontinuuma reprezentacije govora čak i u ortografskoj transkripciji. Moguće je, na primer, da se i rečca *i*, koja je uključena u izlaz modela zaista čuje u nekoj meri odvojeno od kraja prethodne reči. Možda se u jednom slušanju čuje, a u drugom ipak ne? Šta da radimo ako je govornica rekla *osnove* a jasno je da je imala nameru da kaže *osnovne*? Kada je *znači* govorni element, a kada nije?

Poseban problem predstavlja pisanje brojeva, skraćenica i stranih naziva. U primeru gore imamo jedan jednostavan slučaj pisanja broja 1 (jedan). Još jedan primer bi bio broj *5 000*, koji možemo da napišemo i kao *5 hiljada* i kao *pet hiljada*. U skraćenice ubrajamo slučajeve tipa *EU*, gde se često dešava da govornik izgovori puno ime *Evropska Unija*, ali da izlaz modela sadrži skraćenicu, što, u principu, nije greška jer bi i govornici ponekad tako zapisali. Dalje, imamo slučajeve tipa *OK*, što može da se napiše tako, ali i kao *okay* ili *okej*. Skraćenica *LGBT* se najčešće ne izgovara kao što je napisano, već kao *el-dži-bi-ti*, dok se skraćenica *RTS* izgovara kao *er-te-es*, ali i kao puni naziv *Radio Televizija Srbije*. U skraćenice ubrajamo i simbole tipa *m²*, *km*, *cm*, *h* i *%*, koji se obično izgovaraju kao pune reči *kvadratnih metara*, *kilometara*, *centimetara*, *sati* i *posto*, ali ih nije pogrešno zapisati ni kao simbole. Strani nazivi tipa *Facebook* / *Fejsbuk*, *Twitter* / *Tviter*, *Viber* / *Vajber* se obično pišu prema engleskom pravopisu, ali ne uvek, tako da su obe varijante zapisa tačne. Tome treba dodati ogroman broj stranih naziva firmi, bendova, pesama, filmova, kao i imena osoba, čija je transkripcija sve manje regulisana pravopisom, tako da je pravopis izvornog jezika vrlo često u upotrebi, mada, opet, ne uvek.

Ovakve varijacije vode gotovo bezgraničnom nijansiranju ortografske transkripcije, što pokazuje da ona nije ni jednostavna ni konzistentna, kako se obično smatra. A iz tog bezgraničnog variranja proizlaze rezultati evaluacije modela koji su vrlo aproksimativni i neuporedivi, do te mere da objektivna evaluacija sistema može da deluje neizvodljivo.

S druge strane, u literaturi se najčešće navode postignute vrednosti mere WER a da se uopšte ne diskutuju odluke o tome koja varijanta transkripcije je korišćena. Kao primer možemo da uzmemo rezultate koje navode Popović et al. (2015), gde isti model postiže WER od 1,86 u jednom slučaju, dok je WER u drugom slučaju čak 48.50, dakle stopa greške je 26 puta veća. Šta možemo da očekujemo od tog modela u nekom trećem slučaju koji nije ni obuhvaćen studijom, na primer, ako želimo da njime transkribujemo neku televizijsku emisiju koja nas zanima?

Striktno govoreći, performanse modela su uporedive ako su modeli obučeni i testirani na istom setu podataka (s tim što podaci za obuku i za testiranje moraju da budu razdvojeni), što obično i važi u okviru jedne studije, pa tako i kod Popović et al. (2015). Problem je, međutim, što u stvarnoj upotrebi podaci neće biti istog tipa, što će dovesti do znatnog pada u odnosu na procenjene performanse. Problem varijacije podataka postaje još veći kod velikih prenosivih modela, koji trenutno daju najbolje performanse, jer podaci na kojima su oni obučeni nisu dostupni javnosti, tako da ni ne možemo da znamo kakvi su.

4. NIJANSIRANJE ORTOGRAFSKE TRANSKRIPCIJE ZA TESTIRANJE MODELA

Naš predlog je da se modeli za konverziju govora u tekst testiraju na kontrolisano varijabilnim ortografskim transkripcijama. Varijabilnost transkripcije se obično vezuje za dijalekte (Ali i dr. 2016, Nigmatulina i dr. 2020), ali ispostavlja se da je potrebna i za visoko standardizovane jezike poput hrvatskog i srpskog. Uzimajući u obzir analizu iznetu u prethodnoj sekciji predlažemo dvodimenzijalnu varijaciju. Prva dimenzija je nivo doslovnosti, gde predlažemo dve varijante iste transkripcije: jednu koja vernije prenosi šta je izgovoreno (doslovna) i jednu koja je bliže važećoj normi (standardna). Varijante R1 i R2 u primeru navedenom gore bi tako bile standardne, dok bi varijante R3 and R4 bile doslone. Druga dimenzija je dužina transkripcije merena u slovnim znakovima. Prema ovom kriterijumu, varijante R1 i R3 su kraće, dok su varijante R2 i R4 duže. Dužina može da ima nekoliko vrednosti u zavisnosti od toga koliko elemenata je ispisano slovima.

