Употреба модела *Whisper Large v3 Sr* за транскрипцију говора на српском језику у програмском језику *Пајтон* на платформи *Гугл колаб*

Научни рад DOI: 10.18485/judig.2025.1.ch27

Никола Јанковић[1], *0000-0003-3484-4220*
Јована Иваниш2, *0009-0006-1225-3461*

Апстракт

У овом раду представљен је скрипт у програмском језику *Пајтон* (енг. *Python*) на платформи *Гугл колаб* (енг. *Google Colab*) који користи фино подешени модел за транскрипцију говора на српском језику *Whisper Large v3 Sr* (https://huggingface.co/Sagicc/whisper-large-v3-sr-cmb), чиме се омогућава бесплатно, квалитетно и једноставно транскрибовање говора на српском језику у текст. Мотивација за креирање овог скрипта проистекла је из недостатка доступних алата који истраживачима пружају једноставно коришћење овог модела без потребе за напредним техничким знањима и значајним рачунарским ресурсима. Овај скрипт омогућује једноставно учитавање аудио-записа, њихову транскрипцију помоћу модела *Whisper Large v3 Sr* и преузимање транскрипције у текстуалном формату. У раду ће најпре укратко бити описан горенаведени модел који скрипт користи, а затим ће детаљно бити описан начин функционисања самог скрипта, уз упутство за коришћење. Такође, биће представљени проблеми које је било неопходно превазићи за успешну примену овог приступа, као што је потреба за аутоматском сегментацијом звучних записа, одређивање оптималних параметара сегментације и омогућавање подршке за различите формате аудио-материјала. Осим тога, биће

[1] Филолошки факултет, Универзитет у Београду, nikolajankovickv@gmail.com
[2] Институт за српски језик САНУ, jovana.ivanis@gmail.com,

представљено неколико приступа везаних за смањивање броја грешака у транскрипцији. На крају, биће приказана евалуација тачности и ефикасности транскрипције, као и статистичка анализа уочених грешака. Сматрамо да овај алат може бити од великог значаја за истраживаче, с обзиром на то да убрзава процес обраде говорних података, омогућава обраду велике количине аудио-материјала у кратком периоду и пружа конзистентан и поновљив метод транскрипције, што је нарочито значајно за научну методологију и поновљивост истраживања везаних за језик. Поред тога, сматрамо да алат може бити користан и другим корисницима, с обзиром на то да омогућава једноставно креирање титлова за видео-садржаје, претварање аудио-бележака у текстуални формат, стварање аутоматски генерисаних транскрипата за особе са оштећеним слухом, као и стварање текстуалних архива аудио-садржаја.

Кључне речи: транскрипција говора, српски језик, Пајтон, Whisper Large v3, Гугл колаб, NLP.

1. Увод

Технологија аутоматског препознавања говора (енг. *automatic speech recognition*, скр. ASR) прошла је од својих почетака у касним 1940-им и 1950-им годинама до данас вишедеценијски процес унапређивања и промена парадигми на којима је заснована (Jurafsky & Martin 2025). У овом процесу, ова технологија сазрела је до те мере да може имати практичну примену у различитим областима (ibid.). Са друге стране, постоји неколико препрека које отежавају примену ове технологије на језицима са мањим бројем говорника у односу на енглески (попут српског језика), нарочито када су у питању решења отвореног кода (енг. *open-source*).

Најпре, сами скупови података на којима се модели за препознавање говора тренирају садрже знатно већи проценат енглеског језика у односу на друге језике. Како наводе Пратап и сарадници (Pratap et al. 2020: 2757), овакви скупови података постоје за друге језике, али су најчешће ограниченог обима, похрањени на различитим местима и ретко су доступни под отвореном лиценцом. Када је у питању српски језик, највећи допринос на овом пољу представља недавно објављени *ParlaSpeech-SR*, који обухвата 896.22 сата говора на српском језику[3] (Ljubešić et al. 2025). Затим, наведени

[3] У питању су транскрипти говора посланика у заседањима Народне скупштине Републике Србије.

фактор доступности адекватних података (односно, њихово одсуство) значајно утиче на перформансе модела када је у питању аутоматско препознавање говора; наиме, како наводе Радфорд и сарадници (Radford et al. 2023: 28499), стопа грешака на нивоу речи (енг. *word error rate*, скр. *WER*) има јаку негативну корелацију са бројем сати транскрибованог говора за одређени језик ($r^2 = 0.83$).

Такође, када је у питању употреба модела отвореног кода за транскрипцију говора, за њихову употребу потребно је познавање напредних софтверских алата, као и поседовање (или изнајмљивање) значајних техничких ресурса.

Горенаведени фактори су били мотивација за овај рад, у оквиру кога смо представили три различите методе примене верзије модела за аутоматско препознавање говора која је фино подешена за српски језик, чиме се истраживачима и другим корисницима омогућује једноставан, бесплатан и транспарентан приступ овој технологији. Такође, у раду је представљена упоредна анализа перформанси сва три приступа, као и језичка својства саме транскрипције која је произашла из њих.

2. Методологија

2.1. Основна методологија процеса транскрипције

Када је у питању фаза планирања система аутоматске транскрипције говора у нашем истраживању и компоненти које би он обухватао, основну одлуку у овом процесу представљао је одабир самог модела за транскрипцију говора, односно за аутоматско препознавање говора. Као основни модел одабран је *Whisper* модел компаније *OpenAI* (Radford et al. 2023), заснован на кодер-декодер трансформер архитектури, која се, како наводе Сузић и сарадници (Suzić et al. 2024: 175), показала поузданом у литератури и у пракси. С обзиром на мали удео српског говора у оригиналном *Whisper* моделу (Radford et al. 2023: 28517), одабрали смо *whisper-large-v3-sr-cmb*[4], верзију наведеног модела која је дообучена за српски језик на следећим скуповима података са српским говором: *Google/Fleurs*[5], *Mozilla/Common Voice*[6] и *JuzneVesti-SR*[7] (Сагић 2024: 221).

