# Neuronsko strojno prevođenje

**Lerga, Košuta Estera**

SVEUČILIŠTE U RIJECI

FILOZOFSKI FAKULTET U RIJECI

Dvopredmetni studij talijanskog jezika i informatike

Košuta Estera Lerga

# NEURONSKO STROJNO PREVOĐENJE

Diplomski rad

MENTOR: doc. dr. sc. Lucia Načinović Prskalo

KOMENTOR: doc. dr. sc. Marija Brkić Bakarić

Rijeka, 2021.

# SAŽETAK

Nedavna otkrića u području umjetne inteligencije utjecala su na sva zanimanja, a donijela su izrazite promjene u radu prevoditelja. Googleova AutoML Translate usluga koja omogućava korisnicima da treniraju Googleove sustave temeljene na neuronskom strojnom prevođenju (GNMT) koristeći svoje vlastite prevoditeljske memorije, smatra se jednim od najznačajnijih novih proizvoda pokrenutih 2018. godine u području neuronskog strojnog prevođenja. Osim analiziranja najboljih načina i alata za stvaranje i oblikovanje prevoditeljskih memorija, rad se bavi analizom rada AutoML Translate platforme na primjeru pet različito pripremljenih prevoditeljskih memorija. Prikazani su rezultati dobiveni u navedenim slučajevima od kojih su većina dosegli izvrstan uspjeh koji je prema Googleovoj bilingvističkoj evaluaciji opisan kao model koji proizvodi visoko kvalitetan neuronski strojni prijevod.

**Ključne riječi:** neuronsko strojno prevođenje, prijevodna memorija, AutoML Translate, Google Cloud

# ABSTRACT

Recent discoveries in the field of artificial intelligence have impacted various professions, including translation, and have resulted in significant changes in the work of translators. One of the most significant new products announced in 2018 is Google's AutoML Translate service, which allows users to train Google's systems based on neural machine translation (GNMT) with their own translation memories. The paper not only examines the best methods and tools for producing and structuring translation memories, but it also examines the AutoML Translate platform using five different translation memories. The preceding examples' findings are reported, with the majority of them great success, which Google's bilingual review describes as models that create high-quality neural machine translation.



**Keywords:** neural machine translation, translation memory, AutoML Translate, Google Cloud

# Zahvala

Mentorica doc. dr. sc. Lucia Načinović Prskalo i komentorica doc. dr. sc. Marija Brkić Bakarić su bile od neizmjerne pomoći pri radu. Zahvaljujem na vodstvu, suradnji i svim korisnim savjetima tijekom izrade ovog diplomskog rada. Također, zahvaljujem mojoj najdražoj obitelji koji su mi neizmjerna potpora u svemu u životu pa tako i studiranju. Hvala Mihaelu, Josipi, Rafaeli, Matei i Joelu za svaki osmjeh potpore i motivacije. Hvala Chiari, Danijeli i Marti na svakom predavanju provedenom na sunčanim terasama. Posebna hvala mami Veri bez koje se ovaj diplomski rad nikada ne bi dogodio i prije svega hvala dragom Bogu za sve blagoslove u životu.

# Sadržaj

# 1. Uvod

Od kad je svijet postao globalno selo, prevođenje je postalo jedna od glavnih alata komunikacije. Prenošenjem ideja s jednog jezika na drugi jezik povezale su se različite kulture, ideologije, običaji, poslovni sustavi i mnoge druge komponente današnjeg društva. Ubrzanim napretkom tehnologije proces prevođenja se postupno mijenjao, a razvojem umjetne inteligencije dosegao je automatsku razinu strojnog neuronskog prevođenja što je tema ovog diplomskog rada.

Kratkim pregledom područja strojnog prevođenja i nedavnog početka korištenja istog u praksi prevođenja vjerskih tekstova u Hrvatskoj preći će se na analizu teme neuronskog strojnog prevođenja općenito te posebice usredotočiti na način rada AutoML Translate platforme kao jedne od najnovijih alata Google Cloud platforme. Ukratko je opisana metodologija rada koja je detaljno analizirana u daljnjim poglavljima počevši od objašnjenja pojmova vezanih za prevoditeljsku memoriju i alate za njezino oblikovanje. Istaknut je program SDL Trados kao jedna od glavnih softverskih komponenti cjelokupnog projekta.

Korpus tekstova na kojima se temelji proces strojnog prevođenja rada je prijevod Biblije Ivana Vrtarića iz 2016. godine i prijevod tekstova Williama Branhama. Budući da je program Trados korišten u postupku prevođenja ovih tekstova, od velike je važnosti analizirati softverske dijelove i načine oblikovanja prevoditeljske memorije u istome. Rad nastavlja s opisivanjem procesa stvaranja prevoditeljske memorije prijevoda cjelokupne Biblije, objašnjavanje postupka poravnavanja iste i alati koji su korišteni u tom procesu.

Sve navedeno služilo je kao priprema za stvaranje dobre i kvalitetno poravnane baze podataka prevoditeljske memorije koja je kasnije korištena u procesu treniranja modela primjenom umjetne inteligencije na AutoML Translate platformi. Opisan je proces treniranja modela krenuvši od postavljanja modela za treniranje, prelazeći na opcije prijenosa paralelnog korpusa te učitavanja iste u sustav, treniranja modela prema potrebama korisnika, evaluacije modela te testiranja modela na novim rečenicama i korištenja kreiranog modela u prevoditeljskoj praksi.

U ovom projektu, postupak treniranja ponovljen je na pet različitih paralelnih korpusa kako bi se dobiveni rezultati usporedili i kako bi se došlo do potencijalnog zaključka o tome koji je najbolji način pripreme prevoditeljske memorije tako da platforma AutoML da što bolje rezultate konačnog modela. Prikazani su dobiveni BLEU rezultati za svaki model, ispisana je statistika broja

učitanih rečenica, kvaliteta poravnavanja prevoditeljske memorije (radilo se o sravnjivanju u paragrafima ili rečenicama), a dodatno je još analizirana i razlika kod učitavanja različitih .tmx i .tsv formata Valja spomenuti da se AutoML sustav svakodnevno usavršava te se očekuje da će u budućnosti davati sve bolje i bolje rezultate.

## 2. Pregled područja

Nagli napredak u području umjetne inteligencije utjecao je na razvitak područja strojnog prevođenja što posebice dokazuju nedavna istraživanja. Ako se fokusiramo samo na 2020. godinu i na prvu polovicu 2021. godine, već primjećujemo promjene kod novih metoda obrade prirodnih jezika i strojnog prevođenja (Šuman, 2021). Nedavna istraživanja su se također fokusirala na neuronsko strojno prevođenje i na izgradnje sustava za neuronsko strojno prevođenje (Saratlija, 2020.). Stručnjaci se također bave ljudskom evaluacijom sustava za neuronsko strojno prevođenje (Mikulić, 2020) te kontrastivnom analizom ljudskoga i strojnoga prevođenja (Višić, 2020.).

Nedavna istraživanja u području institucijskog prevođenja (Krizmanić, 2020.) također ističu važnost sustava za strojno prevođenje, budući da se upravo u praksi institucijskog prevođenja prevoditelji koriste softverskim rješenjima programa Trados iprevoditeljskim memorijama o čemu će kasnije biti više riječi. U ovoj sferi, bitno je također spomenuti važnost istraživanja u vezi utjecaja višejezičnosti vrednovatelja na ljudsku procjenu kvalitete strojnih prijevoda (Ljubas, 2020.).

Osim istraživanja koji se fokusiraju na višejezičnost, znanstvenici su u 2021. godini istraživali neuronsko strojno prevođenje s jednojezičnim prijevodnim memorijama (Cai, 2021.) fokusirajući se također na proučavanje sigurnosti podataka (Jiao, 2021.).

Utjecaji novih tehnologija na prevoditeljsku struku (Gogić, 2020.) nezaobilazna je tema rasprava među prevoditeljima, dok je ovaj rad dokaz napretka i ubrzavanja procesa prevođenja za englesko – hrvatski jezični par na primjerima vjerskih tekstova u Hrvatskoj.

## 2.1. Neuronsko strojno prevođenje

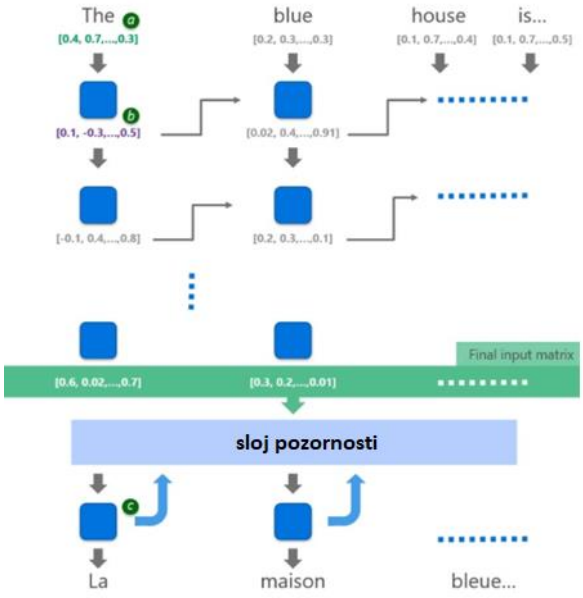
Razvojem umjetne inteligencije postigao se veliki napredak u različitim područjima računarstva pa time i u području strojnog prevođenja - pojavilo se neuronsko strojno prevođenje kao korak dalje od prethodnog statističkog načina prevođenja. Ono što je statistički način koji je prethodio razvoju neuronskog strojnog prevođenja podrazumijevao bilo je bazirano na statistici, odnosno na izračunavanju postoji li i kolika je „vjerojatnost da će se neka riječ prevesti nekom drugom ili da će se prijevodi dviju riječi koje se nalaze jedna pokraj druge također nalaziti jedni pored drugih.“ (Ljubaš, 2018.)

Kod spomenutog statističkog pristupa postojale su tri različite podvrste, a razlikovale su se u jezičnoj analizi teksta. Tri spomenute podvrste su temeljene na sintaksi, frazama i riječima (Yonghui Wu et al., 2016.). Ono što je drugačije kod neuronskog strojnog prevođenja je ideja da se do prevedenog teksta dolazi koristeći neuronske mreže odnosno duboko učenje. Da bi se neuronsko strojno prevođenje moglo definirati kao sekvencionalno predviđanje mora zadovoljiti uvjete da je tekst koji se ulazno obrađuje u obliku rečenice, što znači da se u ulaznom obliku ne obrađuju odlomci ili cijeli dokumenti, a kod izlaznog teksta podrazumijeva se generiranje teksta auto regresivno s lijeva na desno. Kod obrade teksta u neuronskom prevođenju postoje dvije neuronske mreže koje se paralelno koriste na način da prva obrađuje ulazne tekstualne podatke, dok druga daje izlazne rezultate. Radi se o sustavu koji kodira i dekodira te iz sustava izvlači prikaze bilo koje dužine nakon čega se iz prikaza generira točan prijevod (Zhang & Zong, 2020.).

Prilikom treniranja neuronske mreže, svaka riječ je kodirana pomoću vektora veličine 500 dimenzija (a) koji predstavljaju svoje jedinstvene karakteristike unutar određenog jezičnih para (npr. engleski i hrvatski). Na temelju jezičnih parova koji se koriste za treniranje, neuronska mreža samostalno definira što ove dimenzije trebaju biti. (Microsoft, 2021) Mogu se, primjerice, koristiti jednostavne značajke kao što su rod, vrsta riječi (glagol, imenica, pridjev, prilog i drugo), razina formalnosti teksta i slično, ali i bilo koje druge značajke koje proizlaze iz podataka na kojima se mreža trenira, a koji nisu toliko očiti.

Koraci u prevođenju prema (Microsoft, 2021) pomoću neuronske mreže su sljedeći:

1) „Svaka riječ, ili konkretno, 500-dimenzionalni vektor koji ju predstavlja, prolazi kroz prvi sloj "neurona" koji će ga kodirati u 1000-dimenzionalni vektor (b) koji predstavlja riječ u kontekstu drugih riječi u rečenici.
2) Nakon što su sve riječi jednom kodirane u ove 1000-dimenzionalne vektore, proces se ponavlja nekoliko puta, svaki sloj omogućuje bolju kombinaciju 1000-dimenzionalnog prikaza riječi u kontekstu pune rečenice
3) Završna izlazna matrica zatim koristi sloj pozornosti (eng. *attention layer*) pomoću kojega će se definirati koja će se sljedeća riječ iz izvorne rečenice prevesti. Također, isti se izračuni koriste kako bi se potencijalno ispuštale nepotrebne riječi na ciljanom jeziku.
4) Sloj dekodera prevodi odabranu riječ (točnije vektor od 1000 dimenzija koji predstavlja ovu riječ u kontekstu pune rečenice) u najprikladniji termin na ciljanom jeziku. Izlaz posljednjeg sloja (c) zatim se vraća u sloj pozornosti kako bi izračunao koja bi sljedeća riječ iz izvorne rečenice trebala biti prevedena (Slika 1).“



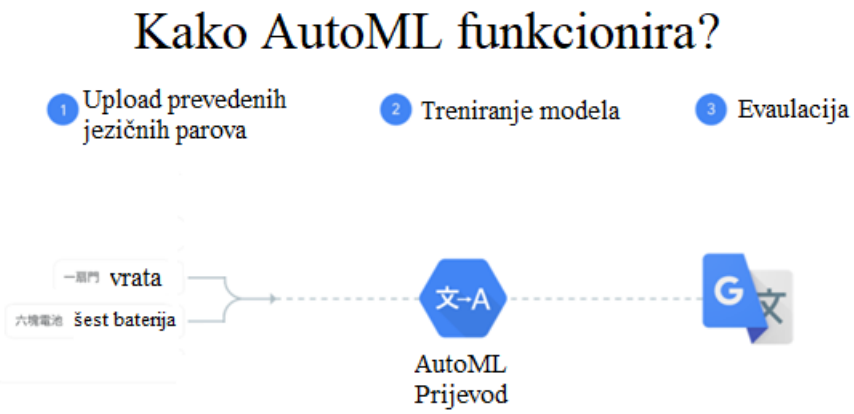

***Slika 1 Grafički prikaz rada neuronskih mreža za konsturkciju strojnog prijevoda[1]***

[1] Microsoft prevoditelj za posao. 2021. *Strojni prijevod-Microsoft Translator za tvrtke*. [online] dostupno na: <https://www.microsoft.com/hr-hr/translator/business/machine-translation/> [pristupila 20.05.2021].

## 2.2. Način rada Google AutoML Translate platforme

AutoML Translation jedan je od Google Cloud alata koji omogućava da programeri, prevoditelji i ostali stručnjaci s ograničenim znanjem o strojnom učenju mogu brzo stvoriti visokokvalitetne modele. Cjelokupni proces rada AutoML-a može se podijeliti u tri etape (prikazano na Slici 2):

- prijenos paralelnog korpusa u određenom jezičnom paru u sustav,
- treniranje modela koristeći AutoML Translation alat,
- evaluacija kreiranog modela i korištenje u praksi.




***Slika 2 Način rada AutoML platforme za prevođenje[2]***

Sustav također omogućuje korištenje API-a za prijevod različitih medijskih sadržaja. Translation API Basic trenutno prevodi tekstove na više od sto jezika. Translation API Advanced nudi iste brze, dinamične rezultate koji se dobivaju s osnovnim i dodatnim značajkama prilagodbe. Prilagođavanje je važno za pojmove ili fraze specifične za domenu i kontekst te formatirani prijevod dokumenta. S druge strane, API za prevođenje medija isporučuje audio prijevod u stvarnom vremenu s velikom točnošću i pojednostavljenom integracijom. Također može se

[2] Google Cloud. 2021. *AutoML Translation beginner's guide*. [online] dostupno na: <https://cloud.google.com/translate/automl/docs/beginners-guide> [pristupljeno 20 May 2021].

poboljšati korisničko iskustvo s prevođenjem s malim vremenom kašnjenja (Perron & Furnon, 2021.).

Značajke koje se spominju kao prednosti korištenja Google Cloud AutoML Translate-a su:

- omogućeno prevođenje izrazito velikog broja jezičnih parova,
- automatska detekcija jezika,
- podrška kreiranja glosara,
- visoka skalabilnost,
- jednostavna integracija,
- pristupačna cijena.

Unaprijed trenirani model sadržan u API-ju za prijevod, podržava više od stotinu jezika, od afričkog do zulu kao što je prikazano u Tablici 1.