Prva dimenzija omogućava da se u doslovnoj varijanti ispišu reči onako kako su izgovorene (npr. *osnove*) dok se u standardnoj varijanti ovakvi slučajevi mogu ispraviti (npr. *osnovne*). Na taj način nijedna opcija ne mora da se uzme kao greška. Druga dimenzija omogućava da se pokriju najčešći slučajevi pisanja brojeva, skraćenica i stranih izraza. Kombinacija ove dve dimenzije daje nekoliko potencijalno tačnih transkripcija (slično primeru gore).

U postupku testiranja modela izmerimo WER u odnosu na svaku ponuđenu varijantu i uzmemo najbolju vrednost (najniži skor). Na taj način omogućavamo

modelima da budu testirani na sličnoj verziji podataka na kakvoj su obučavani, a istovremeno i saznajemo za koju verziju su bolje obučeni.

Da bi se dobila bolja slika o tome kako modeli rade, potrebno je izračunati vrednost WER za svaki iskaz. Konačna vrednost može da bude prosečna vrednost svih najboljih skorova, ali može da se uzme i drugi kriterijum. Na primer, možemo da ograničimo izbor po konzistentnosti i zahtevamo da se svaki model evaluira na jednoj ili više zadatih varijanti u svakoj instanci. Moguće je definisati i drugačije kriterijume u zavisnosti od potreba korisnika. Ono što je bitno za princip evaluacije koji predlažemo jeste da se uvede fleksibilnost koja omogućava bolji uvid u performanse. Iako naš predlog ne omogućava iscrpno beleženje svih mogućih nijansi, omogućava ipak realiniju procenu performansi pod različitim okolnostima.

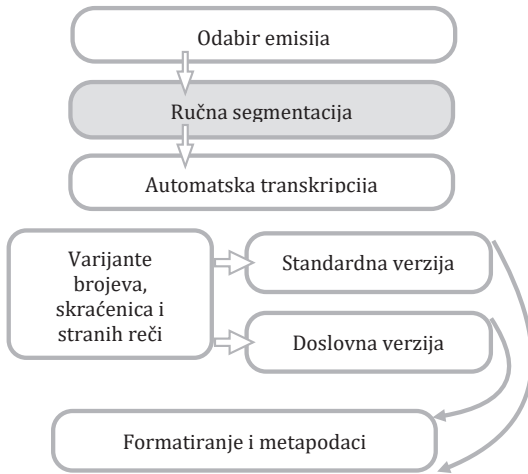
4.1 Projekat Mak na konac

Rukovodeći se dizajnom nijansirane ortografkse transkripcije tim istraživača na ljubljanskom Institutu Jožef Štefan u saradnji sa beogradskim udruženjem ReLDI Centar i programom *Jezik i prostor* Univerziteta u Cirihu izradio je prvi skup podataka za fleksibilnu evaluaciju modela za konverziju govora u tekst na hrvatskom i srpskom. Ovaj projekat, pod nazivom *Mak na konac*, finansirali su zajednički slovenačka jezička infrastruktura CLARIN.SI, kroz centar za proučavanje južnoslovenskih jezika CLASSLA, i program *Jezik i prostor* Univerziteta u Cirihu. Udruženje ReLDI je bio glavni izvođač zadataka anotacije i kontrole kvaliteta podataka. Tim je brojao ukupno 9 članova (petoro anatora, koordinatorka i troje istraživača). Izrada skupa podataka je trajala šest meseci (1. novembar 2023. - 30. april 2024.), dok je evaluacija modela u trenutku pisanja ovog rada i dalje u toku. Prvi prelinarni rezultati će biti prikazani na kraju ove sekcije.

