Metodologija izrade višejezičnog paralelnog korpusa na osnovu onlajn digitalnih uputstava za upotrebu: korpus Hilti uputstava

Naučni rad DOI: 10.18485/judig.2025.1.ch29

Nikola Janković[1], *0000-0003-3484-4220*

Apstrakt

U radu je predstavljena metodologija izrade opsežnog višejezičnog paralelnog korpusa uputstava za upotrebu na osnovu digitalnih uputstava za upotrebu uređaja kompanije Hilti. Cilj ovog rada je doprinos boljim resursima za istraživanja vezanim za mašinsko prevođenje, teoriju prevođenja, ali i za nastavu prevođenja i druge vrste lingvističkih istraživanja. Polazna tačka metodologije bilo je korišćenje uniformne strukture HTML uputstava za svaki jezik na zvaničnoj Hilti onlajn bazi uputstava. Dalji koraci su obuhvatali identifikovanje HTML elemenata koji sadrže relevantne delove teksta, njihovu doslednu numeraciju za svaki jezik, identifikovanje i čišćenje pogrešno paralelizovanih tekstova i čuvanje obrađenih tekstova u formatu TMX, tako da je engleski izvorni jezik u svakoj datoteci. U proseku, za svako uputstvo postoji oko 20 ciljnih jezika, a sam korpus sadrži oko 7 miliona reči. Prednosti kreiranog korpusa uključuju visoku preciznost kada je u pitanju podudaranje segmenata između jezika, značajan broj reči, kao i veliki broj jezika obuhvaćenih korpusom. U radu će biti detaljno opisani koraci preuzimanja, obrade i strukturiranja podataka, tehnički izazovi sa kojima smo se suočili i kako su rešeni, kao i osnovni statistički podaci vezani za jezike u korpusu. Smatramo da se ovakav tip korpusa može koristiti za unapređivanje resursa za prevođenje, teoriju prevođenja, nastavu prevođenja, kao i za različite vrste analiza stručnog jezika.

Ključne reči: paralelni korpusi, mašinsko prevođenje, jezički alati, stručni jezik, TMX

[1] Filološki fakultet, Univerzitet u Beogradu, nikolajankovickv@gmail.com

1. Uvod

Pod terminom „paralelni korpusi“ najčešće se podrazumevaju korpusi koji sadrže tekstove na jednom prirodnom jeziku (takozvanom „izvornom jeziku“), uparene sa njihovim prevodom na drugom, takozvanom „ciljnom“ jeziku, iako ovaj termin može obuhvatati i korpuse tekstova na različitim jezicima koji su uporedivi na osnovu određenih kriterijuma poput oblasti znanja (Lefer, 2020:257-258). Osim toga, moguće je da paralelni korpus sadrži dva različita prevoda na isti jezik i u tom slučaju se izvorni i ciljni jezik poklapaju. Posle njihove pojave tokom 1990-ih godina prošlog veka, paralelni korpusi su postali vredan resurs u oblastima kontrastivne lingvistike, bilingvalne leksikografije, nastavi stranih jezika, nastavi prevođenja, ekstrakciji termina, mašinskom prevođenju i računarski potpomognutom prevođenju (engl. computer-aided translation), učenju parafraza, projekciji jezičkih alata na druge jezike, kao i mnogim drugim primenama u obradi prirodnih jezika (Bañón et al., 2020:4555; Lefer, 2020:258).