[4] https://huggingface.co/Sagicc/whisper-large-v3-sr-cmb
[5] https://huggingface.co/datasets/google/fleurs
[6] https://huggingface.co/datasets/mozilla-foundation/common_voice_9_0
[7] http://hdl.handle.net/11356/1679

Следећа одлука у истраживању тицала се платформе на којој бисмо омогућили учитавање и бесплатну употребу овог модела, без потребе за инсталацијом додатног софтвера и без потребе за напредним хардверским ресурсима. Као најбоље решење показала се платформа *Google Colab*[8], осмишљена као истраживачки пројекат за развој прототипа модела машинског учења, а која пружа временски ограничен приступ окружењу за *Jupyter notebook*[9], у коме се може вршити интерактивни развој и извршавање програма (Bisong 2019: 59), а чији су хардверски ресурси довољни за покретање различитих модела вештачке интелигенције, укључујући и *whisper-large-v3-sr-cmb*.

Први корак у процесу израде скрипта било је омогућавање учитавања аудио-фајлова путем графичког интерфејса, што платформа *Google Colab* подржава помоћу функције *files.upoload()* у оквиру своје *Python* библиотеке.

Затим, било је потребно омогућити подршку за различите фајлова и прилагодити их формату који очекује *Whisper* модел, који је трениран на моно аудио-фајловима са фреквенцијом одабирања од 16,000 Hz (Radford et al. 2023: 28494). Ово је постигнуто обрадом улазног аудио-записа помоћу *ffmpeg* библиотеке (преко *pyDub*[10] библиотеке), при чему је, такође, извршена нормализација нивоа звука на -20 dBFS.

Када је у питању сама употреба модела за предикцију (енг. *inference*), *Whisper* модел подржава контекст од 30 секунди, због чега је за сегменте који трају дуже потребно извршити сегментацију (Radford et al. 2023: 28501). У *transformers* библиотеци платформе *HuggingFace*, која подржава позивање *Whisper* модела, постоје два мода[11]: први, заснован на „исечцима" (енг. *chunks*), у коме се аудио-запис дели на више краћих сегмената са малим преклапањем између сегмената, и други, секвенцијални алгоритам, у коме се аудио-запис дели на сегменте од 30 секунди са „клизајућим прозором" (енг. *sliding window*), који користи транскрипцију из претходног сегмента као додатни контекст, што доприноси већој тачности. У процесу тестирања ова два приступа, други, секвенцијални алгоритам, показао се као тачнији приступ јер је садржао мање грешака на крајевима сегмената у односу на први приступ. Како би крајња транскрипција

[8] https://colab.research.google.com/
[9] https://jupyter.org/
[10] https://github.com/jiaaro/pydub
[11] https://huggingface.co/openai/whisper-large-v3

била што квалитетнија, у нашем раду смо за предикцију помоћу *transformers* библиотеке одабрали секвенцијални приступ.

Приликом анализе транскрипције добијене овим приступом, био је приметан већи број грешака на крајевима реченица, које су типичне за грешке у сегментацији. Из овог разлога, одлучили смо да тестирамо други приступ, у коме бисмо применили напреднији вид сегментације. С обзиром на то да је *Whisper* модел (неправилно) транскрибовао и сегменте говора који су садржали звуке који нису говор (као што је музика), одлучили смо да у новој методологији за транскрипцију применимо моделе за детекцију гласовне активности (енг. *voice activity detection*, скр. *VAD*), који омогућују боље препознавање почетка и краја говора од једноставног приступа заснованог на тишинама у звучном сигналу (Behre et al. 2022; Ding et al. 2020: 433). У овој фази, тестирали смо популарне моделе отвореног кода за детекцију говора као што су *WebRTC VAD*[12], *Silero VAD*[13], и *Pyannote* (Bredin et al. 2020), при чему је *Pyannote* у нашим тестовима остварио најбоље резултате. Како бисмо омогућили упоредивост са претходним приступом, заснованим на аутоматској сегментацији помоћу *transformers* библиотеке, извршена је иста нормализација нивоа звука као у претходном приступу и задржана иста фреквенција одабирања звука. Такође, како бисмо осигурали компатибилност овог приступа транскрипцији са *Whisper* моделом, применили смо адаптивну сегментацију, у којој су најпре добијени сегменти говора на основу *VAD* модела, а затим је на сваки сегмент који је трајао дуже од 30 секунди примењена функција *split_on_silence* модула *pyDub* уз почетне параметре минималног трајања тишине (*min_silence_len*) од 500 милисекунди, прага тишине *(silence_thresh*) од -40 *dBFS,* и задржавања сегмента тишине (*keep_silence*) од 200 милисекунди; у случају неуспешне сегментације, вршено је до 5 покушаја аутоматске сегментације уз скраћивање минималног трајања тишине за 100 милисекунди и повећањем прага тишине 2 dBFS. Уколико и након ове процедуре сегментација није произвела сегменте адекватне дужине, сегмент је подељен на најмањи број једнаких делова који су краћи од 30 секунди. Такође, како би се смањила вероватноћа одбацивања речи на самим крајевима сегмента од стране модела, приликом сегментације додато је 100 милисекунди тишине на почетак и крај сваког сегмента. Наведени приступ сегментацији звука довео је до

[12] https://github.com/wiseman/py-webrtcvad
[13] https://github.com/snakers4/silero-vad

смањења броја грешака у транскрипцији, што је детаљније описано у другом делу рада.