***Tablica 1 - AutoML Translate podržani jezici u svibnju 2021. godine[3]***

| Language Pair | Language Codes |
|---|---|
| Afrikaans <-> English | af <-> en |
| Albanian <-> English | sq <-> en |
| Arabic <-> English | ar <-> en |
| Azerbaijani <-> English | az <-> en |
| Bengali <-> English | bn <-> en |
| Bulgarian <-> English | bg <-> en |
| Catalan <-> English | ca <-> en |
| Chinese (Simplified) <-> English | zh-CN * <-> en |
| Chinese (Traditional) <-> English | zh-TW <-> en |
| Croatian <-> English | hr <-> en |
| Czech <-> English | cs <-> en |
| Danish <-> English | da <-> en |

[3] Google Cloud. 2021. *Language support for custom models | AutoML Translation Documentation*. [online] dostupno na: <https://cloud.google.com/translate/automl/docs/languages> [pristupila 8.5.2021].

| | |
|---|---|
| Dutch <-> English | nl <-> en |
| Estonian <-> English | et <-> en |
| Finnish <-> English | fi <-> en |
| French <-> English | fr <-> en |
| Galician <-> English | gl <-> en |
| Georgian <-> English | ka <-> en |
| German <-> English | de <-> en |
| Greek <-> English | el <-> en |
| Gujarati <-> English | gu <-> en |
| Haitian Creole <-> English | ht <-> en |
| Hebrew <-> English | iw <-> en |
| Hindi <-> English | hi <-> en |
| Hungarian <-> English | hu <-> en |
| Icelandic <-> English | is <-> en |
| Indonesian <-> English | id <-> en |
| Italian <-> English | it <-> en |
| Japanese <-> English | ja <-> en |
| Korean <-> English | ko <-> en |
| Latvian <-> English | lv <-> en |
| Lithuanian <-> English | lt <-> en |
| Malay <-> English | ms <-> en |
| Marathi <-> English | mr <-> en |
| Norwegian <-> English | no <-> en |
| Persian <-> English | fa <-> en |
| Polish <-> English | pl <-> en |
| Portuguese <-> English | pt <-> en |

| | |
|---|---|
| Punjabi <-> English | pa <-> en |
| Romanian <-> English | ro <-> en |
| Russian <-> English | ru <-> en |
| Serbian <-> English | sr <-> en |
| Slovak <-> English | sk <-> en |
| Slovenian <-> English | sl <-> en |
| Spanish <-> English | es <-> en |
| Swahili <-> English | sw <-> en |
| Swedish <-> English | sv <-> en |
| Thai <-> English | th <-> en |
| Turkish <-> English | tr <-> en |
| Ukrainian <-> English | uk <-> en |
| Urdu <-> English | ur <-> en |
| Vietnamese <-> English | vi <-> en |
| Welsh <-> English | cy <-> en |

Što se tiče automatske detekcije jezika, ova značajka je od posebne koristi kod analize originalnog teksta za prevođenje u slučaju kada korisnik nije upoznat s jezikom ili pismom izvornog teksta. U tom slučaju AutoML ima mogućnost automatskog otkrivanja o kojem se izvornom jeziku radi. Platforma također omogućuje podršku kod izrade pojmovnika po potrebi korisnika, što prevoditelju omogućuje da održi konzistentnost i duh jezika u prevedenom sadržaju.

***Tablica 2 - Način rada AutoML Translation platforme - podržane značajke sustava[4]***

| | Api za prevodenje Basic | Napredni API za prijevod | AutoML prijevod |
|---|---|---|---|
| | Početak rada | Početak rada | Početak rada |
| Vrste sadržaja | Tekst, HTML | Tekst, HTML, DOCx, PPTx, XLSx, PDF | Tekstualna poruka |
| Primarne akcije | Prijevod teksta | Prijevod teksta | Model treniranja, upravljanja i prevođenja teksta |
| Prepoznavanje jezika | | ✓ | |
| Glosar | | ✓ | |
| Skupni prijevodi | | ✓ | |
| Oblikovani prijevod dokumenta (skupno i mrežno) | | ✓ | |
| Prijevodi s prilagođenim modelima | | ✓ | |
| Izrada prilagođenih modela | | | |
| Prijevod općim modelima | | ✓ | |
| Integrirani REST API | | ✓ | |
| Integrirani GRPC API | | | |
| Račun servisa | | ✓ | |
| API ključ | ✓ | | |
| HTML | ✓ | ✓ | |
| Prijevod neograničenog broja znakova dnevno | | ✓ | |
| Podržava više od 100 jezičnih parova | | ✓ | ✓ |

U ovom projektu korištene su značajke iz posljednjeg stupca Tablice 2. Prema vrsti sadržaja, korišten je tekst kao podržan oblik, a kao primarna aktivnost korišteno je treniranje modela i prevođenje teksta. Također je za projekt odabrana opcija kreiranje korisničkog računa, prevođenje s modelom izrađenim po potrebi korisnika, kreiranje i izgradnja korisničkog modela

[4] Google Cloud. 2021. *Cloud Translation | Google Cloud*. [online] dostupno na: <https://cloud.google.com/translate#section-7> [pristupila 2.5.2021.]

AutoML platforma omogućuje skaliranje prijevodnih proizvoda. Na korisniku je da postavi pravila rada u prevoditeljskom timu, dok se skupni prijevod može koristiti s Google Cloud Storageom kako bi se smanjila složenost tijeka rada pri prevođenju dugih ili više tekstualnih datoteka u većim timovima.

Sljedeća značajka je jednostavna integracija tj. Google REST API koji se lako koristi za prijevod, što znači da se ne mora izdvajati tekst iz dokumenta; već se sustavu samo pošalje HTML datoteka, a korisnik natrag dobiva prevedeni tekst.

Što se tiče naplaćivanja, prijevod se naplaćuje po znaku, čak i ako je znak zapisan pomoću više bajtova. AutoML Translation naplaćuje učenje modela, upotrebu predviđanja po znakovima i pohranu. Plaća se samo ono što se u sustavu koristi, dok neiskorištene opcije sustava ostaju nenaplaćene. Cjenik treniranja modela na AutoML Translation platformi prikazan je u Tablici 3, a cjenik prevođenja za istu platformu u Tablici 4.

***Tablica 3 Cijenik treniranja modela na AutoML Translation platformi u svibnju 2021. godini***[5]

| Broj parova za treniranje | Tipično vrijeme treniranja |
|---|---|
| Manje od 1000 | 2 - 3 sata |
| 1000 - 10,000 | 2 - 3 sata |
| 10,001 - 100,000 | 4 - 5,5 sati |
| 100,001 - 1,000,000 | 5 - 7 sati |
| 1,000,001 - 10,000,000 | 6 - 12 sati |
| Više od 10.000.000 | 12 sati ili više |

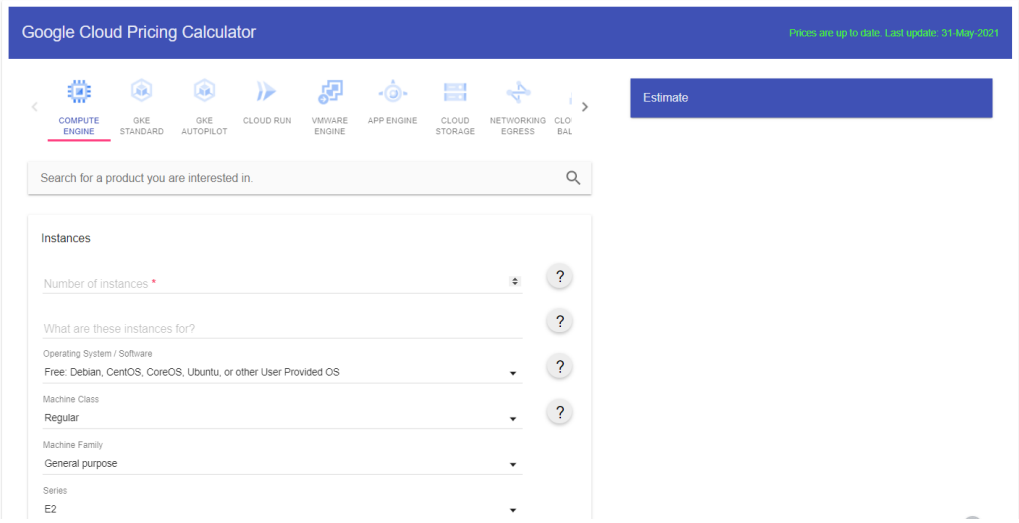
| AutoML Sku | Po satu treniranja | Maksimalno po treniranju |
|---|---|---|
| AutoML treniranje | $45 | $300 |

[5] Google Cloud. 2021. *Pricing | AutoML Translation Documentation | Google Cloud*. [online] dostupno na: <https://cloud.google.com/translate/automl/pricing> [pristupljeno 2.5.2021.]

***Tablica 4 Cijenik prevođenja na AutoML Translation platformi u svibnju 2021. godini***

| Značajka | Mjesečna upotreba | Cijena |
|---|---|---|
| Prijevod teksta (TXT, HTML i XLSX formati) | 0-500k | Besplatno |
| | 500k-250M | $80 za milijun znakova |
| | 250M-2,5B | $60 za milijun znakova |
| | 2,5B-4B | $ 40 za milijun znakova |
| | >4B | $ 30 za milijun znakova |
| Prijevod dokumenta (pretpregled) pomoću v3 API-ja za prevođenje u oblaku (podržava DOCX, PPTX i PDF formate) | sva upotreba | $ 0.25 po stranici |

Tablice 3 i 4 prikazuju cjenik koji je aktualan za svibanj 2021. godine. Platforma zahtjeva unos podataka bankovne kartice prije samog korištenja zbog čega je kod planiranja troškova projekta bitno uzeti u obzir sve komponente koje će za izvedbu biti korištene. Platforma također nudi opciju procjene cijene Google Cloud projekta na web stranici Google Cloud Pricing Calculator (Slika 3) koja omogućava korisnicima lakše planiranje budžeta projekta.




***Slika 3 Google Cloud Pricing Calculator - procjena cijene Google Cloud projekta[6]***

[6] Google Cloud. 2021. *Google Cloud Platform Pricing Calculator*. [online] dostupno na: <https://cloud.google.com/products/calculator> [pristupila 21.5.2021.].

# 3. Metodologija

## 3.1. Problem i predmet istraživanja

Predmet istraživanja rada je neuronsko strojno prevođenje na primjeru korištenja AutoML Translate platforme Google Cloud sustava za englesko – hrvatski jezični par. Problematika kojom se ovaj rad bavi je kreiranje prevoditeljske memorije kroz različite alate za poravnavanje postojećih izvornih i prevedenih tekstova te stvaranje iste pomoću programa za potpomognuto prevođenje i strojno prevođenje. Osim navedenog, provodi se analiza rada AutoML Translate platforme koja primjenom umjetne inteligencije omogućava korisniku stvaranje i treniranje modela izrađenih prema potrebama korisnika te se dolazi do zaključaka vezanih za rad platforme za odabranu jezičnu kombinaciju.

## 3.2. Svrha i cilj istraživanja

Svrha i cilj istraživanja je odrediti najbolje postupke pripreme paralelnog korpusa na osnovu koje AutoML Translate obrađuje podatke i trenira model za neuronsko strojno prevođenje kako bi isti dao što bolje rezultate primjenjive u prevoditeljskoj praksi.

## 3.3. Radna hipoteza

Provedeno je treniranje pet različitih modela na AutoML Translate platformi od kojih su modeli bili trenirani na:

- prijevodnoj memoriji tekstova Williama Branhama poravnanoj na:
    - razini rečenice
    - razini odlomka
- prijevodna memorija Biblije poravnana na:
    - razini stihova
- prijevodna memorija tekstova Williama Branhama i Biblije poravnana na:
    - razini rečenice

- razini odlomka

Pretpostavlja se da će najbolji rezultati dati sustav treniran na stopostotno točno poravnanoj prevoditeljskoj memoriji kojoj svaki rečenični segment na originalnom engleskom jeziku pored sebe ima odgovarajući prijevod na hrvatski jezik, a sadržaj je tekstova jednog autora i ne dolazi do kombiniranja sa prijevodom Biblije.

## 3.4. Opis korpusa

Tmx datoteka memorije prijevoda tekstova Williama Branhama sadrži 63728 rečeničnih parova, a struktura prevoditeljske memorije poravnana na način da lijeva strana korespondira prijevodu na desnoj strani. Memorija je nastala koristeći program Trados na način da su se povezale radne memorije više različitih prevoditelja u jednu sveobuhvatnu memoriju koja je iz tmw formata morala biti pretvorena u tmx format. Tsv datoteka memorije prijevoda tekstova Williama Branhama i prijevoda Biblije sadrži 90902 jezičnih parova od kojih su tekstovi Williama Branhama proavnani na razini rečenica, a Biblijski tekstovi na razini stihova, što nerijetko uključuje dvije do tri rečenice u jednom stihu. Sljedeća datoteka memorije je tmx datoteka koja uključuje prijevode tekstova Williama Branhama poravnanima na razini odlomka zajedno s prijevodom Biblije koja je kao i u prethodnom slučaju poravnana na razini stihova. Datoteka se sastoji od 61147 jezičnih parova. Sljedeća datoteka prijevodne memorije se sastoji od 24996 jezičnih parova prijevoda Biblije poravnatih na razini stihova. Posljednja memorijska datoteka korpusa je tmx datoteka koja sadržajno odgovara prvoj prevoditeljskoj memoriji korpusa, međutim, u ovom slučaju se proces poravnavanja izvršio na razini odlomka zbog čega se ova prijevodna memorija sastoji od 36151 jezičnih parova.

## 3.5. Metodologija i tijek istraživanja

Tijek istraživanja se je sastojao od uređivanja prevoditeljskih memorija, korištenja alata za uređivanje baze podataka prevedenih tekstova i pretvorbu prevoditeljskih memorija iz jednih oblika datoteka u druge, treniranja modela koristeći AutoML Translate platformu, te završilo s uspoređivanjem rezultata različitih modela. Kvaliteta određenog modela bila je izražena ponajprije BLEU metrikom, ali također se je uzela u obzir i razlika između BLEU metrike početnog Google modela prirodnog strojnog prijevoda (eng. *NMT – natural machine translation*) i nastalog korisničkog modela u svakom od primjera te su se dodatno komentirale primjene različitih opcija u prevoditeljskoj praksi.

# 4. Prijevodna memorija i alati za njezino oblikovanje

## 4.1. Prijevodna memorija

Prevođenje s jednog na drugi jezik tehnika je koju nije lako staviti u okvire strogih pravila što je krajem 20. stoljeća potaklo potrebu za novim metodama strojnog prevođenja. Postavlja se pitanje je li moguće da stroj uči iz već prevedenih tekstova. Istraživanja na temu strojnog prevođenja otvorila su nova moguća rješenja uglavnom temeljenih na podatcima (eng. data-driven methods). Iz namjere da se na temelju već prevedenih dijelova rečenica dođe do novih prijevoda došlo je do razvijanja strojnog prevođenja temeljenog na primjerima (eng. example-based machine translation). Krajem 20. stoljeća kada se pojavila navedena metoda u istraživanjima, nije bio velik broj istraživača koji su bili zainteresirani istom, međutim danas je upravo ovaj pristup sve u primjeni (Đunđer, 2015.).

Temeljna ideja prijevodne memorije (eng. translation memory) je pohrana jezičnih segmenata na izvornom jeziku i paralelno spremanje prijevoda na drugom jeziku. Ova tehnologija ima široku primjenu u procesu prevođenja koji je računalno potpomognut. Prijevodna memorija osim funkcije pohrane podataka također i pretražuje postojeće prijevode u bazi podataka te nudi prevoditelju adekvatne prijevode pri radu na novim tekstovima (Đunđer, 2015.).

Postupak pripreme prevoditeljske memorije bio je od ključne važnosti u ovom projektu. Bitno je napomenuti da se tijekom izrade diplomskog rada paralelno radilo na stvaranju prevoditeljske memorije prijevoda Biblije te na procesu izdvajanja prevoditeljske memorije tekstova Williama Branhama koja se kroz prevoditeljsko iskustvo automatski pohranjivala u programu Trados.