Međutim, za izradu ove vrste korpusa vezuju se specifične prepreke. Kako navode Lefer i saradnici (Lefer, 2020:268), prvi problem vezan za izradu paralelnih korpusa je problem dostupnosti tekstova. Naime, za određene jezičke parove postoji veoma mali broj paralelnih tekstova, pri čemu tipovi teksta koji se često prevode mogu biti teško dostupni zbog faktora poverljivosti (ibid.). Takođe, kako Varga i saradnici (Varga et al., 2007:247) navode, 40% svetske populacije su maternji govornici jezika sa 100 miliona govornika ili više, dok 4% svetske populacije su maternji govornici jezika sa pola miliona govornika ili manje; najveći broj ljudi na svetu govori takozvane „jezike srednje gustine“ (engl. *medium density languages*), čiji broj maternjih govornika je između ova dva broja. Međutim, kako isti autori ističu, broj paralelnih strana na Internetu koji uključuje jezike srednje gustine je iznenađujuće mali (ibid:247). Sa druge strane, sastavljanje paralelnih korpusa iz tekstualnih izvora koji nisu nastali u elektronskom (engl. *digital-born*) formatu podrazumeva optičko prepoznavanje teksta (engl. *optical character recognition*, skr. OCR) koje značajno otežava proces obrade tekstova zbog neizbežnih grešaka koje uvodi (Lopresti, 2009), dok paralelni dokumenti u elektronskom formatu u specijalizovanim domenima, kao što je pomenuto, često nisu dostupni zbog pitanja poverljivosti. Takođe, jedan od najzahtevnijih koraka u kreiranju paralelnih korpusa predstavlja segmentacija tekstova i njihovo uparivanje (na nivou paragrafa, rečenice ili reči), što najčešće zahteva korak ručne provere. Iz navedenih razloga smatramo da sajtovi sa veb-stranama na različitim jezicima, koje imaju identičnu HTML strukturu, predstavljaju

veoma vredan resurs za efikasnu izradu kvalitetnih paralelnih korpusa. U ovom radu predstavljen je Hilti-UM korpus (prema engleskoj sintagmi *Hilti User Manuals*), napravljen na osnovu onlajn uputstava za upotrebu za uređaje kompanije Hilti, a čija metodologija izrade može biti korisna osnova za izradu sličnih korpusa, naročito u oblasti stručnih jezika.

2. Proces izrade Hilti-UM korpusa

U fazi istraživanja izvora za građu paralelnih korpusa stručnog jezika koji bi uključivali srpski jezik, početna hipoteza je bila da pogodan format predstavljaju višejezična uputstva za upotrebu u mašinski čitljivom formatu, konkretno PDF. Međutim, iz razloga navedenih u odeljku 5, veb-strane dostupne na više jezika koje imaju identičnu HTML strukturu su se pokazale kao znatno jednostavniji format za obradu, a sajt kompanije Hilti kao najbolji kandidat za izvor građe takve vrste.

Prvi korak u procesu sakupljanja građe predstavljao je pretragu dokumenata na sajtu kompanije Hilti[2] iz kategorija "uputstvo za rukovanje" i "uputstvo za upotrebu", uz filter "srpski jezik".

U narednom koraku je sastavljena lista uputstava dostupnih u formatu HTML pomoću *Python* skripte i modula *Selenium* za automatizaciju pretraživača, pomoću kog su preuzeti i sačuvani u formatu JSON (*lista_uputstava.json*) linkovi za svaki model proizvoda, nazivi modela, kao i jezici za koje je uputstvo dostupno. Na osnovu sačuvanih linkova, za određeni model proizvoda su automatski preuzeta HTML uputstva na svim dostupnim jezicima pomoću Python skripte (*preuzimanje.py*) koja generiše odgovarajuće linkove za preuzimanje. Python skripta koristi prethodno napravljen *Python* rečnik kojim se uparuju nazivi jezika na sajtu sa njihovim jezičkim kodom (npr. "nemački": "de", "norveški": "no", "finski":fi, "ruski":"ru") kako bi dodavanjem odgovarajućih skraćenica za jezik automatski bili generisani linkovi ka uputstvima na svim dostupnim jezicima. Za svaki model proizvoda je na njegov osnovni URL (na primer, https://www.hilti.rs/content/hilti/E4/RS/sr/op-man.html/2201898) dodat svaki odgovarajući jezički kod (npr. "/en", "/sr", "bg"), a sadržaj ovih strana je sačuvan u zasebnom katalogu imenovanom kao i sam model. Obrazac imenovanja datoteka bio je *{naziv modela}_{jezički kod}.htm*, pri čemu su eventualne kose crte (/ i \) zamenjene crticama (-) kako bi nazivi datoteka bili validni. U ovoj fazi pronađeno je 120 modela koji poseduju HTML verziju uputstava, uspešno su preuzete HTML datoteke za 112 modela od

[2] https://www.hilti.rs/downloads.html

kojih je 108 imalo uputstvo na engleskom jeziku, odabranom za izvorni jezik.