Упркос бољој сегментацији, добијена транскрипција је и даље садржала значајан број грешака. Како бисмо истражили могућности аутоматске корекције грешака у процесу транскрипције српског говора, креирали смо трећи скрипт, у коме се примењује неколико метода аутоматске детекције и исправљања грешака у транскрипцији. У овом приступу, након што модел генерише транскрипцију, она се најпре прослеђује модулу *CLASSLA-Stanza*, који представља скуп алата за аутоматску језичку анотацију јужнословенских језика (Ljubešić et al. 2024). Помоћу овог модула транскрибовани текст се дели на реченице, а затим се добијене реченице деле на токене помоћу следећег регуларног израза: „*\w+|[^\w\s]|\s+*”, који раздваја реченицу на речи, интерпункцију и размаке. Појединачни чланови овог низа се затим могу мењати у исправкама, а *join* функцијом се склапа финална реченица, чувајући оригиналну интерпункцију и размаке.

Један од основних елемената у нашем приступу аутоматској корекцији транскрипције представља листа фреквенција српских речи, коју смо за потребе овог рада креирали на основу два извора:

- морфолошког речника *SrpMD* (Krstev & Vitas 2015), чији су облици речи аутоматски сматрани валидним, и којима су такође на основу *SrpMD* речника придружене информације о врсти речи и леми;
- подкорпуса великих веб-корпуса и великих електронских корпуса српског језика, међу којима издвајамо корпусе Кишобран и Знање (Škorić & Janković 2024), одакле су издвојени облици речи који имају више од 10,000 појављивања у обједињеном корпусу. На основу истих корпуса генерисана је и листа фреквенција биграма у обједињеном корпусу.

Ове листе фреквенција учитане су у модул *symspellpy*, који је био основа за анализу грешака у транскрипцији и њихово исправљање. Израда наведених листа фреквенција биће детаљније описана у другом раду. У нашој методологији, свака реч из претходног корака упоређивана је са наведеним листама фреквенција.

Како би се избегле непожељне исправке, примењено је неколико врста критеријума који су изузимали реч из процеса аутоматског исправљања грешака. Примери типова речи које су аутоматски изузете су: облици речи који постоје у *SrpMD*

морфолошком речнику, облици речи чија је фреквенција већа од 100, властите именице (речи које је *CLASSLA-Stanza* модул класификовао као *PROPN*), речи које почињу великим словом а не налазе се на почетку реченице, речи које садрже бројеве, познате скраћенице, једно или два велика слова у средини реченице (ради избегавања измене иницијала), речи које садрже латиницу, речи под наводницима и интерпункција и размаци.

За сваку реч која није задовољавала један од ових критеријума, помоћу модула *symspellpy*, а на основу претходно поменуте листе фреквенција, генерисани су кандидати за замену речи поређани по растојању уређивања (енг. *edit distance*), а затим по учесталости. Како би се погрешне исправке смањиле на минимум, на наведене кандидате је затим примењено неколико конзервативних филтера, као што је висока сличност између речи, усаглашеност оригиналне речи и предложене замене по врсти речи (према *CLASSLA-Stanza* модулу). Након филтрирања, преостали кандидати рангирани су према скору који обједињује фреквенцију предложених речи, скор из једноставног биграмског језичког модела и скор којим се фаворизују речи које су већ присутне у тексту.

С обзиром на то да сви претходни механизми исправљања грешака не узимају у обзир контекст, значајан број исправака се састојао од сличних речи које су семантички биле потпуно погрешне. Из овог разлога, за предложене речи чија је фреквенција мања од 500 применили смо методологију исправки која користи модел *Jerteh-355*, који остварује најбоље резултате у моделовању маскираног језика за српски језик (Škorić 2024). У овом кораку, изградња маскиране реченице вршена је заменом циљане речи токеном <*mask*>, а затим је генерисано 15 предлога модела за дату реч (*top_k=15*). Затим, ови предлози су пролазили сличну процедуру филтрирања и поновног рангирања као и речи у претходном приступу, при чему је у овом случају у обзир скор вероватноће који модел пружа, док је максимално растојање уређивања у овом случају било 1[14]. У финалном кораку, врши се уједначавање предложених речи на основу њихове учесталости у самом тексту. С обзиром на ограничења у простору, у опису методологије дат је преглед само њених најзначајнијих аспеката, док ће више техничких детаља о приступима наведеним у овом раду бити доступно на *GitHub* страници везаној за наше истраживање[15], која ће бити јавно доступна по објављивању рада.

[14] Изузетак је комбинација других фактора који веома позитивно утичу на скор, када је максимално растојање уређивања 2.

[15] https://github.com/nikolakv/whisper-sr-transcription

Такође, напомињемо да су у процесу израде методологије, креирању кода и вршењу анализа коришћени модели вештачке интелигенције компанија *OpenAI* и *Google*, при чему смо вршили ручну проверу креираног кода, а предлоге методологије смо валидирали истраживањем релевантне стручне и научне литературе.

3. Квантитативна и квалитативна анализа грешака

Тачност различитих модела транскрипције, описаних у претходним сегментима рада, проверавали смо користећи аудио-записе преузете са платформе *Јутјуб* (енг. *Youtube*). У питању је звучни материјал у трајању од 2 сата и 32 минута, ексцерпиран из видео-записа документарног филма *Магија истока*[16] и кратких предавања из области биологије, преузетих са образовног канала *Изокренута учионица*[17]. Целокупан звучни материјал анализиран је у оквиру сваког појединачног модела, те ће анализа грешака бити представљена с освртом на тип транскрипционог модела који је коришћен.