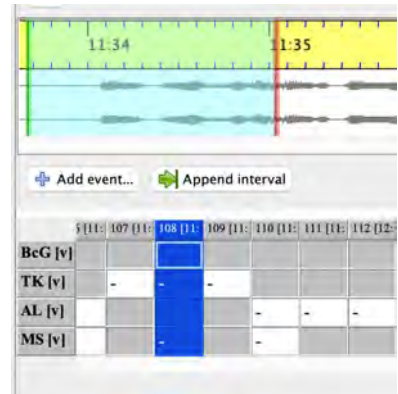
Ukupno je obrađeno oko 15 sati govornog materijala preuzetog iz tri izvora (oko 5 sati po izvoru):

- SR1: Radijske emisije u produkciji *Peščanik* (Beograd),
- SR2: Televizijska emisija *15 minuta* u produkciji *Južne vesti* (Niš),
- HR1: Radijska emisija *Ponedjeljkom u 3PM* u produkciji *Radio Student Zagreb* (Zagreb).

Plan je u početku bio da u sastav korpusa uđe i još jedan hrvatski izvor koji bi predstavljao južnije varijante govora (Split), ali do ovog trenutka nismo uspeli da obezbedimo saglasnost za upotrebu podataka. Za sve ostale izvore dobili smo saglasnost medijskih kuća, tako da će podaci biti slobodno dostupni i objavljeni putem infrastrukture CLARIN.SI nakon završetka evaluacije. Priprema podataka se odvijala u nekoliko koraka (Slika 2), koje opisujemo u nastavku ove sekcije.



Slika 2: Pregled toka izrade korpusa Mak na konac. Faza ručne segmentacije je prikazana na Slici 3



Slika 3: Prikaz predobrade audio snimaka za unos automatske transkripcije sa ručno unetim granicama segmenata u programu EXMARaLDA. Svaki red označava po jednog govornika koji učestvuje u emisiji. Crtice pokazuju koji govornik se čuje u datom segmentu.

Nakon odabira emisija tako da se uključe sadržaji sa što raznovrsnijim govornicima (različitog pola, uzrasta i zanimanja), prvi anotatorski zadatak je bio da se audio snimak segmentira na iskaze slične primeru na Slici 1. Za ovaj postupak koristili smo program EXMaRALDA (Schmidt and Wörner 2014), koji nudi opciju ručnog poravnanja govora i teksta. Tačnije, program omogućava da se obeleže vremenske tačke u audio snimku gde se završava jedan segment i počinje drugi (*time stamps*). Zadatak anotatora ili anotatorke je da, slušajući audio snimak, ručno obeleži granice. U principu, nakon obeležavanja granice, transkripcija se unosi ručno u odgovarajuće polje koje se formira unosom granice segmenta. Umesto toga, mi smo pribegli polu-automatskom unosu.

U slučaju SR1, već su postojali transkripti emisija, ali oni nisu bili poravnati sa audio snimcima na nivou segmenta. Zato je bilo potrebno uvesti postojeće transkripte u program EXMaRALDA i ručno dodati segmente i time automatski i poravnanje. Tokom ovog postupka ispostavilo se da su transkripti bili dosta slobodni i vrlo često nisu bili ni blizu nivoa doslovnosti koji je neophodan za evaluaciju modela. Sva odstupanja su zato ručno ispravljena. U narednom koraku ručno smo uzorkovali segmente kod kojih nema preklapanja. Takođe, prilikom uzorkovanja vodili smo računa o zastupljenosti govornika tako što smo u uzorak uključivali približno jednaku količinu govora (oko 10 minuta po govorniku).

U slučaju SR2 i HR1, transkripti nisu postojali pa smo ovde umesto ručne transkripcije prvo uneli automatsku, tj. konvertovali smo audio snimke u tekst pomoću modela koji su bili dostupni na početku projekta. Anotatorski zadatak je prvo bio da se odrede granice segmenata, s tim što se umesto transkripcije u zadato polje unosi samo jedan znak (crtica) kao što je pokazano na Slici 3. U ovom primeru vidimo preklapanje govornika u obeleženom segmentu, dok u prethodnom i narednom segmentu nema preklapanja. Nakon segmentiranja, uzorkovali smo segmente po istom principu kao i u SR1. Uzorci su zatim poslani na automatsku obradu u kojoj su crtice zamenjene izlazom modela.

OVAKO DOBIJENI DOKUMENTI SU DALJE ANOTIRANI U DVA KORAKA. U prvom koraku smo ispravljali prvobitnu transkripciju tako da dobijemo dosledno standardnu verziju. Takođe, dodavali smo varijante pisanja brojeva, skraćenice i stranih reči. U drugom koraku smo dodavali govorne elemente u kopije standardnih transkripcija. Na taj način smo za svaki audio snimak dobili dve transkripcije, standardnu i doslovnu (prema prvoj dimenziji varijacije), dok su varijante pisanja brojeva, skraćenica i stranih reči (druga dimenzija) unete u obe transkripcije.