Pomoću druge Python skripte (*paralelizacija.py*), koristeći modul *Beautiful Soup*, obrađena je sadržina svih HTML datoteka. Python skripta iz odgovarajućeg odeljka (HTML elementa čija klasa ima vrednost *o-editorial-section*, a nalazi se u elementu *main*) ekstrahuje sve HTML elemente koji pripadaju sledećim tipovima: *p, h1, h2, h3*, i *li*. Tekstualni sadržaj ekstrahovanih elemenata (koji odgovaraju pasusima) u izvornoj i ciljnoj datoteci potom se zapisuje u *Python* liste. S obzirom na to da je struktura HTML datoteka identična za sve jezike u okviru jednog modela, elementi tako dobijenih Python lista se zatim mogu jednostavno upariti,.

Iz prethodno sačuvane JSON datoteke (*lista_uputstava.json*) sa osnovnim podacima vezanim za uputstva, prema prethodno utvrđenom šablonu automatski je generisan URL za englesku verziju HTML uputstva (koja je uzeta kao izvorni tekst), i za verziju na drugom jeziku (koja je uzeta kao ciljni tekst).

U finalnoj fazi paralelizacije generiše se TMX datoteka u kojoj su upareni pasusi iz engleskog i odgovarajućeg ciljnog jezika. Jedan par pasusa se nalazi u elementu *tu* (engl. translation unit = jedinica prevođenja), koji sadrži po jedan element *tuv* za izvorni jezik (engleski, sa atributom *xml:lang="en"*) i ciljni jezik sa odgovarajućom vrednošću atributa *xml:lang*. Ovaj par elemenata *tuv* sadrži upareni tekst originalnih pasusa iz izvornog i ciljnog jezika. TXM datoteke se čuvaju u potkatalogu *txm* odgovarajućeg kataloga modela.

3. Validacija paralelizacije

Po obavljenoj paralelizaciji pasusa, odnosno generisanju TMX datoteka, bilo je potrebno izvršiti validaciju njenog kvaliteta za svaki model i svaki jezički par. S obzirom na veliki broj modela, jezičkih parova i pasusa u svakom dokumentu, ručna provera paralelizacije izvršena je za sve srpsko-engleske parove, dok je za ostale jezičke parove validacija izvršena programskim putem u dve faze pomoću dve *Python* skripte (*validacija.py* i *uzastopni_loši_segmenti.py*).

Prva faza programske validacije podrazumeva validaciju uparenih pasusa (obuhvaćenih elementima *tu*) na osnovu nekoliko kriterijuma (koristeći fajl *validacija.py*). Validan paralelizovan par je morao da zadovolji sledeće kriterijume:

- Postoji element *tuv* i za izvorni i za ciljni jezik, i nijedan od njih nije prazan.

- Ukoliko postoje brojevi u izvornom elementu, identični brojevi moraju postojati i u ciljnom elementu.
- Ukoliko se u izvornom elementu nalazi neki od sledećih simbola: %, ©, ®, ™, $, €, £, ili ¥, isti simbol mora postojati i u ciljnom elementu.
- Ukoliko je tekst u izvornom elementu duži od 10 reči, razlika u dužini izvornog i ciljnog teksta ne sme biti veća od 50%. Ovaj odnos izračunat je pomoću sledećeg izraza: *min(len(source_text), len(target_text)) / max(len(source_text), len(target_text))*.

Ova provera obuhvatala je jezike koji koriste ćirilično ili latinično pismo. Nije bilo moguće automatski validirati hebrejski, japanski, jermenski, korejski, kineski i kazaški jezik, i samim tim je provera TMX datoteka, kojima je jedan od navedenih jezika bio ciljni, automatski preskočena i one su premeštene u zaseban katalog.

Rezultati ove analize sačuvani su u JSON datoteku (*rezultati_validacije.json*), uključujući podatke koje su datoteke (ne)obrađene, a za svaku obrađenu datoteku i informacije o elementima koji su prekršili neki od prethodno navedenih kriterijuma validnog paralelizovanog para. Informacije o ovakvim elementima obuhvataju identifikacioni broj elementa, kriterijum koji je prekršen, izvorni tekst i ciljni tekst i one pružaju neophodni kontekst za dalju doradu kriterijuma, kako bi se obuhvatili svi nepravilno paralelizovani tekstovi i smanjio broj lažno pozitivnih rezultata.