С обзиром на то да модел због „кодер-декодер трансформер архитектуре" емитује токене који могу чинити више од једне фонеме, као и да је токенизатор *Whisper* модела преузет из GPT-2 модела, који је у највећој мери трениран на енглеском језику, упркос подршци за више језика, оваква архитектура модела може довести до различитих грешака и формирања речи које у испитиваном језику не постоје (Radford et al. 2023: 28495; Suzić et al. 2024: 175).

Наш скуп података за валидацију садржао је сегменте са говором дијалекта, као и прекључивање кодова (енг. *code switching*), који су садржали већи број грешака у односу на друге сегменте. Ова чињеница је у сагласности са резултатима претходних истраживања, према којима ове врсте језичке варијабилности доприносе већем броју грешака у транскрипцији (Errattahi et al. 2018: 32).

3.1. Квантитативна анализа грешака

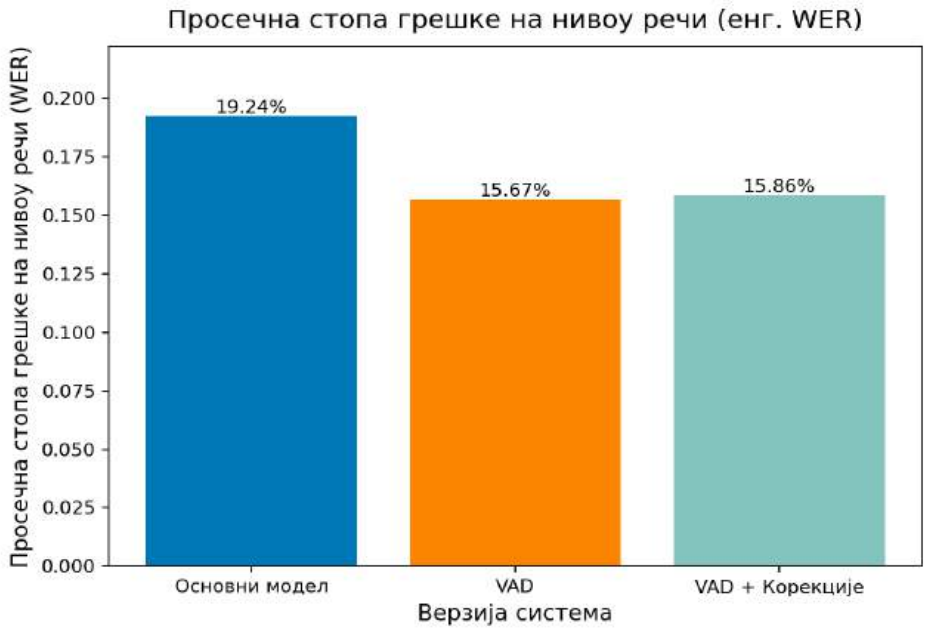
Како бисмо омогућили објективно и конзистентно упоређивање различитих приступа које смо користили, применили смо квантитативну анализу грешака, у којој смо помоћу *Python* скрипта упоређивали финалну верзију транскрипције у сва три приступа. Овај корак обухватао је израчунавање стопе грешке на нивоу речи (енг. *word error rate*, скр. *WER*) упоређивањем ручно

[16] https://www.youtube.com/watch?v=8xAlB_TH3nY
[17] https://www.youtube.com/user/irinabiologija

транскрибоване верзије са аутоматски транскрибованим верзијама: *WER = (S + D + I) / N*, односно дељењем збира измењених речи (*S*), избрисаних речи (*D*) и уметнутих речи (*I*) са укупним бројем речи у тексту (*N*).

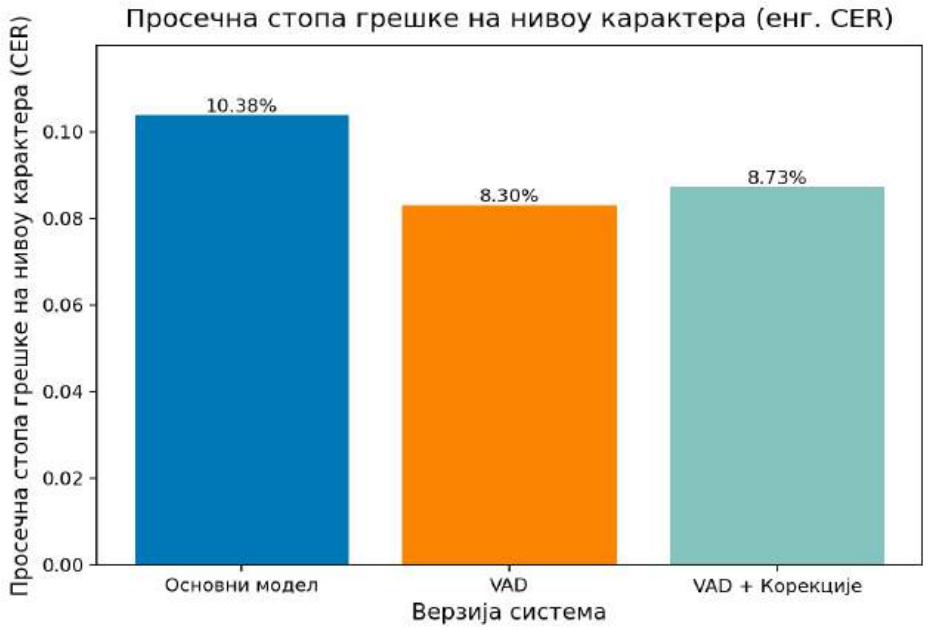
Помоћу *Python* библиотеке *jiwer* вршено је упоређивање грешака само за облик речи (енг. *plain WER*), где је пре упоређивања најпре вршена нормализација уклањањем интерпункције, сажимање размака и конверзија великих у мала слова. Осим грешака на нивоу речи, вршено је и израчунавање стопе грешке на нивоу карактера (енг. *character error rate*, скр. *CER*).