U finalnom koraku podaci su formatirani tako da je za svaki izvor formirana po jedna tabela, gde svaki red sadrži jedan segment (oko 3 000 redova, tj. Segmenata po izvoru). Svaki red sadrži sledeća polja:

1. Identifikator segmenta
2. Identifikator govornika/ce
3. Lokacija odgovarajućeg audio snimka
4. Standardna transkripcija (sa varijantama)
5. Doslovna transkripcija (sa varijantama)

Svakoj od tri glavne tabele pridružena je i po jedna pomoćna tabela koja sadrži metapodatke govornika/ce:

1. Identifikator govornika/ce
2. Dužina uzorka
3. Lokacija izvornog audio snimka
4. Lokacija punog poravnatog transkripta
5. Naziv emisije
6. Internet adresa emisije
7. Ime govornika/ce
8. Pol
9. Približan uzrast
10. Zanimanje

Ključ preko koga se povezuju ove tabele je identifikator govornika/ce. Pošto je zamisao da se izmeri WER skor na svakom segmentu, onda je jasno da pi-druživanje metapodataka omogućava višeslojne analize uspešnosti modela. Tako možemo da ustanovimo da li demografske karakteristike utiču na performanse, a možemo i da ispitamo ostala svojstva segmenata (specifična leksika, konstrukcije, način izražavanja). Takođe, možemo da posmatramo na koji način dostupnost podataka na hrvatskom utiče na performanse na srpskom i obratno, to jest da li je uputno razdvajati podatke na ovim jezicima ili ne.

Ovakve analize dalje omogućavaju objektivnu evaluaciju modela u odnosu na poželjne vrednosti umesto pokušaja da se dobije univerzalna mera kvaliteta izlaza. Dakle, umesto da nastojimo da sve modele rangiramo na jednoj, univerzalnoj skali kvaliteta koju propisuje jedna tačna referenca, možemo da odredim skup kriterijuma koji su nama bitni i prema njima ocenimo modele u direktnom poređenju na onsovu nekoliko referenci. Možda nam nije bitno da li model meša srpski i hrvatski, ali nam je bitno da pouzdano i dosledno prepoznaje brojeve. Takođe, možda nam više odgovara model koji uvek pomalo greši, ali nikad mnogo, dok nam ne odgovara model koji neke segmente konvertuje savršeno dok u nekim pravi ogromne greške. Ako nam nije uopšte bitno gde model greši, već samo želimo da imamo ukupan skor, onda možemo da izračunamo srednju vrednost WER tako što ćemo za svaki segment uzeti najbolji WER. Na taj način smanjujemo prostor za uticaj arbitrarnih odluka na rezultat evaluacije.

WER	SR1		SR2		HR1	
	M1	M2	M1	M2	M1	M2
Najbolji	22,6	23,8	15,2	19,4	16,2	27,9
Najlošiji	28,8	29,9	21,7	26,1	25,9	37,0

Tabela 1: Preliminarna evaluacija dva modela na skupu podataka *Mak na konac*.

4.1 Primer poređenja dva modela na osnovu korpusa *Mak na konac*¹²

Kao primer korišćenja korpusa navešćemo jedno jednostavno poređenje dva modela na nivou ukupnog skora WER na sva tri podkorpusa *Mak na konac* na osnovu preliminarne evaluacije.

- **M1**: Aktuelna verzija modela Whisper (Radford et al. 2023) dostupna na repozitorijumu Hugging Face (whisper-large-v3).¹³ Ovaj model se smatra trenutno najboljim rešenjem.

¹² Modele su testirali istraživači na Institutu Jožef Stefan.

¹³ <https://huggingface.co/openai/whisper-large-v3>

- **M2:** Ista verzija modela Whisper dodatno obučena na samo srpskim podacima, takođe dostupna na repozitorijumu Hugging Face (whisper-large-v3-sr-combined).¹⁴ Zbog dodatnih podataka za srpski, očekivano je da ovaj model pokaže bolje performanse nego polazni M1.

Tabela 1 sadrži rezultate ove evaluacije, koja pokazuje da polazni model M1 daje bolje performanse u svim slučajevima. Suprotno očekivanju, dodatna obuka za srpski ne daje bolje rezultate na srpskom, ali zato znatno pogoršava rezultate na hrvatskom. Takođe vidimo, da su rezultati generalno najbolji kod SR2, na šta je najverovatnije uticala dostupnost podataka za obuku iz istog izvora (Rupnik i Ljubešić 2022). Na kraju, vidimo i dosta izraženu razliku između najboljeg i najlošijeg skora, što pokazuje raspon variranja skora iz arbitrarnih razloga. Najbolji rezultat M2 na SR1 i SR1 je bolji od najlošijeg rezultata M1. To znači da bi evaluacija u odnosu na jednu referentnu transkripciju, koja favorizuje izlaz M2, dovela do pogrešnog zaključka da dodatna obuka za srpski rezultira boljim performansama. Dostupnost varijabilnih referenci omogućava bolje poređenje modela iako performanse nije moguće tačno pozicionirati na univerzalnoj skali uspešnosti.