JSON datoteka sa rezultatima analize paralelizacije (*rezultati_validacije.json*), proistekla iz prethodne faze validacije paralelizacije, iskorišćena je u potonjoj fazi validacije . Na osnovu nekoliko ponovljenih testiranja različitih kriterijuma, najbolji rezultati prilikom filtriranja pogrešno paralelizovanih datoteka postignuti su primenom kriterijuma uzastopnih pogrešnih elemenata. Naime, pomoću *Python* skripte *uzastopni_loši_segmenti.py,* datoteke koje su sadržale 5 ili više uzastopnih pogrešnih elemenata su automatski označene kao pogrešno paralelizovane. Na osnovu ovog kriterijuma, kreirana je finalna JSON datoteka (*finalna_validacija.json*) sa listom nepravilno paralelizovanih datoteka, kao i listom datoteka čija je obrada automatski preskočena i koje su zatim premeštene u zasebne kataloge. Finalni validirani korpus obuhvatio je 54 različita modela sa ukupno 800 TMX datoteka.

4. Analiza korpusa Hilti-UM

Posle faze izrade TMX datoteka i njihove validacije, pomoću *Python* skripte (*analiza_korpusa.py*) urađeno je nekoliko jednostavnih vrsta analiza vezanih za jezik sadržan u samom korpusu.

Glavni katalog korpusa sadrži kataloge pojedinačnih modela mašina čija su uputstva paralelizovana. Obilaskom potkataloga glavnog kataloga korpusa pomoću odgovarajuće *Python* skripte, kreiran je indeks TMX datoteka, gde je za svaku TMX datoteku zabeležen ciljni jezik na osnovu atributa *xml:lang* elemenata *tuv* (engleski jezik je bio izvorni jezik u svim datotekama).

Po kreiranju indeksa TMX datoteka izvršena je analiza sadržaja samih datoteka na nivou tokena (posebno reči) i segmenata. Kada je u pitanju analiza na nivou reči, s obzirom na veliki broj jezika koji je zastupljen u korpusu, bilo je potrebno uzeti u obzir specifičnosti kodiranja različitih pisama i njihovih karaktera kako bi se poboljšala tačnost tokenizacije, odnosno segmentacije teksta na reči. Na osnovu smernica navedenih u aneksu Unikod konzorcijuma o segmentaciji teksta (Unicode Consortium, 2025) za segmentaciju na nivou reči je primenjen Unikod regularni izraz:

$$\p{L}[\p{L}\p{M}\p{Nd}\p{Pc}''] * \qquad (RI\ 1)$$

Regularni izraz (RI 1) pronalazi tekst koji odgovara sledećim pravilima:

- Počinje slovom (\p{L}) posle kojeg ne sledi nijedan ili sledi bar jedan (*) karakter koji može biti:
  - slovo (\p{L}),
  - znak interpunkcije za povezivanje (engl. *combining diacritics*) (\p{M}),
  - cifra[3] (\p{Nd}),
  - znak vezne interpunkcije (engl. *connector punctuation*) (\p{Pc}), poput crtice (-),
  - apostrof (’').

Pomoću *Python* biblioteke lxml (konkretno, klase *xml.etree.ElementTree* i funkcije *iterparse*), obrađeni su svi segmenti u

[3] S obzirom na tehničku prirodu tekstova u korpusu Hilti-UM, određeni broj termina i oznaka sadrži i slova i brojeve, zbog čega je ovaj skup karaktera dodat kao moguć sastavni deo unutar reči. Cifre na početku nisu uključene u reč, čime se izbegava da se mere, poput “8mm”, beleže kao reči.

TMX datotekama pojedinačno, beležeći pritom broj segmenata, broj reči, kao i same reči koje se pojavljuju za određeni jezik. S obzirom na to da je engleski jezik bio izvorni jezik u svim datotekama, ovi podaci su se za engleski jezik sakupljali na osnovu podataka za izvorni jezik samo jedne TMX datoteke za svaki od modela, te broj dokumenata za engleski jezik odgovara broju modela čija su uputstva paralelizovana. Analiza je pokazala da validiran korpus obuhvata 3.439.354 tokena u okviru 800 datoteka, u kojima se pojavljuje 26 jezika.

Za svaki jezik, izračunat je broj dokumenata u kojima je jezik korišćen, broj segmenata, broj tokena, broj jedinstvenih reči i prosečan broj reči po segmentu (Tabele 1.1 i 1.2).