На фигурама 1 и 2 приказана је просечна стопа грешке на нивоу речи (енг. *plain WER*) и на нивоу карактера [18](енг. *CER*), редом, за сва три приступа.



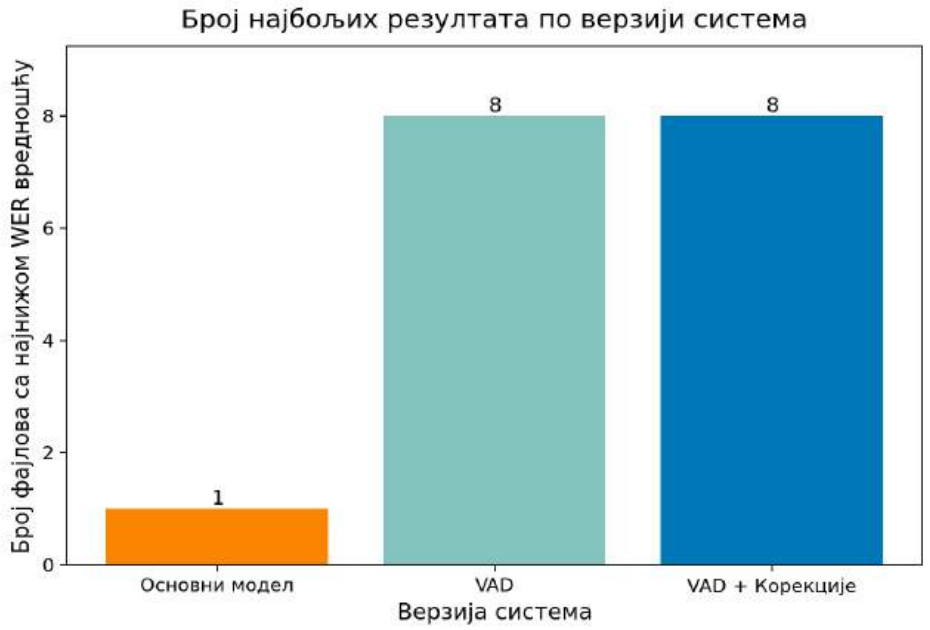
Фигура 1: просечна стопа грешке на нивоу речи (енг. WER) за сва три приступа

[18] https://github.com/jitsi/jiwer



Фигура 2: просечна стопа грешке на нивоу речи (енг. WER) за сва три приступа

Као што се може видети на фигурама 1 и 2, наша два модификована приступа остварила су значајно боље резултате у односу на оригиналну сегментацију у подразумеваној процедури за транскрипцију *transformers* библиотеке. Међутим, најмањи број грешака (како на нивоу речи, тако и на нивоу карактера) забележен је са *VAD* моделом без корекција транскрипције (15.67% на нивоу речи и 8.3% на нивоу карактера), док је приступ са корекцијама транскрипције имао нешто већу просечну стопу грешака (15.86% на нивоу речи и 8.73% на нивоу карактера). Са друге стране, као што се може видети на фигури 3, од 17 фајлова у валидационом скупу података, оба приступа са нашим модификацијама имала су исти број фајлова у коме су остварила најбољи резултат од сва три тестирана (при чему напомињемо да дужина фајлова није узета у обзир). Ови подаци указују на то да примењена методологија исправљања грешака у сегментацији у неким типовима текста доводи до бољих резултата, али да и даље постоји значајан број нежељених исправки које у одређеним случајевима доводе до већег укупног броја грешака.



Фигура 3: Број фајлова у којима је одређена верзија имала најмању стопу грешке у односу на остале верзије

С обзиром на то да WER не осветљава појединачне језичке аспекте који су специфични за различите приступе аутоматској транскрипцији (и аутоматским исправкама које се над њом врше), поред квантитативне анализе, извршили смо и квалитативну анализу транскрибованог текста добијеног у сва три приступа.

3.2. Квалитативна анализа грешака

3.2.1. Грешке на нивоу речи или фразе

Највећи недостатак претварања говора у писани текст коришћењем ових модела представља изостанак одређених речи. Иако би коришћење ових модела знатно умањило време које лингвисти користе за транскрипцију, изостављање делова текста онемогућило би коришћење ових модела за области у којима је важна свака изговорена реч, попут прагматике, фонетике, дијалектологије итд.