U prvom procesu stvaranja prevoditeljske memorije bitno je bilo točno poravnavanje originalnog teksta Biblije prijevoda Kralja Jakova s hrvatskim prijevodom Ivana Vrtarića iz 2016. godine kako ne bi došlo do krivog poklapanja lijeve i desne strane, točnije izvornog teksta s prevedenim tekstom. Uz pomoć korištenih alata koji su objašnjeni u nastavku rada, potrebna je bila ljudska intervencija i korekcija određenih automatskih funkcionalnosti programskih alata kako bi poravnavanje bilo točno. Budući da je kod izrade prevoditeljske memorije Biblije izvorni tekst već postojeći, a također i hrvatski prijevod, bilo je potrebno učitati izvorni i odredišni tekst u program Excel na način da broj stiha Biblije s lijeve strane odgovara broju stiha Biblije s desne

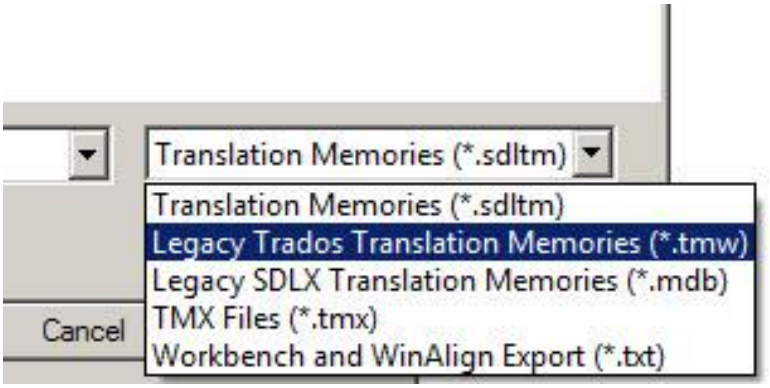
strane. Na taj se je način, korištenjem Excel Query opcija došlo do kranjeg produkta poravnane memorije.

U drugom navedenom procesu stvaranja prevoditeljske memorije tekstova Williama Branhama ponajviše se koristio program Trados u kojem prevoditelji rade na vjerskim tekstovima duži niz godina te nove prijevode automatski unose u kreiranje prevoditeljske memorije. Međutim u ovom se je slučaju arhivirana prijevodna memorija morala uređivati i mijenjati formate. Tada je bilo bitno dobiti krajnji produkt u tsv ili tmx datotekama budući da su one podržane od strane Google AutoML za treniranje modela za neuronsko strojno prevođenje. Memorija se poravnavala u rečeničnim oblicima kako bi se dobila što točnija semantička analiza teksta te kako bi trenirani model mogao u fazi evaluacije dati rezultate što točnijih izraza specifične terminologije vjerskih tekstova.

### 4.1.1. Prijevodna memorija sustava Trados

Kada se govori o prevoditeljskoj memorij, govori se o memoriji koja predstavlja bazu podataka pohranjenih rečenica, odlomaka ili segmenata teksta koji su ranije prevedeni. Svaki unos ili segment u prijevodnoj memoriji uključuje izvorni jezik, poznat kao *izvor* i njegov prijevod, poznat kao *cilj*. Ti se parovi nazivaju prijevodnim jedinicama ili TU (eng. *translation units*). Prijevodne memorije koriste se s prijevodnim softverom, kao što je Trados Studio, i automatski predlažu pohranjena identična ili slična podudaranja kako se prevode novi dokumenti. To znači da rečenice, odlomci ili segmenti teksta koji su prethodno prevedeni nikada više ne trebaju biti iznova prevedeni. Prijevodne memorije dramatično poboljšavaju kvalitetu, brzinu, dosljednost i učinkovitost svakog prevoditeljskog posla, a koriste se kao dio alata za računalno potpomognuto prijevođenje kao što je Trados Studio.

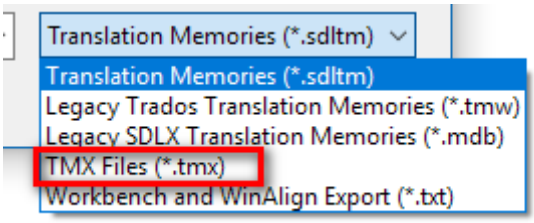
Prilikom otvaranja dokumenta koji se prevodi ili izvorne datoteke, prijevodna memorija provjerava je li bilo koji sadržaj ranije već preveden i traži stopostotna podudaranja - identična podudaranja ili djelomična podudaranja - slična, ali ne i točna podudaranja koja se pojavljuju u novoj izvornoj datoteci. Svaki prethodno spremljeni prevedeni tekst smješta se unutar odgovarajućeg ciljnog segmenta unutar prozora za uređivanje teksta. Dok prevoditelj radi na izvornoj datoteci, svi prijedlozi iz prijevodne memorije mogu biti prihvaćeni ili nadjačani novim alternativama. Ako se prijevodna jedinica (segment izvornog i ciljnog teksta) ručno ažurira, može se spremiti u prijevodnu memoriju za buduću upotrebu. Na sličan način, svi segmenti u ciljanoj datoteci bez podudaranja trebali bi biti prevedeni ručno. Jednom prevedeni, oni se tada mogu dodati u sve veću prijevodnu memoriju.



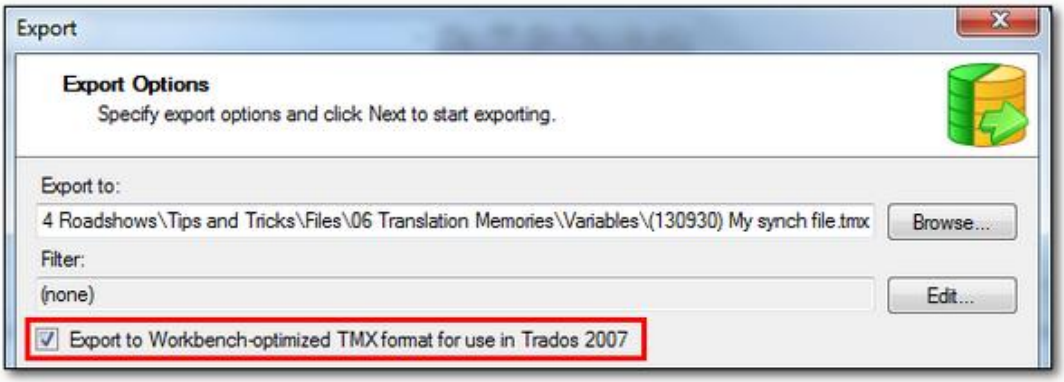

***Slika 4 Trados prijevodna memorija – tmw***

Kod korištenja softvera Trados Studio, prijevodne memorije stvara profesionalni prevoditelj i ponovno koristi sadržaj koji je prethodno preveden u tmw datoteci (Slika 4) kako bi se poboljšala brzina, kvaliteta i dosljednost budućih prijevoda. Novije verzije programa pokušavaju implementirati strojno prevođenje kao mogućnost dobivanja automatskog prijevoda bez ikakvog ljudskog unosa. Trados Studio nudi opcije strojnog prevođenja za ograničeni broj jezika, međutim kod istoga se dodatno radi na uređivanju prijevoda od strane prevoditelja.

Strojno prevođenje se može također koristiti zajedno s prijevodnom memorijom kako bi se povećala ukupna brzina prijevoda. Takva se kombinacija, primjerice, koristi kada prijevodna memorija nema dovoljno podataka za dovršavanje segmenta teksta, pa pružatelj strojnog prijevoda, kao što je RWS Language Cloud, može dati prijedlog za neprevedeni segment. Prevoditelj zatim naknadno uređuje rezultat strojnog prijevoda i sprema dovršeni segment u prijevodnu memoriju za ponovnu upotrebu. Prijevodna memorija pohranjuje segmente teksta kao prijevodne jedinice. Segment se može sastojati od rečenice ili odlomka. Prijevodna memorija (*eng. translation memory* ili TM) sadrži i izvornu i prevedenu verziju svakog segmenta za ponovnu upotrebu. Poput prijevodne memorije, baza termina je paralelni korpus koja se može pretraživati. Međutim, baza termina sadrži popis višejezičnih pojmova i pravila u vezi s njihovom uporabom. Pojmovi mogu biti bilo što specifično za tvrtku, kao što je način na koji se prevode nazivi robnih marki ili imena proizvoda, ili ih ne treba prevesti. Prevoditelji obično koriste i prijevodnu memoriju i bazu termina u Trados Studiju (Lewey, 2021).

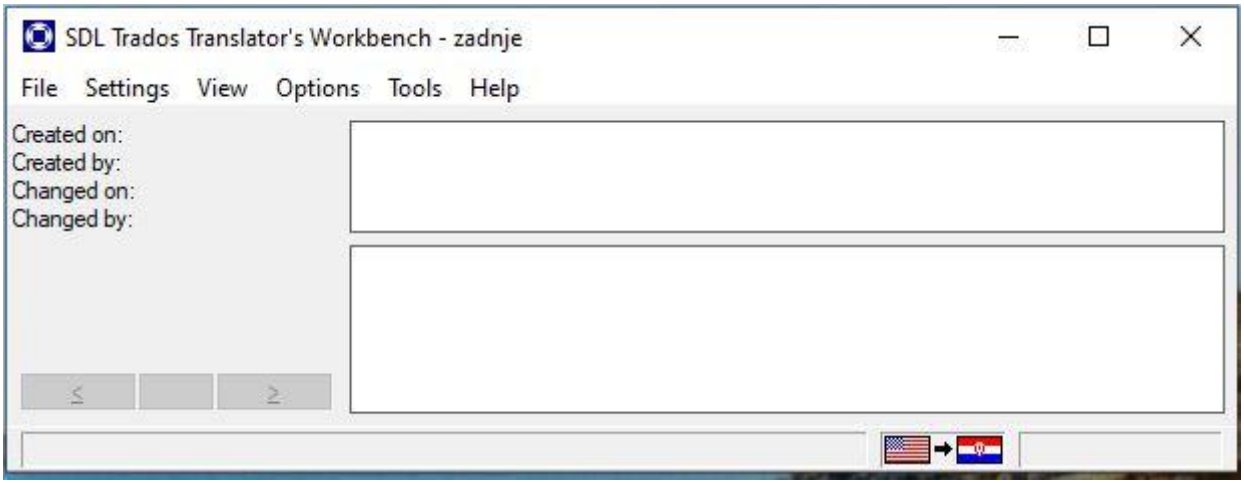



***Slika 5 Trados prijevodna memorija – tmx***




***Slika 6 Trados prijevodna memorija - Export odgovarajućeg TMX formata za korištenje u starijim verzijama programa***

Ono što je u provedbi projekta ovog rada bilo od velike važnosti je mogućnost izvoza prevoditeljske memorije iz programa Trados u odgovarajući TMX format (Slike 5 i 6) koji se može koristiti i u starijim verzijama programa. Važnost ove opcije leži i u tome da se kod povezivanja prevoditeljskih memorija više prevoditelja u jednu smislenu cjelinu često verzije programa u praksi ne poklapaju pa je od velike važnosti imati fleksibilnost kod pretvorbe i formatiranja kako bi povezane memorije stvarale koherentnu cjelinu.

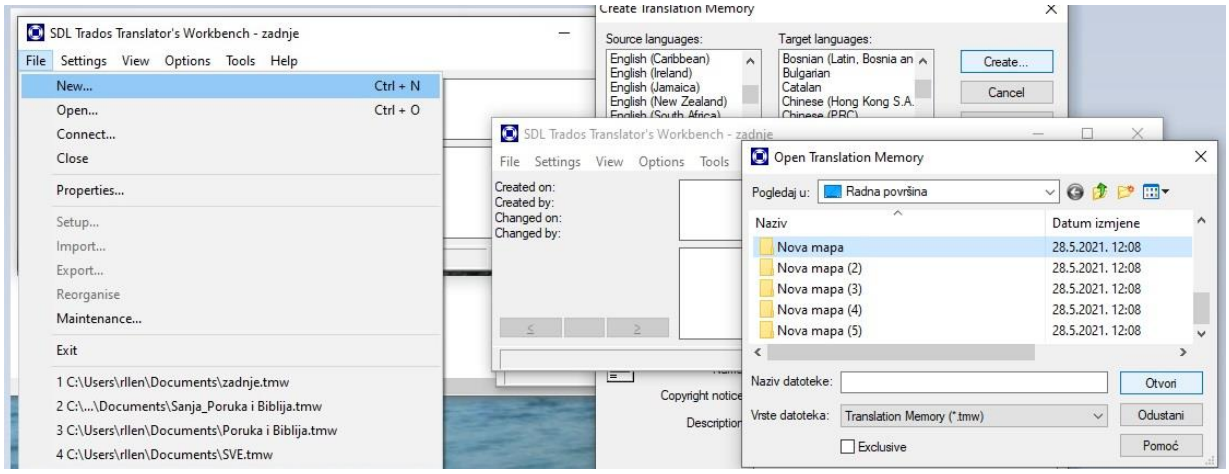



***Slika 7 SDL Trados Translator's Workbench***

SDL Trados Workbench (Slika 7) radni je prostor programa u kojem se mogu stvarati nove prevoditeljske memorije, otvarati postojeće, a ponajviše se koristi za prevoditeljsku konkordancu

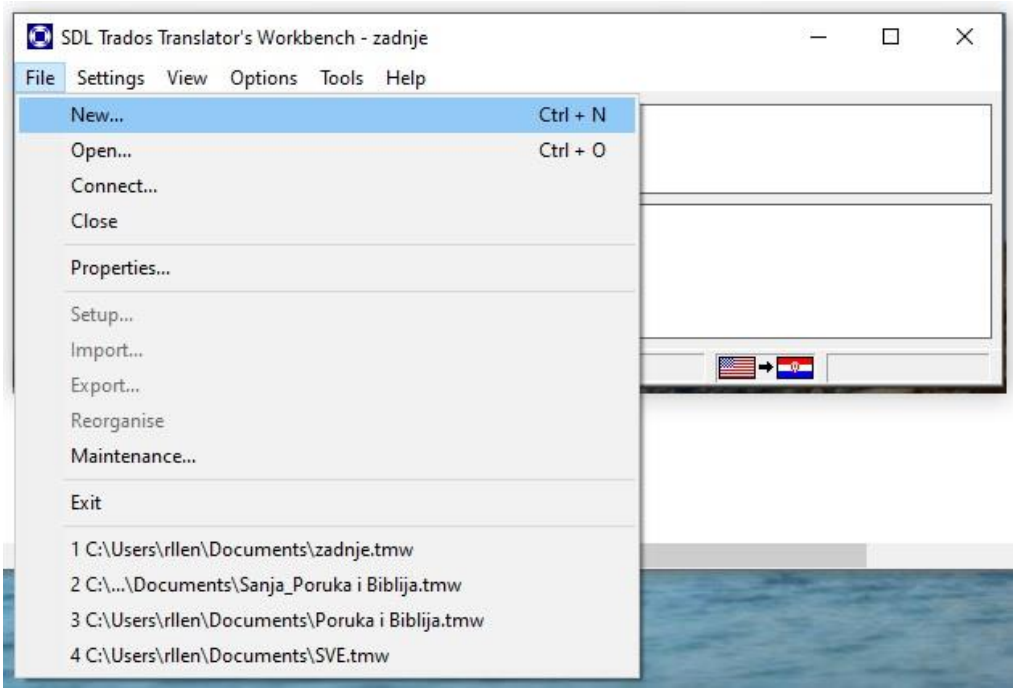
tj. pretraživanje pojma u prevedenim rečenicama odabrane prevoditeljske memorije što je prevoditeljima od velike pomoći budući da se direktno dobiva informacija ne samo o značenju riječi, već o njezinom korištenju i kontekstima u kojima je već u selektiranoj memoriji prevedena.

Otvaranje postojeće prevoditeljske memorije u SDL Trados Translator's Workbench-u ili stvaranje nove preduvjet je za nastavak korištenja Trados SDL opcija u prevođenju. Kod otvaranja postojeće memorije u ovom slučaju govorimo o otvaranju tmw datoteke u koju su prethodno pohranjeni prijevodi sličnog sadržaja i terminologije. Kako bi se ta memorija mogla koristiti za stvaranje modela primjenom umjetne inteligencije na Google Cloud platformi neophodno je tmw memoriju prebaciti u tmx format što je moguće kroz opciju Export u SDL Trados Translator's Workbench-u (prikazano na Slici 8).