5. ZAKLJUČAK

Nagli razvoj tehnologije za konverziju govora u tekst otvara brojne mogućnosti za inovacije i u sferi praktične primene i u sferi empirijskog istraživanja jezika. Uz sve veću ponudu računarskih modela, neophodno je uspostaviti kriterijume za njihovo objektivno poređenje uz što manji upliv arbitrarnih faktora. Ovaj rad je pokazao da se objektivnost u evaluaciji modela može znatno poboljšati ako se umesto jedne referentne transkripcije koristi skup referentnih varijanti. Predstavili smo i prvi takav skup podataka za evaluaciju modela za hrvatski i srpski, čija izrada je upravo završena. Ovi resursi predstavljaju značajan doprinos primenjene lingvistike efikasnijem praćenju budućih pomaka u tehnologiji obrade jezika.

LITERATURA

- Ali, A., Bell, P., Glass, J., Messaoui, Y., Mubarak, H., Renals, S., and Zhang, Y. (2016). The mgb-2 challenge: Arabic multi-dialect broadcast media recognition. In: *2016 IEEE Spoken Language Technology Workshop (SLT)*, 279–284. IEEE.
- Babu, A., Wang, C., Tjandra, A., Lakhota, K., Xu, Q., Goyal, N., Singh, K., Platen, P.V., Saraf, Y., Pino, J.M., Baevski, A., Conneau, A., & Auli, M. (2021). *XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale*. Interspeech.

¹⁴ <https://huggingface.co/Sagicc/whisper-large-v3-sr-combined>

- Jelinek, F. (2009). ACL Lifetime Achievement Award: The Dawn of Statistical ASR and MT. *Computational Linguistics*, 35(4), 483–494.
- Jurafsky, D. and Martin, J. H. (2024). *Speech and Language Processing* (3rd ed. draft), Online (<https://web.stanford.edu/~jurafsky/slp3/>)
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8), 707–710. Original in: *Doklady Akademii Nauk SSSR* 163(4), 845–848.
- Polovina, V. (1987). *Leksičko-semantička kohezija u razgovornom jeziku*. Beograd : Filološki fakultet Beogradskog univerziteta.
- Polovina, V. (2015). Tradicija i inovacija - dve strane digitalne humanistike. In (ed. Александра Вранеш, Љиљана Марковић): *Дигитална хуманистика : тематски зборник у 2 књиге*. Књ. 1, 49–57.
- Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal: Association for Computational Linguistics, 392–395.
- Popović, B., Ostrogonac, S., Pakoci, E., Jakovljević, N., and Delić, V. (2015). Deep Neural Network Based Continuous Speech Recognition for Serbian Using the Kaldi Toolkit. In: *Proceedings of the Speech and Computer: 17th International Conference Proceedings*.
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I.. (2023). Robust speech recognition via large-scale weak supervision. In: *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, 28492–28518.
- Rupnik, P. and Ljubešić, N. (2022). *ASR training dataset for Serbian JuzneVesti-SR v1.0*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042 (<http://hdl.handle.net/11356/1679>).
- Schmidt, T. and Wörner, K. (2014), “EXMARaLDA”, In: *Handbook on Corpus Phonology*. Oxford University Press, 402-419.

Tanja Samardžić

HOW TO MEASURE THE PERFORMANCE OF COMPUTATIONAL MODELS
FOR SPEECH TO TEXT CONVERSION?

Summary

The evaluation of computational models for speech-to-text conversion has become especially needed in the context of the latest technological advances, which have led to the real usability of these models and a strong market competition. This paper shows why objective evaluation is a challenging problem, but also that solving this problem is not impossible if we approach performance measurement from a new angle. Instead of a strict objective evaluation in relation to one given solution, our proposal is a flexible evaluation on a variable test data set. We also present an example of a variable data set, which consists of transcribed speech samples with a total duration of about 15 hours.

Key words: speech-to-text conversion, spoken language corpus, computational models, evaluation, orographic transcription

Tanja Samardžić
URPP Language and Space, University of Zurich
tanja.samardzic@uzh.ch