Речи и фразе које модел углавном прескаче јесу такозвани *језички испуњивачи*, речи које попуњавају паузе и празнине у говору, у тренуцима када говорник покушава да осмисли свој говор, не може да се сети шта је желео да каже, жели да исправи нешто у свом говору

или га допуни, као што су *неуобличена вокализација*, *поштапалице*, *понављање*, *надопуњавање* итд. (Кликовац, Чудомировић 2016: 7). Дакле, у потпуности су изостављене поштапалице, попут *значи*, *мислим*, *овај*, *па*, *некако*, *овако*, које, иако понекад означене као *језички шунд* (Фекете 1997), указују на важну комуникативну функцију и сигнализирају да говорник размишља о планирању свог исказа, па самим тим нису ни комуникативно сувишне (Чудомировић 2020: 52). Уочљиво је и изостављање дискурсних маркера, као што су: *такође*, *у ствари*, *међутим*, *заправо*, *наравно*, *управо*, *односно*, *рецимо*, *иначе*, *дакле* итд. Такође, исправљање говорника модел транскрибује тако што узима само један облик: ***па*** *на* ***неким*** *неком > на неком*; ***у јако хладним*** *у јако сувим пределима > у јако сувим пределима*; ***највећи*** *највећа маса > највећа маса.*

Поред тога, модел тежи економичности исказа, те изостављa, сажима и скраћује поједине речи и конструкције: то је било **више** практиковање > то је било практиковање; између Тимока, **исто тако** и Дунава > између Тимока, **али** и Дунава; нећу **чак то** ни на телевизији > нећу ни на телевизији итд. У таквим ситуацијама најчешће се изостављају неодређене и показне заменице (неки, некако, овај, онај), показне речце којима се уводе појмови (ево, ето, ено) и већ поменути испуњивачи. Имајући у виду особине скупова података на којима се модели за транскрипцију говора најчешће тренирају, а у којима се тежи језичкој економији, са једне стране, а да процес сегментације може довести до грешака у транскрипцији на крајевима појединачних звучних исечака, са друге стране, није јасно да ли су одређене грешке узроковане самим својствима модела или неправилном сегментацијом (некадашњу ботаничку башту > неботаничку башту; ако ћемо још да идемо > хућемо да идемо).

3.2.2. Грешке на нивоу гласа или групе гласова

Понекад због недовољно разговетног говора, затвореног или отвореног изговора вокала, долази до супституције, редукције или пак додавања гласа који није изговорен. Разлоге за то можемо потражити у поставкама *Whisper* модела, будући да су токени којима он располаже, а који могу садржати једну или више фонема, преузети из одређеног *ChatGPT* модела, те тако настају речи које у српском језику не постоје (Suzić et al. 2024: 175). Стога, како би се избегле грешке које подразумевају речи које не постоје у српском језику, у трећи модел транскрипције имплементиран је *фреквенцијски речник*, чија је методологија примене описана у претходном делу рада.

Примери који указују на то да је имплементирање фреквенцијског речника допринело тачности транскрипције јесу оне речи које су у прва два типа транскрипције, оригиналном моделу и моделу са ВАД сегментацијом, транскрибоване са одређеним фонетским грешкама, док у трећем моделу таквих грешака нема: *екстремитети* > *екстр**и**митети* (1), *екстр**и**митети* (2), *екстремитети* (3); *кичмењацима* > *кичме**ј**ацима* (1), *кичме**ј**ацима* (2), *кичмењацима* (3). Међутим, иако је имплементација модела повећала тачност транскрипције, постоје и речи које су тачне у оригиналном типу транскрипције, а погрешне у другом и трећем, попут *оповргнути* > *оповргнути* (1), *поврнути* (2), *оповр**к**нути* (3).

Будући да правилности у појављивању грешака у гласовним сегментима нема, тешко је одредити који тип грешака је учесталији, те да ли модел више греши када су у питању вокали или консонанти. Прво, када су у питању вокали, грешке у транскрипцији су бројне, што понекад можемо приписати неадекватном изговору говорника. Честе су супституције (*арх**е**типске* > *арх**и**типске*, *круне* > *круну*, *хладноће* > *хладноћа*, *попречном* > *попричном*) и редукције вокала (*смисао* > *смиса*), а понекад долази и до уметања или удвајања вокала (*почнимо* > *поч**и**нимо*, *усред* > ***уу**сред*). Грешака има и међу консонантима, посебно уколико су у питању речи у којима је извршена нека гласовна алтернација: *ишчупају* > *и**с**чупају* (1), *и**ш**чупају* (2, 3); *лисну дршку* > *лис**т**ну др**ж**ку* (1), *лис**т**ну др**ж**ку* (2), ***листу** дршку* (3). Нарочито се истичу примери супституције одређеног гласа његовим звучним односно безвучним парњаком: *ис**к**уставa* > *изгуставa*, *хла**д**ноће* > *хла**т**ноће*, *уз**д**ужном* > *ус**т**ужном*, ***п**раинатом* > ***б**рајантом*. Додатне примере грешака наводимо у табели:

Табела 4: Примери грешака с осврtом на тип транскрипционог модела

Оригинална верзија	Whisper (1)	VAD (2)	VAD + Jerteh (3)
*к**о**тлина*	*к**у**тлина*	*к**у**тлина*	*к**у**тлина*
*смањ**у**ју*	*смањ**а**ју*	*смањ**у**ју*	*смањ**у**ју*
изблиза	*изблиза**н***	*изблиза**к***	*изблиза*
кловн	*клов**е**н/клоун*	*клов**е**н/клоун*	*клов**е**н/клоун*
сведоче	*седоче*	*сл**е**доче*	*сл**е**доче*
аутохтона	*аутол**т**она*	*аутохтона*	*аутох**к**тона*
багрем	*багрем*	*багре**н***	*багре**н***
кестен	*кестен*	*кесте**р***	*кес**п**ен*
поменули	***у**поменули*	***у**поменули*	*поменули*
отров	*от**воров***	*от**воров***	***отвор пров***
успут	*успут**ом***	*успут**ом***	*успут**ном***
кисеоник	*кисони*	*кисони*	*кисони*
спољашње	*спољаш**ће***	*спољ**еш**ће*	*спољаш**ће***
причвршћен	*прич**ајћен***	*прич**ај***	*прич**ај***

Истоветан говорни сегмент понекад се транскрибује исправно, а некада није могуће одгонетнути његово право значење. Пример *што зглавкара чини зглавкаром* понавља се два пута у току аудио-записа. Први пут резултат транскрипције јесте: *спод изглавкара*, *чини изглавкаром* (1), *спод злавка рачини и злавка ром* (2), *спод зглавка рачини и зглавка рома* (3), док је сличан сегмент, када је други пут поменут, исписан тачно: *шта то зглавкара чини зглавкаром* (1, 2, 3). Такође, реч *нерватура* први пут је транскрибована као: *на ратвору* (1, 2), *на раствору* (3), док други пут имамо тачну реч – *нерватур**у*** (1, 2, 3).