S obzirom na to da je u fazi pretrage uputstava na sajtu firme Hilti primenjen filter kojim se prikazuju samo uputstva koja postoje i na engleskom i na srpskom jeziku, ova dva jezika su zastupljena u najvećem broju modela (54 za engleski i 43 za srpski jezik), pri čemu je razlika u broju modela (54 – 43) posledica toga što 9 TMX datoteka na srpskom jeziku nije prošlo fazu validacije paralelizacije. Izuzimajući jezike isključene iz ove verzije korpusa zbog nemogućnosti validacije paralelizacije, kao i jezike za koje je validacija paralelizacije bila (gotovo) uvek neuspešna (zbog drugačijeg rasporeda sadržaja dokumenta), jezici sa najmanje pojavljivanja u modelima su danski, estonski i norveški (18 TMX datoteka). Prosečan broj dokumenata u kojima se koristi određeni jezik je 32,8. Kada je u pitanju broj tokena po jezicima, najveći broj tokena je očekivano prisutan u engleskom jeziku, dok je broj tokena za srpski potkorpus manji od onog za španski, francuski, italijanski, portugalski, rumunski i grčki uprkos većem broju dokumenata sa srpskim jezikom.

Tabela 1.1: Broj tokena po jezicima

Jezik	Broj tokena
en	236006
es	210535
fr	191069
it	190065
pt	186244
ro	185904
el	182446
sr	175374

Tabela 1.2: Broj dokumenata po jezicima

Jezik	Broj dokumenata
en	54
sr	43
es	42
hr	42
ro	41
it	40
nl	40
pt	40

nl	163977		bg	39
bg	162414		el	39
hr	161062		fr	39
sl	151499		sk	37
pl	137287		sl	37
sk	135374		cs	36
cs	132637		pl	35
uk	121691		de	31
de	117787		hu	31
hu	116690		uk	28
tr	93104		tr	26
sv	64986		lt	21
lt	64923		lv	20
lv	62343		sv	20
no	59817		fi	19
da	59404		da	18
fi	51414		et	18
et	50134		no	18
prosečno	**133237.9**		**prosečno**	**32.8**

Koristan uvid u leksiku u specijalizovanom domenu, kao što su uputstva za upotrebu različitih vrsta aparata, predstavljaju frekvencijske liste. Kako bismo izdvojili najčešće termine u engleskom i srpskom jeziku koji su česti u ovoj vrsti tekstova, ali ne i u opštem jeziku, koristili smo liste frekvencija generisane na osnovu velikih korpusa engleskog, odnosno srpskog jezika. Za engleski jezik korišćena je lista najfrekventnijih 333.333 reči iz Guglovog korpusa (en. *Google corpus*) (Norvig, 2009), dok je za srpski jezik korišćena lista frekvencija od preko 2 miliona srpskih reči, dobijena na osnovu srpskih potkorpusa velikih veb-korpusa i velikih elektronskih korpusa srpskog jezika[4], među kojima izdvajamo korpuse Kišobran i Znanje (Škorić & Janković, 2024). U nedostatku specijalizovane liste frekvencija učestalih reči koje imaju malu informativnu vrednost (engl. *stopwords*) za srpski jezik, koristili smo jednostavno filtriranje najčešćih 1241 reči iz navedenih lista frekvencija. Ovaj broj je odabran na

[4] https://huggingface.co/datasets/procesaur/Vikipedija, https://huggingface.co/datasets/procesaur/kisobran, https://huggingface.co/datasets/procesaur/znanje, https://huggingface.co/datasets/oscar-corpus/oscar

osnovu istraživanja u kome su Marovac i saradnici (Marovac et al., 2021) izdvojili 1241 najučestaliju reč male informativne vrednosti u srpskom jeziku. Pritom ne odbacujemo mogućnost da su neki validni termini isključeni iz liste zbog jednostavnosti naše metodologije. Tabele 2.1 i 2.2 ilustruju liste najčešćih termina za srpski i engleski jezik u korpusu Hilti-UM.