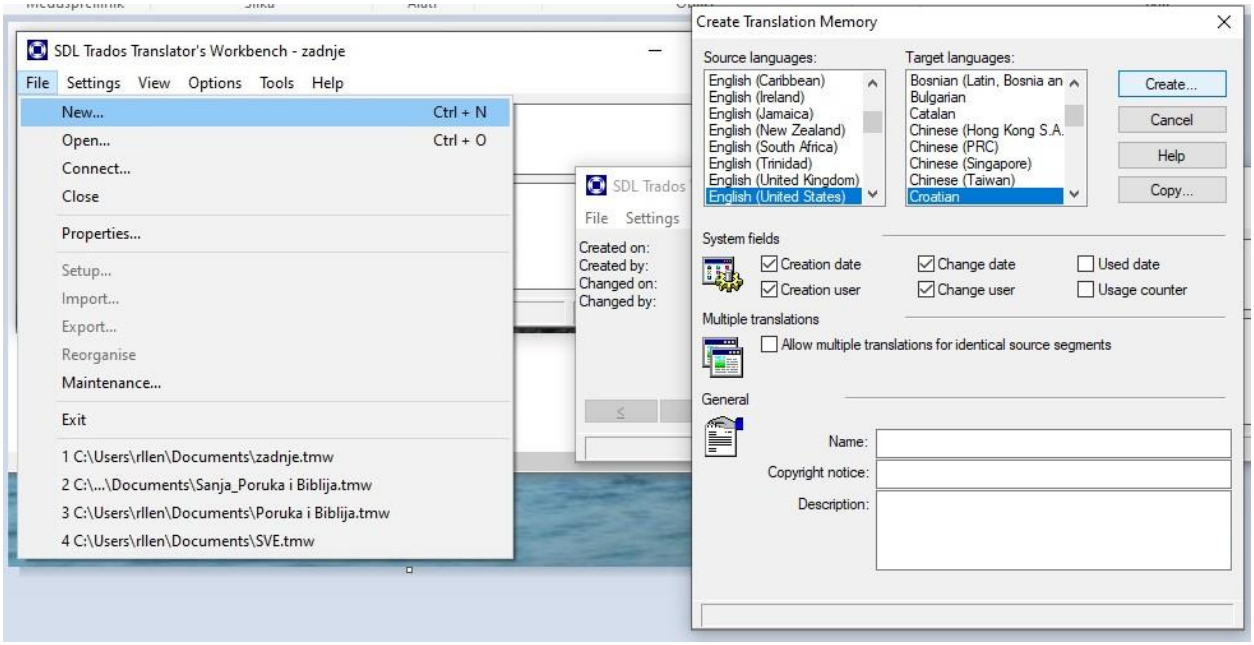



***Slika 8 SDL Trados Translator's Workbench - otvaranje postojeće prevoditeljske memorije***

Stvaranju nove prevoditeljske memorije (Slike 9 i 10) u programu, prethodi detaljna analiza jezične strukture tekstova na kojima se prevodi. Kontekstualna i jezična analiza teksta dozvoljavaju bolju i efikasniju odredišnu prevoditeljsku memoriju koja će terminologijom pripadati istoj jezičnoj tematici te na taj način dobivamo uredne prevoditeljske memorije koje kasnijim korištenjem daju bolje rezultate u neuronskom strojnom prevođenju budući da modeli trenirani na takvim memorijama imaju jasnu strukturu i koncept koji omogućava kvalitetno učenje budućeg modela.



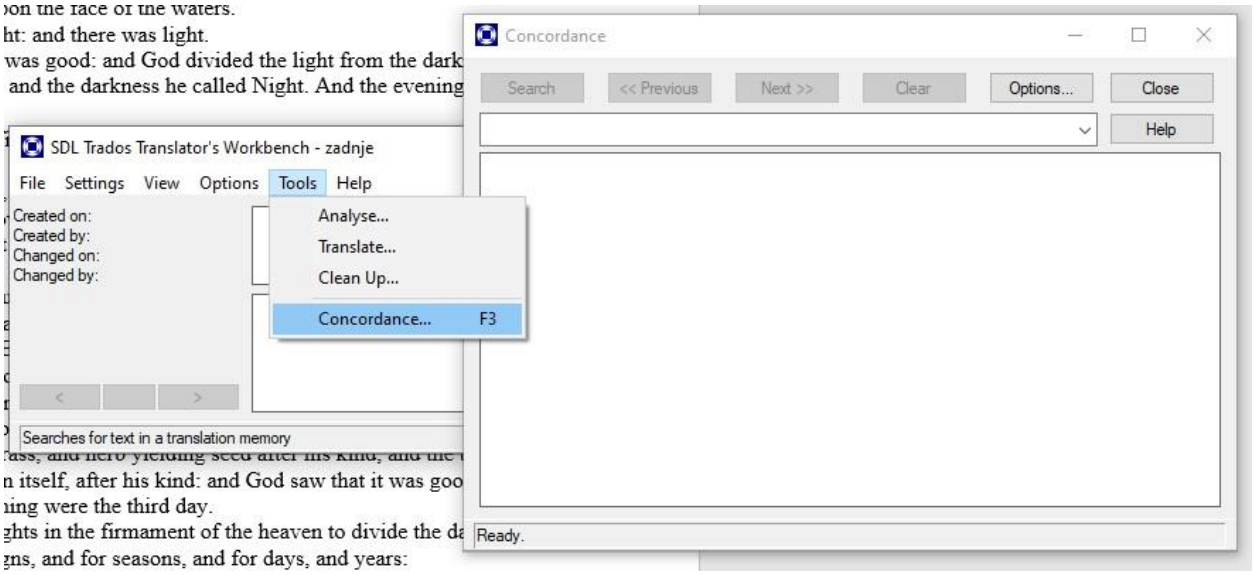

***Slika 9 SDL Trados Translator's Workbench - stvaranje prevoditeljske memorije 1***




***Slika 10 SDL Trados Translator's Workbench - stvaranje prevoditeljske memorije odabir izvornog i odredišnog jezika***

Drugi razlog zbog kojeg je dobro definirati karakteristike prevoditeljske memorije unaprijed je mogućnost dodavanja višestrukih prijevoda za isti izvorni segment i određivanje sistemskog dijela polja u stvorenoj memoriji za ubuduće korištenje. Također je bitno napomenuti spominjanje izvornih autora stvaranja datoteka pod opcijom *Copyright notice* zbog zaštite autorskih prava.

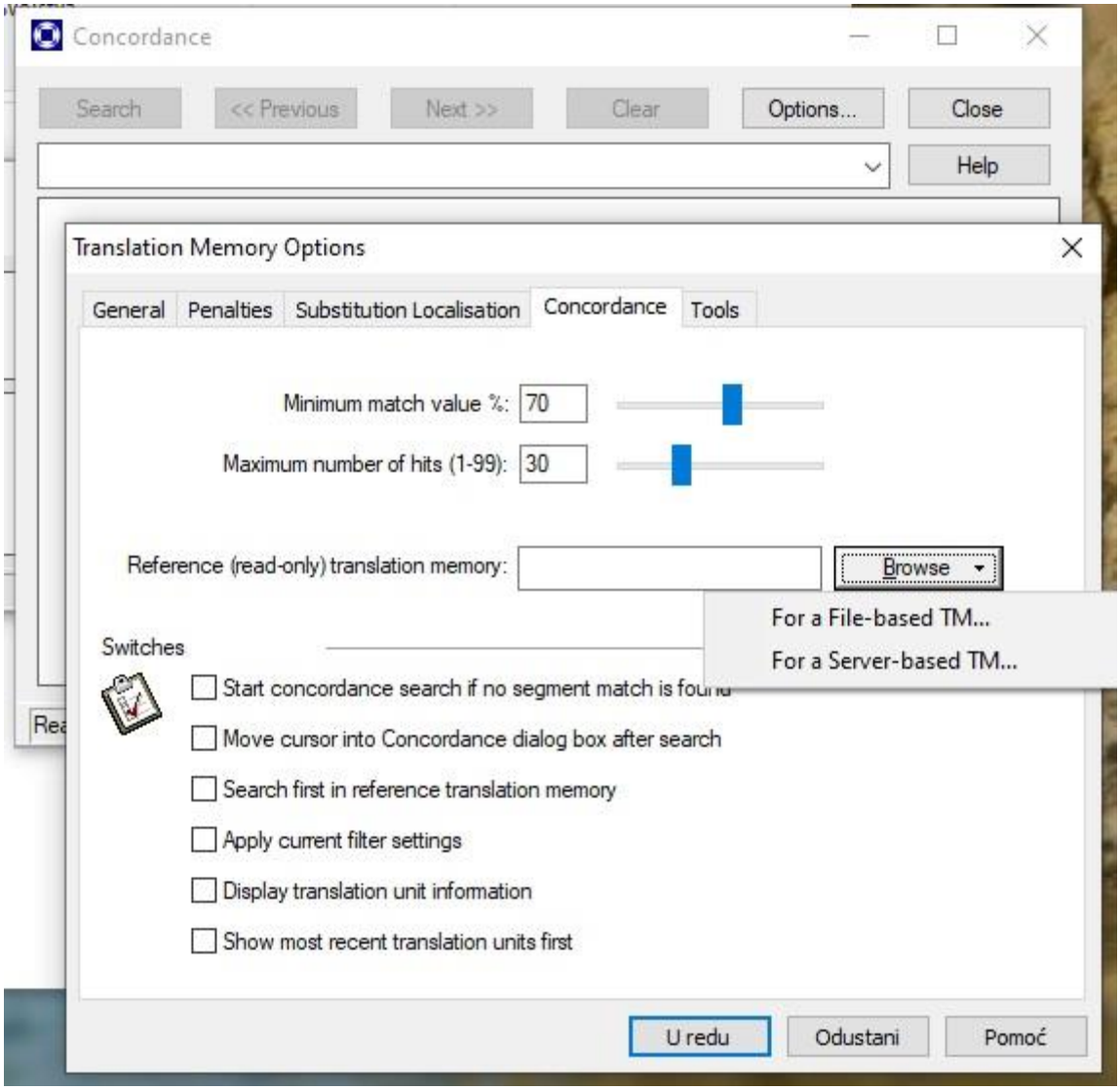
SDL Trados nudi korisnicima opciju pretraživanja pojmova u odabranoj prevoditeljskoj memoriji. Bitno je spomenuti opciju konkordance (Slika 11) budući da je često korišteni alat kod prevoditelja.Njezino ispravno i brzo pretraživanje memorije nije temeljeno samo na automatskim postavkama softverskog sustava, već prevoditeljima pruža i mogućnost podešavanja opcija prema vlastitoj potrebi.




***Slika 11 SDL Trados Translator's Workbench – Kokordanca***

Radni prostor softverskog rješenja za prevoditelje sustava Trados sugerira prevoditeljima na osnovi prevoditeljske memorije moguće prijevode za rečenice čije segmente uspije pronaći u memoriji. Kod korištenja ove opcije u praktičnom prevođenju, mogu se odabrati dva načina povezivanja s bazičnom bazom podataka prijevodne memorije, a to su ili memorija bazirana na datoteci ili memorija bazirana na serveru. Kad se odabere jedna od opcija, prevoditelj u nastavku može odrediti na osnovi kolikog postotka podudaranja želi da mu se tijekom prevođenja sugeriraju

nove mogućnosti za prijevode. Program također nudi određena podešavanja za pretragu konkordance i referenciranje na prevoditeljsku memoriju koja su vidljiva na Slici 12.




***Slika 12 SDL Trados Transaltor's Workbench - opcije podešavanja konkordance***

### 4.1.2. Priprema prijevodne memorije Biblije korištene u istraživanju

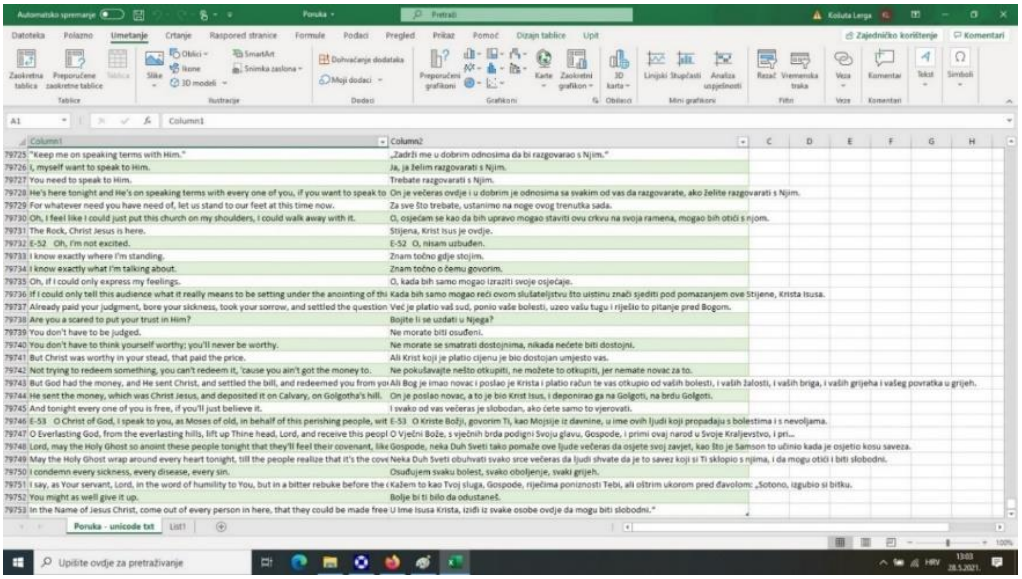
Stvaranje prevoditeljske memorije za prijevod Biblije bio je jedan od zahtjevnijih zadataka cjelokupnog istraživanja. Za početni temeljni tekst na engleskom jeziku odabran je prijevod Biblije Kinga Jamesa budući da je to jedini prijevod Biblije citiran u tekstovima Williama Branhama. Hrvatski prijevod Biblije korišten u ovom slučaju je prijevod Ivana Vrtarića iz 2016. godine za koji sam prevoditelj u predgovoru piše: „Ovaj prijevod Biblije rađen je s čvrstim uvjerenjem da je tekst izvornih rukopisa Staroga i Novoga zavjeta (autografa) sačuvan u prijepisima tradicionalnoga teksta koji su nam do danas sačuvani Božjim proviđenjem, a ne da su nadahnuti i nepogrešivi bili samo izvorni rukopisi. Jer: "Riječi GOSPODINOVE *su* riječi čiste; *kao* srebro u zemljanoj peći pretopljeno, sedam puta pročišćeno. Ti ćeš ih čuvati, GOSPODINE; ti ćeš ih očuvati od ovoga naraštaja zauvijek" (Psalam 12,6.7).

Kao uzor i pomoć u prevođenju s izvornih jezika poslužili su znameniti povijesni prijevodi utemeljeni na tradicionalnim tekstovima Staroga i Novoga zavjeta: *The King James Bible*, češka *Kralická*, španjolska *Reina Valèra* i ruska *Sinodalna* Biblija. Od koristi su bili i neki drugi strani prijevodi koji ne slijede istovjetnu tekstualnu tradiciju, primjerice njemačka nerevidirana *Elbelfelderska* i još neki slavenski, engleski i njemački prijevodi. Namjera je bila sačiniti Bibliju na hrvatskom književnom jeziku koja bi bila što točniji prijevod izvornih tekstova, i to prijevod načinjen prema načelu formalne, a ne dinamičke ekvivalencije (parafraza). Poštovanje prema božanskom Autoru i Njegovoj riječi vodilo je prevoditelja da prihvatljivim smatra samo načelo krajnje doslovnosti i točnosti, a ne prenošenje poruke teksta." (Vrtarić, 2021)

### 4.1.2. Korištenje Excel Queryja za rad na prevoditeljskoj memoriji

Kod izrade prevoditeljskih memorija u praksi se pokazalo da je Excel osim što je općepoznato izrazito dobar za obradu brojčanih podataka, također dobar i kod rada s tekstovima. Način na koji je podatkovno organizirana struktura Biblije također je utjecao na odabir upravo Excel-a za rad na prevoditeljskoj memoriji.