Уколико је одређена реч кандидат за исправљање, а само се у једном гласу разликује од неке речи која постоји у фреквенцијском речнику, модел ће преузети ту реч, иако је семантички потпуно различита, што представљаману овог приступа: *врхова* > *вр**к**ова* (1), *вр**к**ова**на*** (2), ***врбована*** (3); *пауколике* > ***о**колик**и*** (1), ***о**колик**и*** (2), ***околини*** (3); *згазите* > ***сх**азите* (1), ***с**газите* (2), ***спазите*** (3). Како би се избегао овакав тип грешака, било би пожљено, у наредним имплементацијама модела, применити напреднији систем корекције који би узимао у обзир контекст. Са друге стране, предност модела увиђа се када модел уноси реч која је у оригиналу непотпуна: ***спец**и*... *за кретање* > ***специјално*** *за кретање.*

Неадекватна је и употреба великог слова, будући да модел понекад адекватно препознаје примере у којима је велико слово неопходно (*Дунав*, *Србија*, *Венера*), понегде погрешно наводи мало (*Бечу* > ***б**ечу*, *Рајац* > ***р**ајац*, *Перуна* > ***п**еруна*), а понегде велико слово (*пирамиде* > ***П**ирамиде*, *српски и словенски* > ***С**рпски и **С**ловенски*, *чудо* > ***Ч**удо*) итд.

Када је у питању транскрипција *антропонима* и *топонима*, модел у све три верзије подједнако греши, с обзиром на то да су властите именице изостављене из исправки. Примери попут *Дунав*, *Србија*, *Балкан*, *Венера*, *Дионис*, *Хиландар*, *Ирина* и сл. исправно су транскрибовани, уз поштовање правила о великом слову, док типичне грешке увиђамо код презимена и појединих топонима: *Дамњановић* > *Дамљановић*, *Демљановић*, ***Бе**мљановић*; *Минх* > *Мин*, *Мих*; *планине Ртањ* > ***Плање Нертањ***; *Зајечару* > *Зај**а**чару*, а понекад је реч која означава име потпуно изостављена: *хотел Рамонда* > *хотел*.

Интересантна је, такође, транскрипција изговора на енглеском језику, будући да су речи махом транскрибоване тако да подражавају српски изговор енглеских речи, попут *vein* > *вејн*; *phylum* > *фајлом*; *flower* > *флауер*; *midrib* > *мид риб* (1, 2), *миди рибе* (3), *petiole* > *петијал* (1), *петијол* (2, 3); *ovary* > *овери* (1), *оувери* (2), *овери* (3) и др., уз изузетак појединих речи које су преведене или постоје у сличном облику у српском језику (*water* > *вода*, *transports* > *транспорт*, *minerals* > *минерал*).

3.2.3. Грешке на нивоу сегментације текста

Када су у питању адекватна сегментација текста и поштовање интерпункцијских правила српског језика, сви модели показали су добре резултате. Најфреквентније грешке су: недостатак размака између појединих реченица, иако су оне одвојене тачком на крају прве и великим словом на почетку друге реченице, употреба малог слова на почетку реченице и неадекватна употреба зареза, премда су синтаксичке функције које се у српском језику одвајају зарезима углавном адекватно сегментиране: *причамо о листу***,** *биљном органу*.

Поред поменутих одступања, речи су понекад подељене на два дела или су две речи спојене у једну: *сунђери* > *су ђери*, *суњери*, *су њуђари*; *машту свих путника* > *маштус их путника*; *јутарњи сунчеви зраци* > *јутарњисунчеви зраци*; *са Ртња* > *свртња* итд. Наведене грешке могу бити узроковане погрешном сегментацијом од стране *VAD* модела, једноставнијом логиком за сегментацију, која је активирана уколико је одређени сегмент био дужи од 30 секунди и након употребе *VAD* модела, или особинама самог *Whisper* модела.

4. Закључак

У овом раду описали смо три различита приступа транскрипцији српског језика уз помоћ *Гугл Колаб* платформе и *whisper-large-v3-sr-cmb модела*, при чему смо постигли значајно смањење у просечној стопи грешке у односу на подразумевани приступ са *transformers* библиотеком. Сматрамо да приступ сегментацији са *VAD* моделом примењен у овом истраживању представља искорак у транскрипцији говора на српском језику када су у питању решења отвореног кода. Са друге стране, анализе везане за аутоматске исправке транскрипције указују на то да да су за поуздан систем аутоматског исправљања грешака потребне методе које би узимале контекст у обзир, затим пречишћене и проширене листе фреквенција, као и експерименти са различитим комбинацијама параметара алата описаних у овом раду. Такође, бројне грешке у иницијалној транскрипцији указују на то да је неопходно додатно обучавање модела за препознавање говора за српски језик. Упркос овим изазовима, сматрамо да наведени приступи транскрипцији могу бити корисни истраживачима, али и другим корисницима за потребе транскрипције српског језика.