Tabela 2.1: Lista najčešćih termina u srpskom potkorpusu

	reč	frekvencija
1	baterije	1927
2	akumulatorske	1600
3	alata	1197
4	alat	1066
5	bateriju	1027
6	akumulatorsku	953
7	baterija	921
8	koristite	712
9	nemojte	677
10	hilti	656
11	proizvod	650
12	električni	629
13	električnog	589
14	back	552
15	akumulatorskih	519

Tabela 2.2: Lista najčešćih termina u engleskom potkorpusu

	reč	frekvencija
1	battery	3528
2	tool	3196
3	batteries	1836
4	instructions	1213
5	hilti	889
6	switch	841
7	operating	827
8	accessory	797
9	dust	749
10	clean	659
11	remove	629
12	injury	579
13	damage	552
14	operation	550
15	damaged	522

Korpus Hilti-UM će biti dodat na platformu „Bibliša“[5] kako bi bio dostupan u formatu pogodnom za jezičke analize i upotrebu u prevodilačkoj praksi.

5. Prepreke pri izradi korpusa Hilti-UM

U ovom odeljku predstavićemo glavne prepreke sa kojima smo se susreli u izradi korpusa Hilti-UM, kao i rešenja koja su primenjena za njihovo prevazilaženje.

Najpre, kada je u pitanju faza sakupljanja građe za paralelni korpus, uputstva za upotrebu različitih aparata se ubrajaju u lakše dostupne resurse stručnih paralelnih tekstova na različitim jezicima, uključujući srpski. Prva prepreka, sa kojom smo se susreli prilikom pretrage sajtova različitih proizvođača raznih vrsta proizvoda (mašina, aparata) na srpskom tržištu i

[5] http://biblisha.jerteh.rs/

analize njihove dostupnosti, je format uputstava. Naime, najveći broj pronađenih uputstava, dostupnih onlajn, nalazi se u formatu PDF koji nije mašinski čitljiv. Otuda je neophodan proces optičkog prepoznavanja karaktera (raščitavanja teksta) u uputstvima, što predstavlja veliki izazov u višejezičnim dokumentima ove vrste. Takođe, prisustvo slikovnih uputstava u PDF dokumentima uslovljava da redosled istih pasusa na različitim jezicima u određenim slučajevima bude različit, što za posledicu zahteva dugotrajan proces ručnog uparivanja pasusa. Imajući u vidu ova ograničenja, datoteke u formatu PDF nisu tretirane kao poželjni izvor građe za ovakav korpus.

U nastavku istraživanja sajtova proizvođača mašina, ustanovili smo da neki od njih nude uputstva za upotrebu u formatu HTML, kao posebne stranice svog sajta. Među ovim firmama, sajt kompanije Hilti je u ponudi imao najveći broj višejezičnih uputstava za različite modele u formatu HTML. Ručnim upoređivanjem HTML strukture uputstava na različitim jezicima i za različite modele, utvrdili smo da je struktura uputstava u najvećem broju slučajeva istovetna, što je u velikoj meri omogućilo automatsko sakupljanje građe opisano u odeljku 2.

U ovoj fazi, bilo je važno razmotriti pravne i etičke faktore sakupljanja građe na Internetu u svrhe izrade korpusa, naročito kada je u pitanju automatsko sakupljanje građe sa veb-sajtova (engl. *web scraping*). Naime, kako navode Krotov i saradnici (Krotov et al., 2020:558-559), osim poznavanja alata za preuzimanje i obradu podataka sa Interneta, istraživači moraju imati u vidu i načine upotrebe ovih podataka na zakonit i etičan način i preduzmu proaktivne mere da osiguraju da je njihovo sakupljanje podataka na Internetu u skladu sa pravnim i etičkim standardima. U ovom radu, autori su analizirali 147 naučnih radova vezanih za tehničke, etičke, i pravne aspekte preuzimanja podataka sa Interneta za naučne svrhe i predložili Etički i pravni okvir za automatsko preuzimanje podataka sa Interneta (en. *Legality and Ethics Framework for Web Scraping*) koji istraživači mogu pratiti pri donošenju odluka o (automatskom) sakupljanju podataka sa Interneta i koji obuhvata različite kriterijume u formi pitanja koja se tiču ilegalnog pristupa i korišćenja podataka, kršenja ugovora, prouzrokovanja štete i poslovnih tajni. Izrada korpusa Hilti-UM je vršena u skladu sa navedenim okvirom, zadovoljavajući pritom sve navedene kriterijume u ovom okviru. Faktori koji su omogućili da, prema našem istraživanju, ovi kriterijumi budu zadovoljeni su sledeći:

- Protokol *robots.txt* sajta *hilti.rs*, koji reguliše aktivnosti automatskog pretraživanja i preuzimanja na sajtu, dozvoljava pristup stranici za preuzimanje uputstava za upotrebu,

- Sama uputstva su javno dostupna i namenjena širokoj divulgaciji,
- Odeljak „Uslovi i politika“ sajta *hilti.rs*, koji se tiče prodaje i pružanja usluga Hilti proizvoda, ne pominje sadržaj na samom sajtu niti njegovo preuzimanje,
- Osim autorskih prava vezanih za prodate proizvode u dokumentu „Uslovi i politika“, kao i napomene vezane za Hilti zaštitni znak u uputstvima za upotrebu, u našem istraživanju samog sajta, dokumenta „Uslovi i politika“, kao i samih uputstava nismo našli eksplicitno naglašavanje autorskih prava o podacima na sajtu,
- Automatsko preuzimanje uputstava sa sajta Hilti je vršeno uz programirano pauziranje između zahteva kako se ne bi vršilo opterećenje servera,
- Proizvod ovog istraživanja je nekomercijalni paralelni korpus namenjen naučnom istraživanju, koji će biti prezentovan u vidu izdvojenih paralelnih segmenata, pomoću kojih se ne može rekonstruisati originalni tekst.

Kao što Krotov i saradnici (ibid.:567) naglašavaju da njihov rad nije namenjen da bude pravni savet, tako i autor ovog rada ističe da podaci navedeni u ovom radu ne mogu predstavljati pravni savet i da bi istraživači u slučaju pravnih nedoumica trebalo da se konsultuju sa pravnikom oko relevantnog zakonskog okvira i legalnosti konkretne vrste sakupljanja podataka.

U fazi paralelizacije, bilo je potrebno kreirati metodologiju serijalizacije HTML datoteka tako da njihov redosled bude identičan u svim jezicima. Inicijalna analiza datoteka pokazala je da se tekst nalazi u elementima *p, h1, h2, h3,* i *li*, sa izuzetkom oznaka segmenata poput „NAPOMENA“ ili „UPOZORENJE“, koje su se nalazile u elementima *div*, koje je bilo teže preuzeti s obzirom na to da većina elemenata *div* na sajtu ne sadrži tekst. Iz ovog razloga, odlučili smo da izuzmemo ove oznake segmenata iz preuzimanja.

S obzirom na to da su uputstva za isti model na različitim jezicima u najvećoj meri imala identičnu strukturu HTML elemenata, odabrali smo programsko generisanje TMX 1.4 datoteka pomoću *Python* skripti. Napominjemo da su u procesu izrade metodologije i pisanja kôda u našem istraživanju korišćeni modeli veštačke inteligencije (VI) kompanija OpenAI, Antropik i Gugl, posle čega je kôd ručno proveren, a hipoteze su istražene u relevantnoj literaturi. TMX datoteke, generisane pomoću modela VI, validirane su prema DTD šemi za TMX 1.4[6] koju je izradila nekadašnja Asocijacija za lokalizaciju industrijskih standarda (engl.

[6] http://www.ttt.org/oscarStandards/tmx/tmx14.dtd

Localization Industry Standard Association) (Savourel, 2002). Zaglavlje u prvobitnoj verziji TMX datoteka nije bilo u adekvatnom formatu, pa su sve datoteke ispravljene kako bi postale validne u odnosu na DTD šemu.

Jedna od najvećih prepreka u izradi korpusa Hilti-UM bila je validacija paralelizacije segmenata u TMX datotekama. S obzirom na veliki broj datoteka i na stepen poznavanja različitih jezika koji su sadržani u korpusu, ručna provera celog korpusa nije bila moguća. Iz ovog razloga, odlučili smo se za ručnu proveru celokupne srpsko-engleske paralelizacije i programsku validaciju ostalih jezičkih parova. Metodologija programske validacije je opisana u odeljku 3, a kao što je već pomenuto, jezici koji nisu zasnovani na ćiriličnom i latiničnom pismu nisu mogli da budu uključeni u ovu validaciju, zbog čega je odlučeno da se izdvoje iz korpusa. Međutim, smatramo da opisana metodologija može biti korisna istraživačima koji se bave izdvojenim jezicima da sami izrade (i validiraju) slične paralelne korpuse uputstava i na tim jezicima.