Biblijski tekstovi organizirani su u knjige koje sadrže poglavlja i brojeve stihova unutar poglavlja. Ovakva organizacija idealna je za oblikovanje u ćelijama, kao što to omogućava Excel, budući da je od logičnih jedinica prevoditeljske memorije jasno vidljivo da broj poglavlja knjige u izvornom tekstu mora odgovarati broju poglavlja biblijske knjige u prevedenom tekstu. Također, svaki stih koji je određen brojkom s lijeve strane, u vizualnom prikazu u ćelijama, mora odgovarati identičnom broju s desne strane. Na ovaj način strukturiran tekst brzo se poravnao, međutim neophodno je bilo ljudska intervencija u nekoliko slučajeva u kojima je došlo do spajanja stihova u istu ćeliju što se moglo uočiti i brzo ispraviti da se dobije poravnani korpus originalnog teksta Biblije na engleskom jeziku i hrvatskom prijevodu iz 2016. godine. U projektu se Excel koristio kod izrade prijevodne memorije Biblije budući da je na raspolaganju bio i izvorni tekst i cjelokupni prijevod Biblije na hrvatskom jeziku Ivana Vrtarića. Slika 25 prikazuje gotov proces poravnavanja memorije na primjeru Biblije. Stih s lijeve strane odgovara prijevodu stiha s desne strane te se poklapaju dijakritički znakovi izvornog i odredišnog teksta.



| Column1 | Column2 |
|---|---|
| 79725 "Keep me on speaking terms with Him." | „Zadrži me u dobrim odnosima da bi razgovarao s Njim." |
| 79726 I, myself want to speak to Him. | Ja, ja želim razgovarati s Njim. |
| 79727 You need to speak to Him. | Trebate razgovarati s Njim. |
| 79728 He's here tonight and He's on speaking terms with every one of you, if you want to speak to | On je večeras ovdje i u dobrim je odnosima sa svakim od vas da razgovarate, ako želite razgovarati s Njim. |
| 79729 For whatever need you have need of, let us stand to our feet at this time now. | Za sve što trebate, ustanimo na noge ovog trenutka sada. |
| 79730 Oh, I feel like I could just put this church on my shoulders, I could walk away with it. | O, osjećam se kao da bih upravo mogao staviti ovu crkvu na svoja ramena, mogao bih otići s njom. |
| 79731 The Rock, Christ Jesus is here. | Stijena, Krist Isus je ovdje. |
| 79732 E-52 Oh, I'm not excited. | E-52 O, nisam uzbuđen. |
| 79733 I know exactly where I'm standing. | Znam točno gdje stojim. |
| 79734 I know exactly what I'm talking about. | Znam točno o čemu govorim. |
| 79735 Oh, if I could only express my feelings. | O, kada bih samo mogao izraziti svoje osjećaje. |
| 79736 If I could only tell this audience what it really means to be setting under the anointing of thi | Kada bih samo mogao reći ovom slušateljstvu što uistinu znači sjediti pod pomazanjem ove Stijene, Krista Isusa. |
| 79737 Already paid your judgment, bore your sickness, took your sorrow, and settled the question | Već je platio vaš sud, ponio vaše bolesti, uzeo vašu tugu i riješio to pitanje pred Bogom. |
| 79738 Are you a scared to put your trust in Him? | Bojite li se uzdati u Njega? |
| 79739 You don't have to be judged. | Ne morate biti osuđeni. |
| 79740 You don't have to think yourself worthy; you'll never be worthy. | Ne morate se smatrati dostojnima, nikada nećete biti dostojni. |
| 79741 But Christ was worthy in your stead, that paid the price. | Ali Krist koji je platio cijenu je bio dostojan umjesto vas. |
| 79742 Not trying to redeem something, you can't redeem it, 'cause you ain't got the money to. | Ne pokušavajte nešto otkupiti, ne možete to otkupiti, jer nemate novac za to. |
| 79743 But God had the money, and He sent Christ, and settled the bill, and redeemed you from you | Ali Bog je imao novac i poslao je Krista i platio račun te vas otkupio od vaših bolesti, i vaših žalosti, i vaših briga, i vaših grijeha i vašeg povratka u grijeh. |
| 79744 He sent the money, which was Christ Jesus, and deposited it on Calvary, on Golgotha's hill. | On je poslao novac, a to je bio Krist Isus, i deponirao ga na Golgoti, na brdu Golgoti. |
| 79745 And tonight every one of you is free, if you'll just believe it. | I svako od vas večeras je slobodan, ako ćete samo to vjerovati. |
| 79746 E-53 O Christ of God, I speak to you, as Moses of old, in behalf of this perishing people, wit | E-53 O Kriste Božji, govorim Ti, kao Mojsije iz davnine, u ime ovih ljudi koji propadaju s bolestima i s nevoljama. |
| 79747 O Everlasting God, from the everlasting hills, lift up Thine head, Lord, and receive this peopl | O Vječni Bože, s vječnih brda podigni Svoju glavu, Gospode, i primi ovaj narod u Svoje Kraljevstvo, i pri... |
| 79748 Lord, may the Holy Ghost so anoint these people tonight that they'll feel their covenant, like | Gospode, neka Duh Sveti tako pomaže ove ljude večeras da osjete svoj zavjet, kao što je Samson to učinio kada je osjetio kosu saveza. |
| 79749 May the Holy Ghost wrap around every heart tonight, till the people realize that it's the cov | Neka Duh Sveti obuhvati svako srce večeras da ljudi shvate da je to savez koji si Ti sklopio s njima, i da mogu otići i biti slobodni. |
| 79750 I condemn every sickness, every disease, every sin. | Osuđujem svaku bolest, svako oboljenje, svaki grijeh. |
| 79751 I say, as Your servant, Lord, in the word of humility to You, but in a bitter rebuke before the | Kažem to kao Tvoj sluga, Gospode, riječima poniznosti Tebi, ali oštrim ukorom pred đavolom: „Sotono, izgubio si bitku. |
| 79752 You might as well give it up. | Bolje bi ti bilo da odustaneš. |
| 79753 In the Name of Jesus Christ, come out of every person in here, that they could be made free | U Ime Isusa Krista, izađi iz svake osobe ovdje da mogu biti slobodni." |

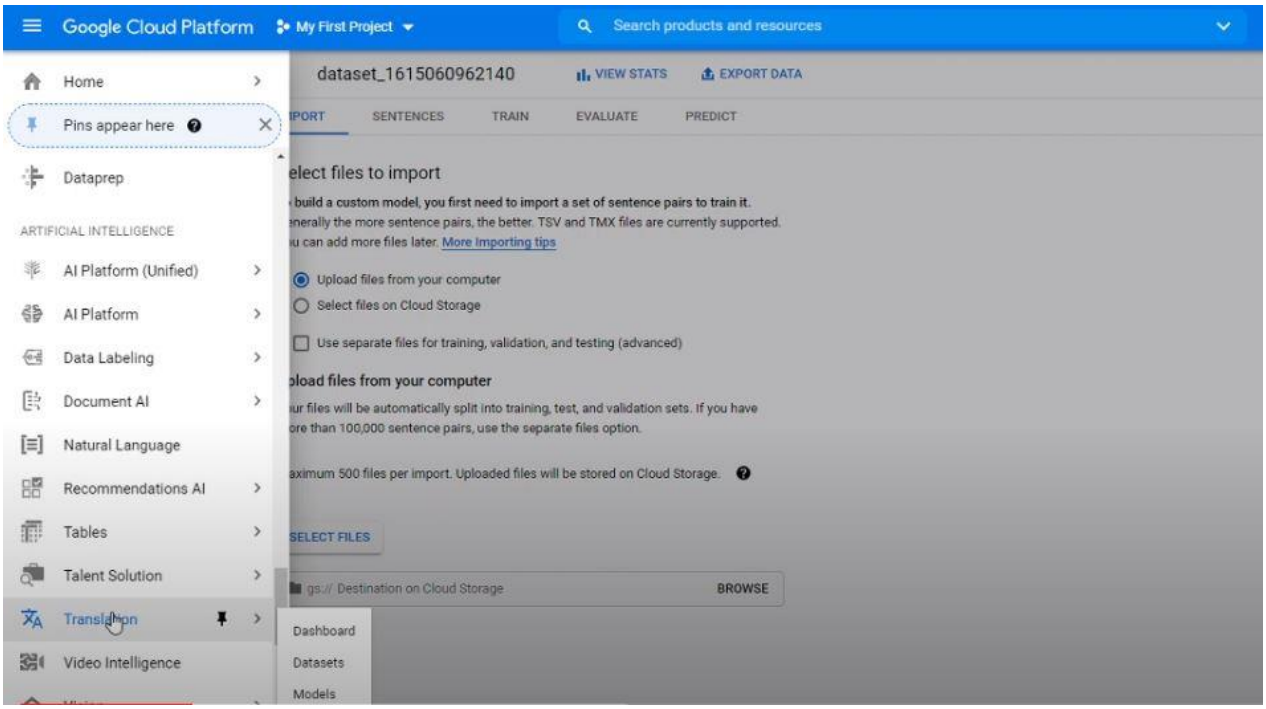
***Slika 13 Excel u radu s tekstom na poravnavanju prevoditeljske memorije***

## 4.3. Opis postupka treniranja modela

### 4.3.1. Postavljanje modela za treniranje

Postupak treniranja modela koristeći AutoML Translate značajku Google Cloud Platforme jednostavan je, poprilično brz kad se uzme u obzir količina podataka koja se elaborira i efikasan. Postupak prijave iziskuje unošenje kartičnih podataka, a nakon prijave u Google Console dobiva se 300 američkih dolara u košarici besplatnog probnog roka u kojem korisnik može odlučiti hoće li nastaviti koristiti platformu o svom trošku ili ne.

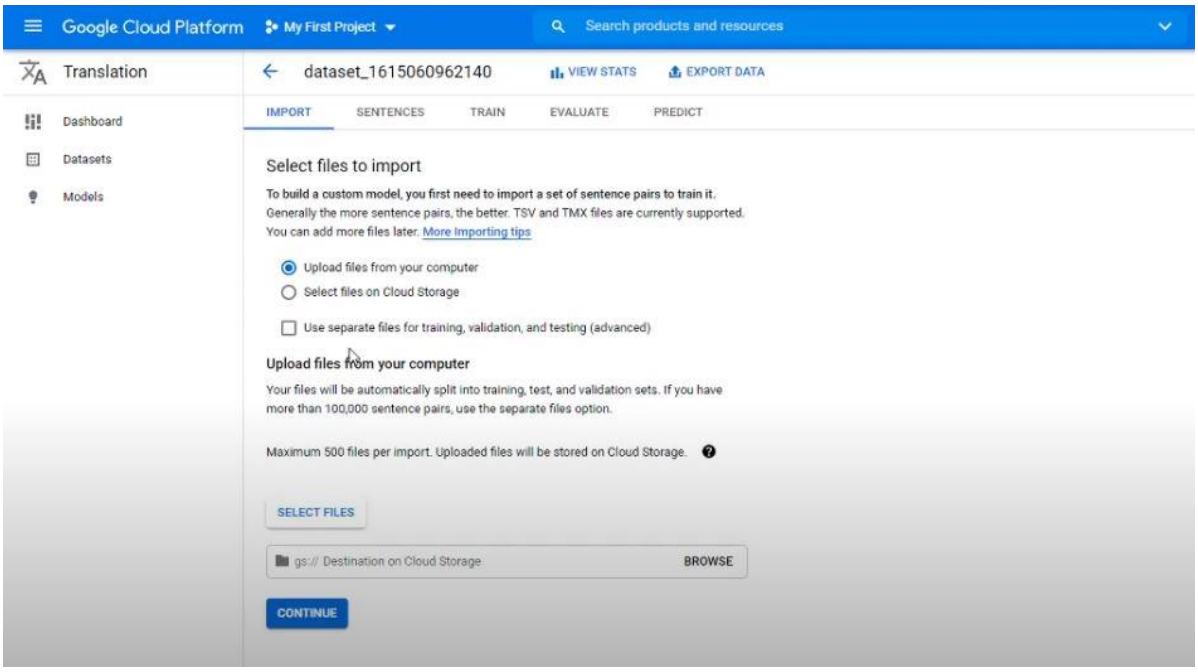
Google Console omogućava različite načine primjene alata s umjetnom inteligencijom uključujući obradu prirodnog jezika, računalni vid i slično. U ovom slučaju, kako je prikazano na Slici 26, u listi smo odabrali Translate opciju za izradu modela za automatsko strojno prevođenje bazirano na našoj bazi podataka tekstova te nastavili klikom na Models.




***Slika 14 Google Cloud Platforma - Odabir opcije izrade modela za strojno prevođenje u navigacijskoj traci***

Nakon odabranoga, ne može se nastaviti korištenje alata dok ne odaberemo opciju Enable API te se tada omogućuje prilaganje baze podataka koja će bit temelj za elaboraciju i trening budućeg modela.

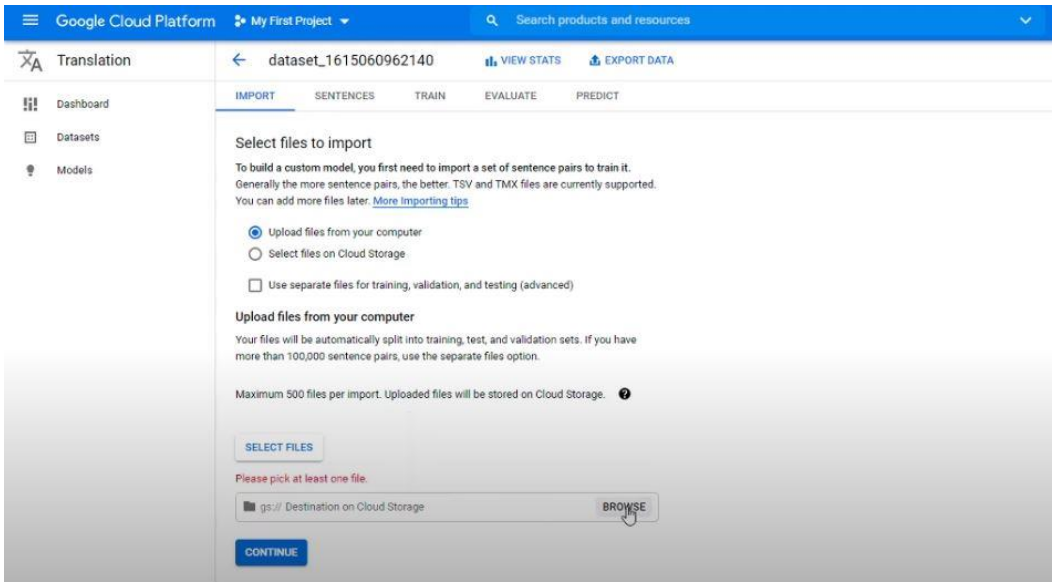
### 4.3.2. Import – prijenos datoteka baze podataka prevedenih jezičnih struktura




***Slika 15 Import - prilaganje baze podataka kao temelja budućeg modela za strojno prevođenje***

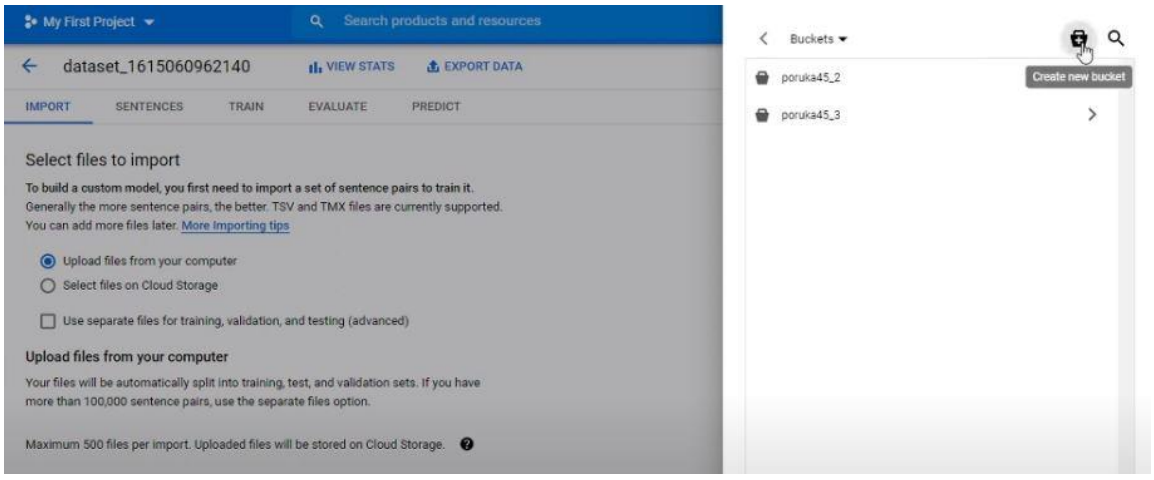
Google Cloud platforma kod prilaganja baze podataka prevedenih tekstova koja će služiti kao temelj budućeg modela nudi opciju prilaganja s uređaja na kojem se koristimo platformom ili prinošenja podataka s Cloud Storage internetskog prostora (Slika 27). U ovom slučaju bazu smo prethodno izradili korištenjem ranije opisanih alata te smo je prilagali iz kompjuterske memorije. Ono što je bitno napomenuti je da Google platforma može raditi na podacima spremljenim kao .tsv datoteke, .tmx datoteke i .csv datoteke, međutim u praksi dolazi do opterećenja stranice kod korištenja .tmx datoteka, a najmanje problema se u praksi pokazalo kod korištenja .tsv datoteka. Također, maksimalno se može predati 500 datoteka na obradu, a one mogu sadržavati prevedene tekstove rastavljene na riječi, fraze, rečenice i odlomke. U nastavku će biti obrazložena statistika različitih modela, od kojih je nekolicina trenirana na više od 64 700 rečeničnih parova.