Литература

[1] Behre, P., Parihar, N., Tan, S. S., Shah, A., Sharma, E., Liu, G., Chang, S., Khalil, H., Basoglu, C., & Pathak, S. D. (2022). Smart Speech Segmentation using Acousto-Linguistic Features with look-ahead. *ArXiv*, *abs/2210.14446*.

[2] Bisong, E. (2019). Google Colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* (pp. 59–64). Apress. https://doi.org/10.1007/978-1-4842-4470-8_7

[3] Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M.-P. (2020). Pyannote.Audio: Neural Building Blocks for Speaker Diarization. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7124–7128. https://doi.org/10.1109/ICASSP40776.2020.9052974

[4] Ding, S., Wang, Q., Chang, S.-Y., Wan, L., & Moreno, I. L. (2020). Personal VAD: Speaker-Conditioned Voice Activity Detection. *The Speaker and Language Recognition Workshop (Odyssey 2020)*, 433–439. https://doi.org/10.21437/Odyssey.2020-62

[5] Errattahi, R., Hannani, A. E., & Ouahmane, H. (2018). Automatic Speech Recognition Errors Detection and Correction: A Review. *Procedia Computer Science*, *128*, 32–37. https://doi.org/10.1016/j.procs.2018.03.005

[6] https://api.semanticscholar.org/CorpusID:253116616

[7] Jurafsky, D., & Martin, J. H. (2025). Automatic Speech Recognition. In *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/15.pdf

[8] Krstev, C., & Vitas, D. (2015). *SrpMD — Serbian Morphological Dictionaries*. https://doi.org/10.57771/j0ge-8e29

[9] Ljubešić, N., Rupnik, P., & Koržinek, D. (2025). The ParlaSpeech Collection of Automatically Generated Speech and Text Datasets from Parliamentary Proceedings. In A. Karpov & V. Delić (Eds.), *Speech and Computer* (pp. 137–150). Springer Nature Switzerland.

[10] Ljubešić, N., Terčon, L., & Dobrovoljc, K. (2024). *CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages*. Institute of Contemporary History. https://doi.org/10.5281/zenodo.13936406

[11] McFee, B., Raffel, C., Liang, D., Ellis, D., Mcvicar, M., Battenberg, E., & Nieto, O. (2015). *librosa: Audio and Music Signal Analysis in Python*, 18–24. https://doi.org/10.25080/Majora-7b98e3ed-003

[12] Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). MLS: a large-scale multilingual dataset for speech research. *Interspeech 2020*. https://doi.org/10.21437/interspeech.2020-2826

[13] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In A. Celikyilmaz & T.-H. Wen (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14

[14] Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning*, 28492–28518. https://proceedings.mlr.press/v202/radford23a.html

[15] Suzić, S., Popović, B., Pakoci, E., Pekar, D., Delić, T., & Delić, V. (2024). *Automatic Speech Recognition in the Serbian Language – A Comparison of Whisper-Based System and Conventional System with Language Model*, 174–177.

[16] Škorić, M., & Janković, N. (2024). New Textual Corpora for Serbian Language Modeling. *Infotheca*, *24*(1), 71–96. https://arxiv.org/abs/2405.09250

[17] Škorić, M. (2024). New Language Models for Serbian. *Infoteka*, *24*(1), 7–28. https://arxiv.org/abs/2402.14379

*

[18] Кликовац, Д., & Чудомировић, Ј. (2016). Скрипта из Лингвистике текста и прагматике за студенте српског језика и књижевности.

[19] Сагић, А. (2024). Употреба модела вештачке интелигенције у процесу дигитализације аудио и видео-грађе. Гласник Народне Библиотеке Србије, 23(26), 217–229.

[20] Фекете, Е. (1997). Поштапалице – елементи језичког шунда. Језик данас, 4, 5–8.

[21] Чудомировић, Ј. (2020). Зашто су „поштапалице" потребне. Књижевност и језик, 67(1), 41–55.

Usage of the *Whisper Large v3 Sr* model for the transcription of serbian spoken language in *Python* programming language on the *Google Colab* platform

Nikola Janković, Jovana Ivaniš

Summary

This paper presents a *Python* script on the *Google Colab* platform, which uses a fine-tuned model for the transcription of speech in Serbian, *Whisper Large v3 Sr* (https://huggingface.co/Sagicc/whisper-large-v3-sr-cmb), which enables free, high-quality and simple transcription of Serbian speech into text. The motivation for creating this script came from the lack of available tools which would allow researchers to use this model in a straightforward way, without the need for advanced technical knowledge and significant computing resources. This script provides a simple method for uploading audio files, transcribing them using the *Whisper Large V3 Sr model*, and downloading the transcription of the files in a textual format. Firstly, the paper briefly describes the aforementioned model used by the script, followed by a detailed description of how the script functions, along with a user manual. The paper will also present problems which we needed to overcome in order to successfully implement this approach, such as the need for automatic audio file segmentation, the determination of the optimal segmentation parameters, and the implementation of support for different audio file formats. Furthermore, several approaches related to the reduction in the number of errors in the transcription will be presented. We believe that this tool can be of significant importance to researchers,

considering that it speeds up the processing of audio data, enables users to proces vast amounts of audio material in a short period of time, and provides a consistent and repeatable transcription method, which is very significant for the scientific methodology and the repeatability of language-related research. We also believe that the tool can be useful to other users, considering the fact that it enables creating subtitles for video content, converting audio notes into text, creating automatically generated captions for the deaf and hard of hearing, as well creating textual archives of audio content.

Keywords: speech transcription, Serbian language, Python, Whisper Large v3, Google Colab, NLP.