6. Zaključak

S obzirom na potrebe istraživača u različitim poljima lingvistike i nastave jezika za kvalitetnim paralelnim tekstovima, naročito kada je u pitanju jezik struke, smatramo da metodologija opisana u ovom radu može biti korisna za izradu sličnih paralelnih korpusa. U odnosu na tradicionalne izvore tekstova, smatramo da veb-sajtovi koji prikazuju javno dostupne stranice sa identičnom strukturom na različitim jezicima predstavljaju znatno lakši izvor za izradu paralelnih korpusa, s obzirom na to da se tekst već nalazi u mašinski čitljivom formatu i da je struktura već uparena na nivou paragrafa. Smatramo da bi ovakvi resursi bili veoma značajni u oblasti prevođenja, nastavi prevođenja, različitim vrstama analize jezika, ali i kao vredni skupovi podataka u oblastima mašinskog učenja i veštačke inteligencije, pri čemu je potrebno razmotriti pravne i etičke aspekte automatskog sakupljanja građe u kontekstu pojedinačnog sajta, kao i ciljeve istraživanja i načine upotrebe resursa koji proishode iz istraživanja.

Bibliografija

[1] Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., & Zaragoza, J. (2020). ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4555–4567). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.417

[2] Krotov, V., Johnson, L., & Silva, L. (2020). Legality and Ethics of Web Scraping. *Communications of the Association for Information Systems*, *47*, 539–563. https://doi.org/10.17705/1CAIS.04724

[3] Lefer, M.-A. (2020). Parallel Corpora. In M. Paquot & S. Th. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 257–282). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_12

[4] Lopresti, D. (2009). Optical character recognition errors and their effects on natural language processing. *International Journal on Document Analysis and Recognition (IJDAR)*, *12*(3), 141–151. https://doi.org/10.1007/s10032-009-0094-8

[5] Marovac, U., Avdic, A., & Ljajić, A. (2021). Creating a stop word dictionary in Serbian. *Scientific Publications of the State University of Novi Pazar Series A: Applied Mathematics, Informatics and Mechanics*, *13*, 17–25. https://doi.org/10.5937/SPSUNP2101017M

[6] Norvig, P. (2009). Natural Language Corpus Data. In T. Segaran & J. Hammerbacher (Eds.), *Beautiful Data: The Stories Behind Elegant Data Solutions* (pp. 219–242). O'Reilly Media, Inc. https://norvig.com/ngrams/ch14.pdf

[7] Savourel, Y. (Ed.). (2002). *TMX 1.4a Specification* [OSCAR Recommendation]. Localisation Industry Standards Association (LISA), OSCAR SIG. https://www.ttt.org/oscarStandards/tmx/tmx14-20020710.htm

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

[9] Unicode Consortium. (2025). *Unicode Standard Annex #29: Unicode Text Segmentation*. https://www.unicode.org/reports/tr29/

[10] Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., & Trón, V. (2007). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing IV* (pp. 247–258). https://doi.org/10.1075/cilt.292.32var

Methodology for Creating a Multilingual Parallel Corpus Based on Online Digital User Manuals: the Hilti User Manuals Corpus

Nikola Janković

Summary

This paper presents a methodology for creating an extensive multilingual parallel corpus of user manuals, specifically based on digital user manuals of the company Hilti. The goal of this paper is to enhance research resources in the fields of machine translation, translation theory, as well as translation teaching and other types of translation research. The starting point of the methodology was the uniform HTML structure of the user manuals for all languages on the official Hilti online database of user manuals. Further steps involved identifying HTML elements which contain the relevant segments of texts, their consistent numeration for each language, identifying and cleaning incorrectly parallelized texts, and storing the processed texts in TMX format, so that English is the source language in each file. On average, for each user manual there are around 20 target languages, and the entire corpus contains around 7 million words. The advantages of this corpus include a high level of precision when it comes to segment matching across different languages, a substantial word count, as well as the significant number of languages contained in the corpus. The paper will describe in detail the steps of scraping, processing, and structuring the data, the technical challenges faced in the process and their solutions, as well as the basic statistical data related to the languages in the corpus. We believe that this type of corpus can be used for improvement of translation resources, translation theory, translation teaching, as well as for different types of analyses of specialized language.

Keywords: parallel corpora, machine translation, language tools, specialized language.