Kod prenošenja datoteka s računala, potrebno je odrediti odredište na Google Cloud Storage internetskom mjestu pohrane podataka (Slika 28).



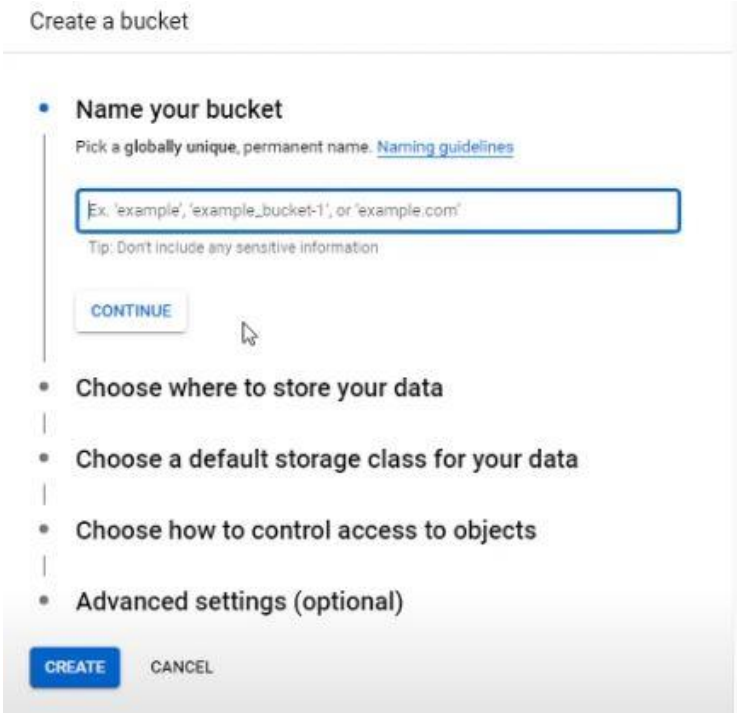

***Slika 16 Import - odabir destinacije na Cloud Storage mjestu pohrane podataka***

Zanimljivo je primijetiti da platforma zahtjeva različita mjesta pohrane za različite jezične parove. Budući da u ovom slučaju govorimo o englesko – hrvatskom jezičnom paru, stvaranje bucketa tj. mape koja je pohranjena na cloud-u koja ima određene specifikacije spomenute u nastavku. Stvaranje novog bucketa prikazano je na Slici 29.



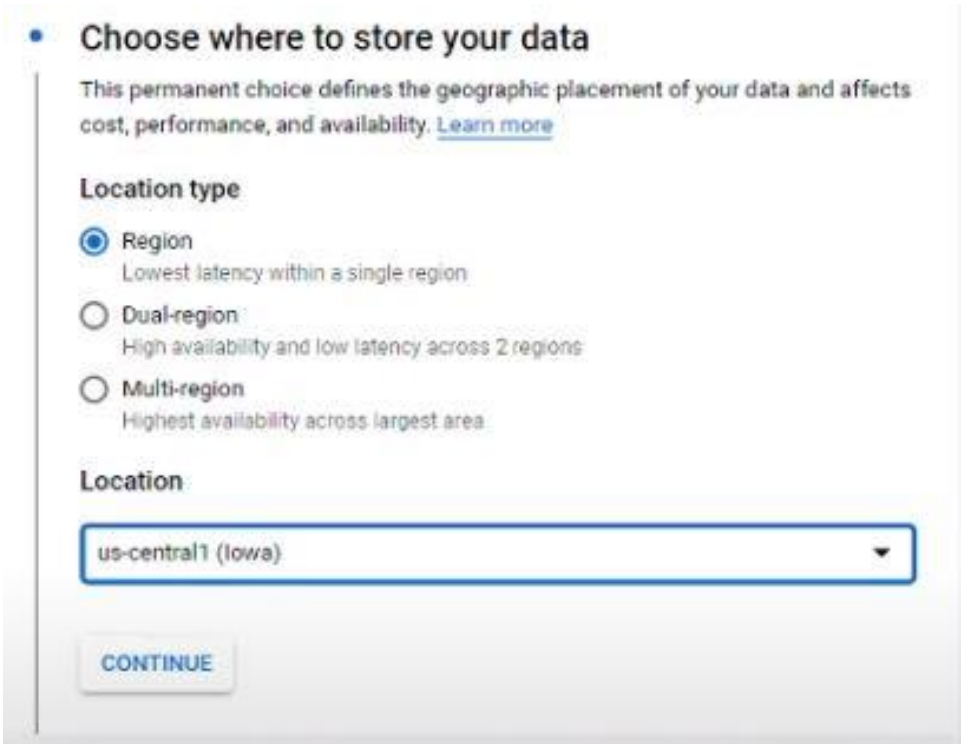

***Slika 17 Import - stvaranje novog bucketa na Cloud Storage-u***

Kod stvaranja bucketa takozvane košarice za pohranu baze podataka prevedenih jezičnih struktura, potrebno je odabrati karakterističnо ime koje do sad nije zauzeto u bazi, odabrati gdje će se spomenuta baza pohraniti (Slika 30), odabrati standardnu klasu za pohranu za našu bazu, odabrati kako kontrolirati pristup objektu, a postoje još i dodatne postavke koje nije neophodno mijenjati kako bi se pohrana podataka neometano izvršila.




***Slika 18 Stvaranje bucketa: opcije za odabir pohrane baze podataka na Google Cloud Storage-u***

Za englesko – hrvatski jezični par Google preporuke su da se odabere Location type Region (Slika 31), a da lokacija pohrane bude us-central1 (lowa). U praksi se pokazalo da spremanje na neku drugu od ponuđenih lokacija iz padajućeg izbornika ne dozvoljava nastavak rada u sustavu, što bi moglo biti podložno promjenama u skoroj budućnosti budući da se platforma neprestano mijenja.




***Slika 19 Cloud Storage: Gdje možemo pospremiti bazu podataka***

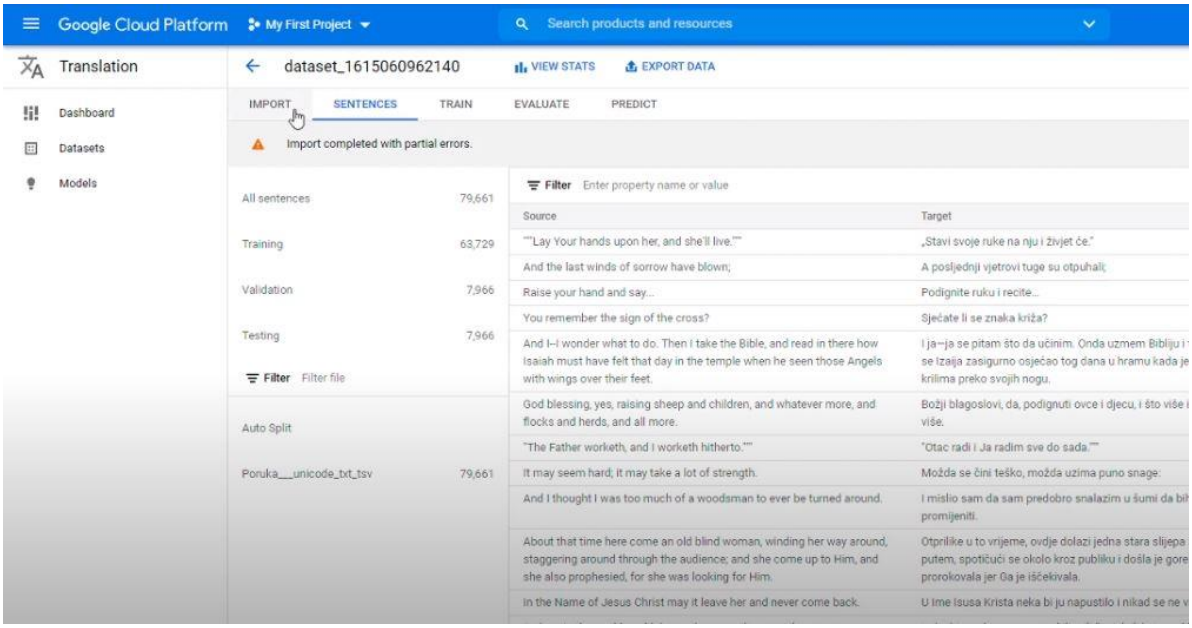
Sljedeći korak kod stvaranja bucketa je odabir standardne klase za pohranu podataka. Kod odabira klase (Slika 32) bitno je dobro odrediti prirodu projekta i budžet budući da odabiru ovisi koliko će dugo i pod kojim cijenama baza moći biti pohranjena. Ovaj korak je također bitan jer se njime ograničuje koliko će se često pristupati podacima pohranjenim na Cloud Storage-u. Standardna opcija prikladna za provedeni projekt odličan je izbor za kratkotrajno spremanje podataka (u našem slučaju je to bilo nekoliko mjeseci) i čest pristup tim podacima. *Nearline* opcija dobra je za pohranu paralelni korpus  kojima se pristupa rjeđe od jednom mjesečno. *Coldline* je opcija koja je najpriklanija za projekte u kojima će se podacima pristupati manje od jednom u kvartalu godine, dok je *archive* opcija prigodna za dugoročnu pohranu podataka kojima će se pristupati manje od jednom godišnje. Kod ovog koraka bitno je odabrati opciju koja će projektu najviše odgovarati, a u našem slučaju pristup podacima je čest, a rok pohrane nekoliko mjeseci zbog čega je standardna opcija najbolje ispunjavala svrhu.


***Slika 20 Standardna klasa pohrane za bazu podataka projekta***

Nakon odabira prethodno opisanih opcija mogli smo stvoriti bucket košaricu te krenuti s učitavanjem podataka kao pripremni materijal za treniranje modela. Trajanje učitavanja za količinu od 63 700 prevedenih rečenica u hrvatsko – engleskoj jezičnoj kombinaciji u našem slučaju trajalo je četiri sata.

### 4.3.3. Sentences – baza prevedenih jezičnih struktura učitanih u AutoML sustav

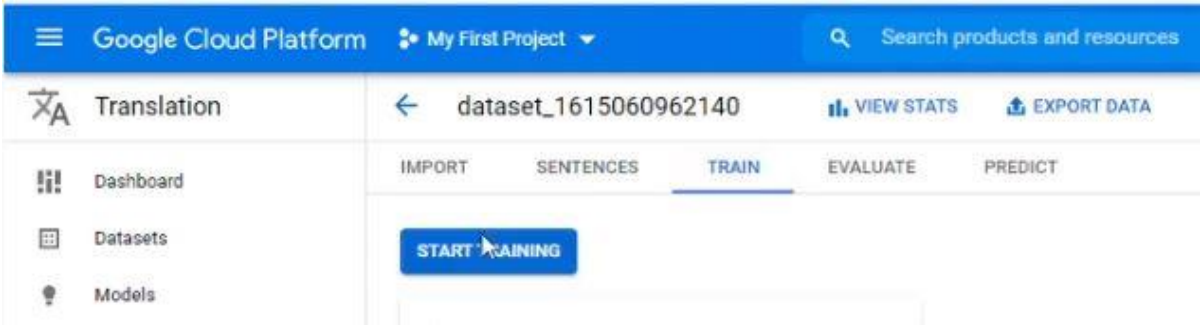
Nakon stvaranja košarice za pohranu baze podataka i učitavanja podataka u projekt, prevedene rečenice možemo vizualizirati u AutoML okruženju (Slika 33). Broj učitanih rečenica se automatski rastavlja na tri komponente: treniranje, validacija i testiranje na temelju čega se primjenom umjetne inteligencije može krenuti s treniranjem modela. Minimalan broj učitanih rečenica za moguće treniranje modela mora biti 10 000 rečenica od kojih platforma automatski bira određeni broj rečenica koje će biti obrađene u procesu validacije i testiranja modela.




***Slika 21 Sentences - učitana baza odataka iz prevoditeljske memorije spremna za treniranje, validaciju i testiranje modela***

### 4.3.4. Train – početak treniranja modela po potrebama korisnika

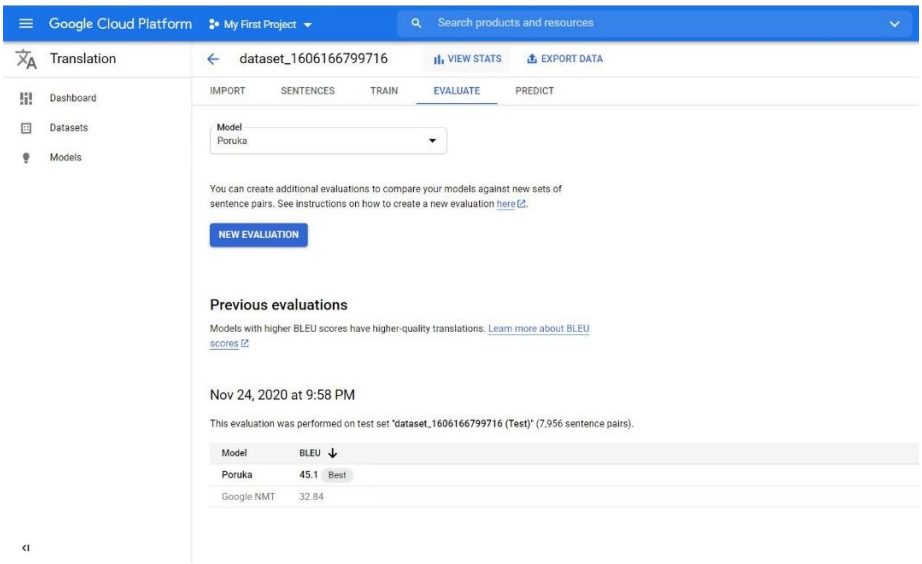
Nakon što su rečenice vidljive u sustavu, memorija podijeljena na komponente treniranja, validacije i testiranja, moguće je krenuti s treniranjem modela (Slika 34). Ono što platforma nudi je provedba testiranja tako da korisnik ne mora ulaziti u kodnu strukturu procesa, već jednim klikom započne treniranje modela za strojno prevođenje temeljenog na primjeni umjetne inteligencije koji koristi bazu podataka odabranu od strane korisnika, što znači da je terminologijom i sadržajem u potpunosti prilagođena potrebama korisnika. Bitno je napomenuti da proces treniranja može biti poprilično dugotrajan proces što je i očekivano kada se uzme u obzir potrebna količina podataka koja se obrađuje u procesu i provedbu svih faza strojnog učenja koje se u tom vremenu događaju. U našem slučaju treniranje se odvijalo u periodu od deset sati što je bilo dovoljno da istrenirani modeli postignu izrazito dobru točnost prijevoda. Ono što je iz perspektive prevoditelja od velike važnosti jest činjenica da nije potrebno poznavanje programskih jezika ili imati stručna informatička znanja kako bi se istrenirao model, već je bitno imati dobru bazu prevedenih tekstova i nekoliko sati strpljenja.




*Slika 22 Start training - treniranje modela za strojno prevođenje u jednom kliku*

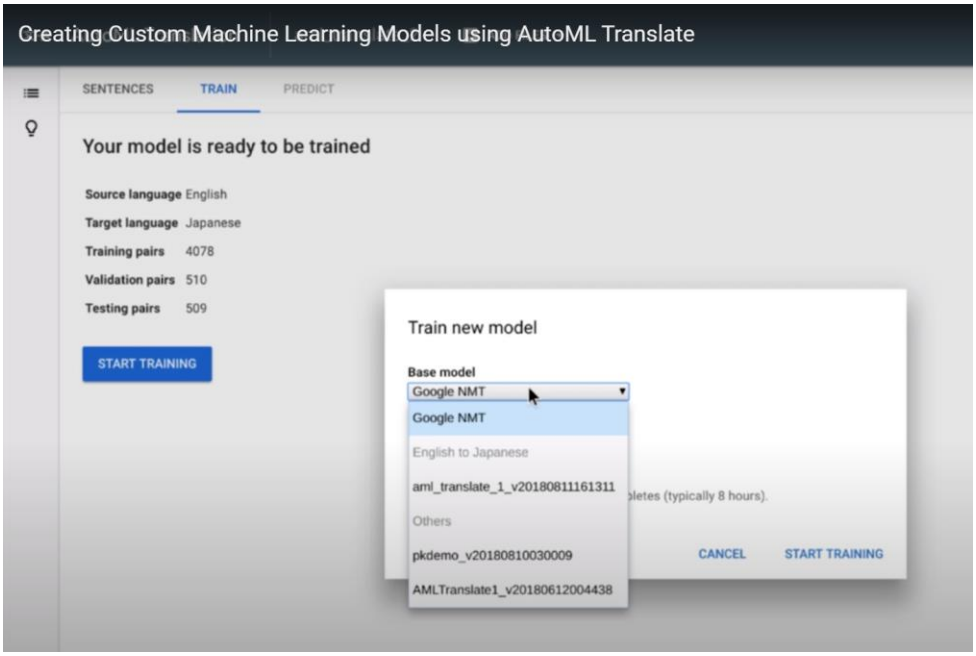
### 4.3.5. Evaluate – automatska evaluacija modela

Google Cloud platforma nudi krajnjem korisniku mogućnost da nakon automatske evaluacije modela ponovi isti proces tako da time dodatno provjeri prvobitno dobivene rezultate automatske evaluacije (Slika 35). U našem slučaju, ponovna evaluacija statističkih rezultata modela bila je identična prvobitno dobivenom rezultatu što je potvrdilo da se proces treniranja događa efikasno bez propuštanja faza elaboracije. Također, automatski možemo usporediti statističke rezultate treniranog modela sa rezultatima standardnog NMT Google modela za englesko – hrvatski jezični par. Ono što je u ovom slučaju iznenađujuće je to da je rezultat pokazao BLEU mjeru za 12,26 veću od one koju ima NMT Google model što se pokazalo izrazito dobrim rezultatom.




*Slika 23 Evaluate - ponuđena opcija ponovne evaluacije modela*

U ovom su se djelu projekta pokazale određene razlike za englesko – hrvatski jezični par o kojima će više riječi biti u poglavlju rezultata tj. statistike treniranih modela. Kada se pokuša objasniti način treniranja modela bitno je spomenuti da je na Slici 36 preuzetoj iz Google AutoML službenog tutorijala vidljivo da kod rada na englesko-japanskom jezičnom paru korisnik sam može odabrati temeljni model (eng. base model) na temelju kojeg će trenirati novi model s vlastitom prevoditeljskom memorijom. Kod englesko-hrvatske jezične kombinacije Google automatski određuje kvalitetu NMT temeljnog modela na koji će dograđivati treniranje specifičnog modela po mjeri korisnika. Također, u praksi se pokazalo da su modeli koji su imali bolji rezultat učinkovitosti (eng. performanse gain/loss) od početnog Google NMT modela u BLEU rezultatu imali lošiji krajnji rezultat od primjerice modela čiji je skok u BLEU rezultatu bio manji, ali im je temeljni model bio bolje početne kvalitete.




***Slika 24 Mogućnost odabira Base modela za englesko - japanski jezični par***

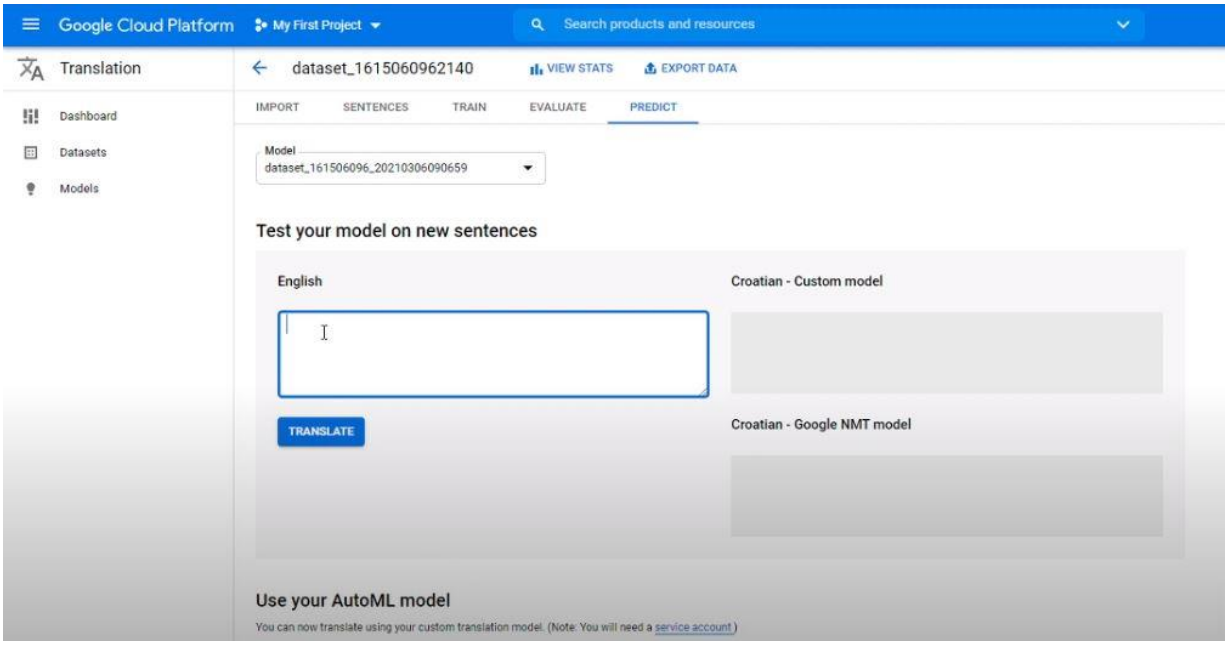
Kod indijskog jezika primjerice prevoditelj može odabrati koji će mu biti base model ili temeljni model na koji platforma dograđuje procesiranje novih podataka, dok za hrvatski jezikto još uvijek nije moguće. Slika 36 prikazuje mogućnost odabira temeljnog Base modela iz padajućeg izbornika za englesko – japansku jezičnu kombinaciju.

Ono što možemo primijetiti kod drugih jezičnih kombinacija, platforma omogućuje nadogradnju modela tako da za temeljni model korisnik može odabrati neki od modela koje je samostalno prethodno trenirao na vlastitoj bazi podataka. Nadamo se da će slične opcije biti uskoro omogućene za hrvatski jezik što će prevoditeljima uvelike poboljšati kvalitetu krajnjih modela i pomoći u radu sa korištenjem strojnog prevođenja u praksi.

### 4.3.6. Predict – testiranje modela na novim rečenicama

Nakon odrađenog cjelokupnog prethodno objašnjenog procesa, model je konačno spreman za korištenje. Predict opcija nudi korisniku mogućnost unosa novih tekstova za prevođenje te daje dva različita prijevoda (Slika 37). Custom model je onaj treniran na podacima korisnika i prethodno objašnjenom procesu, dok je Google NMT model dan za usporedbu kao osnova te je moguće odmah vidjeti razlike u prijevodu.

Osim toga što je navedena opcija korisna za procjenu rezultata rada modela, također je korisno za prevoditelje imati dvije različite opcije koje mogu odabrati u svojim daljnjim prijevodima. Daljnje korištenje modela može se odvijati preko Google AutoML platforme, a moguća je integracija modela u neke druge platforme.




*Slika 25 Predict - testiranje modela na novim rečenicama*

# 5. Rezultati

U ovom poglavlju bit će prikazani statistički podaci za pet različitih modela koji su bili trenirani na različitim varijacijama baze podataka sličnih sadržaja. Radilo se o bazama podataka prevedenih rečenica cjelokupne Biblije i tekstova Williama Branhama u formatima .tmx i .tsv. Statistički podaci prikazuju BLEU rezultat odnosno metriku, koja se dobiva na temelju algoritma za procjenu kvalitete teksta koji je strojno preveden s jednog prirodnog jezika na drugi.

Prema (Tonković, 2019.) „BLEU metrika - Bilingual Evaluation Understudy – je metrika koja nam daje rezultat usporedbe strojnog prijevoda teksta i jednog ili više referentnih tekstova. Ova metrika je danas jedna od najkorištenijih, a zamisao metrike je korištenje težinskog prosjeka uparenih fraza iz strojnog i referentnog prijevoda.“ Formula BLEU metrike je:

$$BLEU = kazna \times \exp \sum \lambda i \; n \; i{=}1 \log preciznosti,$$

a kazna se računa prema sljedećoj formuli:

$$e^{\,1-\, duljinaReferentnogPrijevoda/duljinaStrojnogPrijevoda.}$$

Prema (Tonković, 2019.), „metrika je brza i lako razumljiva, savršen rezultat podudaranja tekstova pomoću ove metrike iznosi 1.0, a potpuno nepodudaranje daje rezultat 0.0. BLEU metrika radi na način da se broje podudarajući n-grami u strojnom prijevodu s n-gramima u referentnom tekstu, gdje je 1-gram svaki „token“, odnosno svaka pojedina riječ, a 2-gram se odnosi na svaki par riječi, itd. Savršeni rezultat, tj. savršeno podudaranje u praksi nije posve moguće, budući da bi to značilo da je strojni prijevod posve identičan referentnom (najčešće ljudskom) prijevodu.“

Tablica 5 pokazuje objašnjenje metrike BLEU rezultata gdje je vidljivo da su modeli koji postignu rezultat manji od 10 skoro beskorisni za korištenje u praksi, modeli s rezultatima u rasponu od 10 do 19 daju prijevode koji su teški za razumjeti te također nisu od velike koristi prevoditeljima. Kod raspona od 20 do 29, dolazi do lakšeg razumijevanja prijevoda, ali su česte gramatičke pogreške, u rasponu od 30 do 40 tek se kreće govoriti o dobrim i razumljivim strojnim prijevodima.

Kod modela koji imaju rezultate u rasponu od 40 do 50 dobivamo prijevode visoke kvalitete, a upravo je u tom rasponu najveća postignuta kvaliteta modela u našem slučaju što će biti vidljivo u nastavku. Modeli koji postignu rezultate u rasponu od 50 do 60 BLEU metrike daju prijevode izrazito velike kvalitete i jezično fluentne, dok su svi modeli s rezultatima većima od 60, češće boljih prijevoda od onog što može proizvesti čovjek prema Tablici 5.

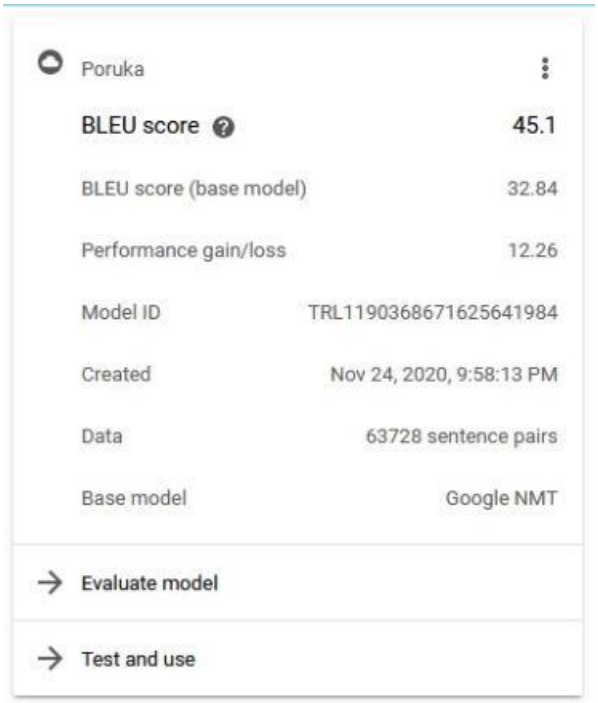
***Tablica 5 Interpretacija BLEU Score metrike*** [7]

| **BLEU rezultat** | **Tumačenje** |
|---|---|
| < 10 | Gotovo beskorisno |
| 10 - 19 | Teško je shvatiti suštinu |
| 20 - 29 | Suština je jasna, ali postoje značajne gramatičke pogreške |
| 30 - 40 | Razumljivi prijevodi |
| 40 - 50 | Visokokvalitetni prijevodi |
| 50 - 60 | Vrlo kvalitetni, adekvatni i tečni prijevodi |
| > 60 | Kvaliteta često bolja od ljudske |

[7] Google Cloud. 2021. *Evaluating models | AutoML Translation Documentation | Google Cloud*. [online] dostupno na: <https://cloud.google.com/translate/automl/docs/evaluate> [pristupila 25.5.2021.]

## 5.1. Model treniran na prijevodnoj memoriji tekstova Williama Branhama poravnanoj na razini rečenice

Prvi model čiji su statistički podaci prikazani na Slici 38, dosegao je BLEU rezultat od 45.1. što se pokazalo kao najbolji strojni prijevod do sada korišten na sličnim tekstovima u za hrvatsko – englesku jezičnu kombinaciju. Rezultat temeljnog base modela iznosio je 32.84 što znači da je novonastali model bolji za 12.26 prema BLEU rezultatu. Količina podataka na kojima je model bio treniran bilo je 63728 rečenica prevedenih tekstova Williama Branhama. U ovom modelu je prijevod Biblije bio izuzet, a temeljni model po automatizmu bio je Google NMT model.

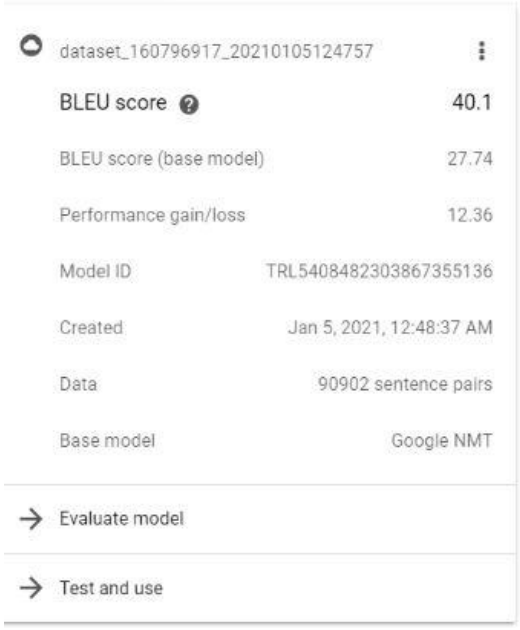



***Slika 26 Statistički podaci za prvi model: BLEU score 45.1***

## 5.2. Model treniran na prijevodnoj memoriji tekstova Williama Branhama i Biblije poravnanoj na razini rečenice

BLEU rezultat dobiven u drugoj provedbi treniranja za 5 je manji od rezultata prvog modela na što je utjecalo nekoliko različitih faktora (Slika 39). Drugi model treniran je na .tsv memoriji koja je imala isti sadržaj kao i tmx memorija u prvom modelu, što su tekstovi Williama Branhama, međutim u ovom slučaju prethodna tmx datoteka konvertirana je u tsv datoteku istog sadržaja. Također tsv datoteci prijevoda dodana je memorija prijevoda cjelokupne Biblije zbog čega je broj rečenica u ovom slučaju narastao na 90 902 prevedene rečenice u englesko hrvatskoj jezičnoj kombinaciji.

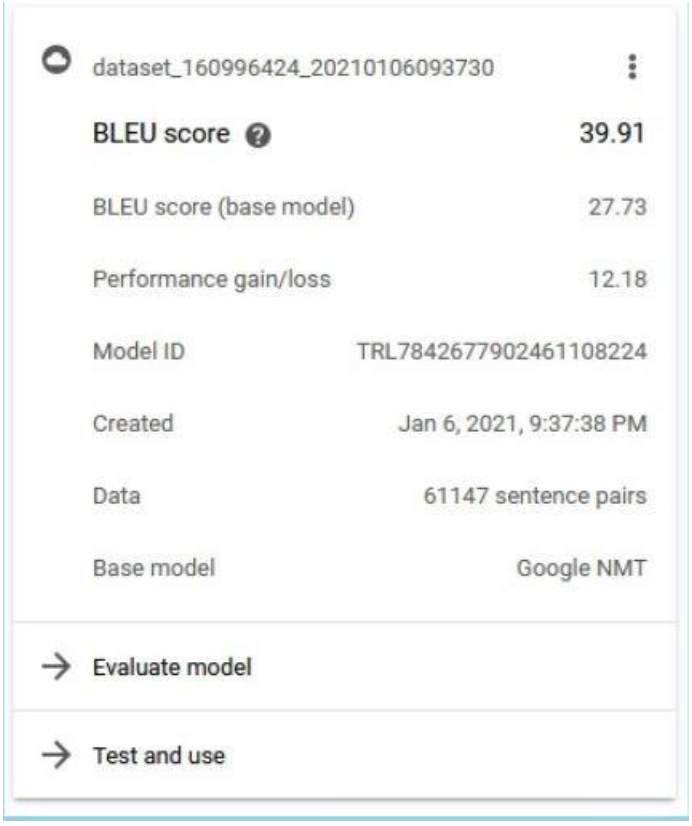
Jedna od promjena koja se pojavila kod treniranja u ovom slučaju bio je BLEU rezultat temeljnog modela na čiji odabir nismo mogli utjecati, već ga je Google automatizmom odredio za hrvatski jezik. Naime prema statističkom izvještaju za temeljni model odabran je Google NMT model, međutim BLEU rezultat je u ovom slučaju 27.74 što je za 5 manje od isto navedenog Google NMT modela u izvještaju modela objašnjenog u prethodnom poglavlju. Razlika između temeljnog modela i novonastalog je slična kao u prethodnom slučaju što je u ovom slučaju 12.36, a konačan rezultat istreniranog modela dosegao je mjeru 40.1.




*Slika 27 Statistički rezultati drugog modela: BLEU rezultat 40.1*

## 5.3. Model treniran na prijevodnoj memoriji tekstova Williama Branhama i Biblije poravnanoj na razini odlomka

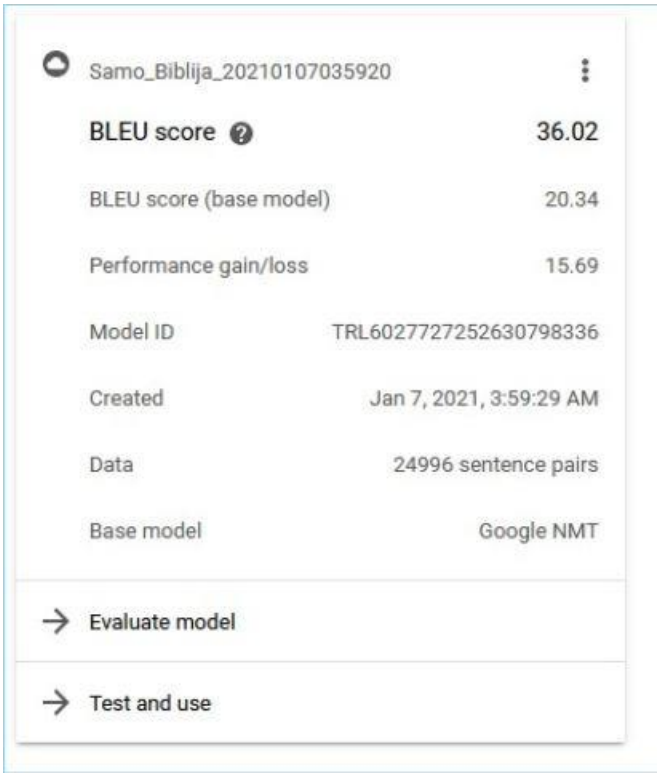
Analiza statističkih podataka trećeg modela (Slika 40), poprilično je slična prvome modelu po dobivenoj kvaliteti i rezultatu, međutim način treniranja se razlikuje u nekim detaljima. Naime, treći model treniran je na nešto manjoj količini podataka, točnije na 61147 rečeničnih parova koji su bili sadržani u tmx datoteci memorije prevedenih tekstova Williama Branhama koji su u prvobitnoj tmx datoteci bili raspoređeni u prevedene paragrafe, a proces pretvorbe u prevedene rečenice odradio se automatizmom na AutoML platformi. U paralelni korpus je također uključena i memorija prijevoda Biblije što je sve zajedno uključeno u 61147 rečeničnih parova. Base model je Google-ov NMT model isti kao u modelu u prethodnom poglavlju s rezultatom 27.73, a razlika u dobitku je 12.18 u korist novonastalog modela.




*Slika 28 Statistički rezultat trećeg modela: BLEU rezultat 39.91*

## 5.4. Model treniran na prijevodnoj memoriji Biblije poravnanoj na razini stihova

Četvrti po redu model bio je treniran samo na .tsv datoteci prijevoda Biblije s 24 996 rečeničnih parova. Kod ovog modela (prikazano na Slici 41) možemo primijetiti najveći skok u kvaliteti s obzirom na temeljni base model odabran automatski, BLEU preformanski dobitak narastao je za čak 15.69, ali je nejasno zašto se u ovom slučaju BLEU rezultat temeljnog modela postavio na 20.34. U statističkom izvještaju možemo vidjeti da se radi o Google NMT modelu, međutim prema rezultatu vidimo da je model manje kvalitete od prethodnih. Uz sve navedeno četvrti model dosegao je rezultat od 36.02 što je i dalje bilo od pomoći prevoditeljima koji su koristili model u praksi, ali se vidjela velika razlika u usporedbi s prvim modelom rezultata 45.1 koji se nastavio sve više i više koristiti u praksi.

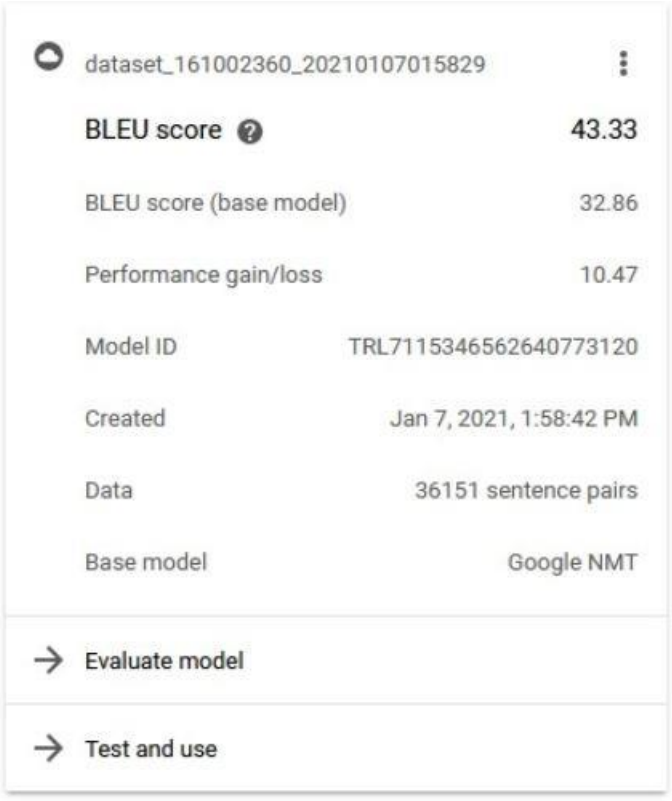



***Slika 29 Četvrti model strojnog prijevoda: rezultat BLEU 36.02***

## 5.5. Model treniran na prijevodnoj memoriji tekstova Williama Branhama poravnanoj na razini odlomka

Peti model u procesu istraživanja statistički je zaokružio logičku cjelinu istraživanja te je sadržajno sličan prvome, osim što je za njegovo treniranje iskorištena memorija prevedenih tekstova Williama Branhama učitana iz .tmx datoteke u kojoj su prevedene cjeline bile razdijeljene u odlomke, također je različita količina prenesenih podataka za treniranje, u ovom slučaju model je koristio 36151 rečenični par.

Kod modela kojeg analiziramo, čiji su statistički rezultati prikazani na Slici 42, možemo uočiti da je BLEU rezultat temeljnog Google NMT modela gotovo jednak kao kod prvog treniranog modela, a iznosi 32.86. Poboljšanje u rezultatu između temeljnog modela i dobivenog modela iznosi 10.47, čime je krajnji model ostvario BLEU rezultat od 43.33, što je samo za 1.77 manje od prvog modela u ovom istraživanju.




***Slika 30 Peti model strojnog prevođenja s rezultatom od BLEU 43.33***

## 5.6. Tablična usporedba BLEU rezultata treniranih modela na AutoML platformi

Tablica 6 prikazuje pregled BLEU rezultata svih modela treniranih tijekom projekta te potvrđuje pretpostavku s početka istraživanja da je najbolje rezultate dao sustav treniran na prevoditeljskoj memoriji u kojoj svaki rečenični segment na originalnom engleskom jeziku pored sebe ima odgovarajući prijevod na hrvatski jezik, a sadržaj je tekstova jednog autora i ne dolazi do kombiniranja sa prijevodom Biblije.

***Tablica 6 Usporedba BLEU rezultata treniranih modela na AutoML platformi***

| Poglavlje u kojem je opisan model | Broj jezičnih parova | BLEU rezultat temeljnog modela | Rezultat učinkovitosti | BLEU rezultat treniranog modela |
|---|---|---|---|---|
| 5.1. | 63728 | 32.84 | 12.26 | 45.1 |
| 5.2. | 90902 | 27.74 | 12.36 | 40.1 |
| 5.3. | 61147 | 27.73 | 12.18 | 39.91 |
| 5.4. | 24996 | 20.34 | 15.69 | 36.02 |
| 5.5. | 36151 | 32.86 | 10.47 | 43.33 |

## 6. Rasprava i zaključak

Kvalitetno organizirana i dobro poravnana prijevodna memorija ključna je kod stvaranja prijevoda visoke kvalitete bilo da se radilo o programsko potpomognutom prevođenju čovjeka ili o strojnom prijevodu. Kada govorimo o neuronskom strojnom prijevodu i treniranju modela koji se koriste umjetnom inteligencijom u izrađivanju što kvalitetnijih automatskih prijevoda, tada je priprema temeljnog materijala podataka za treniranje modela od ključne važnosti. U provedenom eksperimentu u kojem je korištena AutoML Translate platforma dostupna kao alat Google Cloud proizvoda, priprema dobre prevoditeljske memorije za treniranje modela pokazala se ključnom za dobivanje što boljih rezultata.

U pet različitih etapa provedenog eksperimenta dobili su se rezultati u rasponu od najnižeg BLEU rezultata od 36.2, do najvišeg rezultata od 45.1. Važno je spomenuti da su rezultati u rasponu od 30 do 40 riječima interpretirani kao modeli koji daju razumljive i dobre prijevode, dok se za sve rezultate u rasponu od 40 do 50 dobiveni prijevod definira kao prijevod visoke kvalitete. Glavni cilj istraživanja bilo je odrediti najbolji način pripreme i strukturiranja prevoditeljske memorije klijenta kako bi trenirani model bio što bolje kvalitete na AutoML platformi za englesko – hrvatsku jezičnu kombinaciju.

Početna hipoteza istraživanja je bila da će najbolje rezultate dati sustav treniran na poravnanoj prevoditeljskoj memoriji u kojoj svaki rečenični segment na originalnom engleskom jeziku pored sebe ima odgovarajući prijevod na hrvatskom jeziku, a sadržaj je tekstova jednog autora i ne dolazi do kombiniranja sa prijevodom Biblije. Hipoteza se nakon mjeseci rada uloženog u pripremu i strukturiranje prevoditeljske memorije pokazala točnom, što dokazuje da je priprema kvalitetnog korpusa važan čimbenik koji utječe na točnost modela.

Dok je za neke druge jezike poput indijskog, Google AutoML Translate omogućio da sam prevoditelj odabere temeljni model na kojem će korisnički podaci biti nadograđivani, na način da mogu za početni model koristiti neki od prethodno samostalno treniranih modela, za englesko hrvatsku jezičnu kombinaciju to još uvijek u lipnju 2021. godine nije moguće. Nadamo se da će brzi napredak neuronskog strojnog prevođenja pokazati rezultate i na Google AutoML Translate platformi te da će isto biti moguće primijeniti u radu s hrvatskim jezikom. Bez obzira na sve

navedeno, dobiveni modeli se trenutno koriste u prevoditeljskoj praksi kao najbolje rješenje u sferi strojnog prevođenja.

Od treniranih pet modela u projektu, najčešće korišten u praksi je prvi model koji je postigao također i najveći BLEU rezultat. U ovom slučaju se pokazalo da je ljudska evaluacija suglasna računalnoj analizi kvalitete dobivenog modela. U procesu softverski potpomognutog prevođenja kada govorimo o englesko – hrvatskoj jezičnoj kombinaciji možemo reći da je ovo rješenje najbolje koje za sada imamo kod prijevoda vjerskih tekstova, a prevoditeljska praksa je pokazala da je proces editiranja strojno dobivenog prijevoda uvelike ubrzao i olakšao rad prevođenja. Zbog svega navedenog u praksi se pokazalo da su tekstovi na čijim se prijevodima u 2019. radilo mjesec dana, u 2021. godini prevedeni u jednom tjednu prevoditeljskog angažmana što dokazuje da je Googlov opis da model daje prijevode visoke kvalitete uistinu točan.

## 7. Literatura


Cai, D. Y. W. H. L. W. L. a. L. L., 2021.. Neural machine translation with monolingual translation memory. U: s.l.:an.

Đunđer, I., 2015.. *Sustav za statističko strojno prevođenje i računalna adaptacija domene.* Zagreb: Doctoral dissertation, Filozofski fakultet u Zagrebu.

Europarlament, 2021. *Technology to support translation.* [Mrežno]
Available at: https://www.europarl.europa.eu/translation/en/translation-at-the-european-parliament/technology-to-support-translation
[Pokušaj pristupa 18 5 2021].

Gogić, S., 2020.. *Effects of New Technologies on the Translation Profession,* Osijek: Josip Juraj Strossmayer University of Osijek.

Hutchins, J., 2005.. *Towards a definition of example-based machine translation.* s.l., Asia-Pacific Association for Machine Translation.

Jiao, W. X. W. Z. T. S. S. M. R. L. a. I. K., 2021.. Self-Training Sampling with Monolingual Data Uncertainty for Neural Machine Translation. U: s.l.:an.

Koehn, P., 2010.. *Statistical Machine Translation.* Cambridge , Cambridge University Press.

Krizmanić, M., 2020.. *INSTITUCIJSKO PREVOĐENJE,* Rijeka: University of Rijeka.

Lewey, C., 2021. *Translation memory - a core feature of Trados translation software.* [Mrežno]
Available at: https://www.rws.com/translation/software/trados-studio/translation-memory/
[Pokušaj pristupa 22 5 2021].

Ljubaš, S., 2018.. Prijelaz sa statističkog na neuronski model: usporedba strojnih prijevoda sa švedskog na hrvatski jezik. *Hieronymus-časopis za istraživanja prevođenja i terminologije,* Svezak 4, pp. 72-91.

Ljubas, S., 2020.. *Utjecaj višejezičnosti vrednovatelja na ljudsku procjenu kvalitete strojnih prijevoda,* Zadar: Sveučilište u Zadru.

Microsoft, 2021. *Strojni prijevod.* [Mrežno]
Available at: https://www.microsoft.com/hr-hr/translator/business/machine-translation/
[Pokušaj pristupa 20 5 2021].

Mikulić, A., 2020. *Ljudska evaluacija sustava za neuralno strojno prevođenje,* Zagreb: University of Zagreb.

Perron, L. & Furnon, V., 2021.. *Cloud Translation | Google Cloud.* [Mrežno]
Available at: https://cloud.google.com/translate
[Pokušaj pristupa 6 4 2021.].

Reinke, U., 2013.. State of the Art in Translation Memory Technology. *Translation: Computation, Corpora, Cognition,* 3(1), pp. 27-48.

Saratlija, J., 2020.. *Izgradnja sustava za neuralno strojno prevođenje.* Zagreb: University of Zagreb.

Seljan, S. & Pavuna, D., 2006.. Why Machine-Assisted Translation (MAT) Tools for Croatian?. *Proceedings of 28th International Information Technology Interfaces Conference – ITI,* pp. 469-475.

Skadiņš, R. & al, e., 2014. Application of Machine Translation in Localization into low-resourced languages. *Proceedings of the 17th Annual Conference of the European Association for Machine Translation,* pp. 209-216.

Smolej, V., 2021. *Chapter 1. Uvod u program OmegaT.* [Mrežno]
Available at: https://omegat.sourceforge.io/manual-standard/hr/chapter.about.OmegaT.html
[Pokušaj pristupa 22 5 2021].

Šuman, S., 2021. *PREGLED METODA OBRADE PRIRODNIH JEZIKA I STROJNOG PREVOĐENJA.* Rijeka: Zbornik Veleučilišta.

Tonković, K., 2019.. Evaluacija strojnih prijevoda s njemačkoga na. U: U. o. R. D. o. I. Doctoral dissertation, ur. *Evaluacija strojnih prijevoda s njemačkoga na hrvatski jezik.* Rijeka: an., p. 8.

Višić, I., 2020.. *Kontrastivna analiza ljudskoga i strojnoga prevođenja,* Osijek: Josip Juraj Strossmayer University of Osijek.

Vrtarić, I., 2021. *Predgovor drugom izdanju, KRŠĆANSKA CRKVA - BOŽJA RIJEČ.* [Mrežno] Available at: https://www.bozjarijec.com/biblija/stari-zavjet/item/230-00-predgovor-drugom-izdanju.html
[Pokušaj pristupa 23 5 2021].

Yonghui Wu et al., 2016.. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation,* s.l.: Google.

Zhang, J. & Zong, C., 2020.. *Neural Machine Translation: Challenges, Progress and Future.,* s.l.: an.

## 8. Popis slika

# 9. Popis tablica