JEZIKOSLOVLJE
21.2 (2020): 207-235

UDK 81'322.4
81'246
Izvorni znanstveni članak
Prihvaćen za tisak 17.03. 2020.
https://doi.org/10.29162/jez.2020.8

Sandra Ljubas
Sveučilište u Zadru

Utjecaj višejezičnosti vrednovatelja na ljudsku procjenu kvalitete strojnih prijevoda

U ovom se radu predstavlja istraživanje o utjecaju višejezičnosti vrednovatelja na subjektivnu metodu vrednovanja kvalitete strojnih prijevoda. Subjektivnost ove metode najčešće se očituje u niskim razinama slaganja u ocjenama vrednovatelja. U ovom su preliminarnom istraživanju jedna skupina jednojezičnih i jedna skupina dvojezičnih ispitanika vrednovale kvalitetu istih strojno prevedenih segmenata na razinama točnosti i tečnosti. Segmenti su prevedeni alatom Google Prevoditelj. Jednojezični ispitanici uz strojni prijevod na raspolaganje su dobili i referentni ljudski prijevod, dok su dvojezični ispitanici kvalitetu uspoređivali s izvornikom. Cilj je bio utvrditi kako razlike među jednojezičnim i višejezičnim vrednovateljima uvjetuju način provođenja vrednovanja s obzirom na dužinu trajanja postupka vrednovanja, na odstupanja u prosječnoj ocjeni vrednovanja i analizu uzroka koji uvjetuju razlike u vrednovanju. Analizom rezultata utvrđeno je da dvojezični vrednovatelji u prosjeku daju lošije ocjene izlaznim podacima i da im je potrebno više vremena za vrednovanje, ali nije utvrđeno da kod jedne skupine postoji tendencija prema višoj razini konzistentnosti.

Ključne riječi: vrednovanje; strojno prevođenje; višejezičnost; subjektivna procjena.

1. Uvod

Informacijska i komunikacijska tehnologija neizbježan je dio svakodnevice današnjeg prevoditelja želi li udovoljiti zahtjevima brzine i kvalitete koji se stavljaju pred njega. Iako je samo računalo već duže vrijeme ključan njegov alat, još uvijek raste opseg znanja i vještina povezanih s prevoditeljskim tehnologijama kojima prevoditelji budućnosti moraju baratati, a njihovo prihvaćanje te implementacija dovode do povećanja konkurentnosti i kvalitete rada (Kučiš i Seljan 2014: 304). Istraživanje u

kojemu se jedna skupina ispitanika pri prijevodu legislativnog teksta služila samo tradicionalnim referentnim djelima, dok je druga imala pristup *online* prevoditeljskim alatima, a čije su leksičke, ortografske, sintaktičke i stilističke pogreške zatim analizirane, dokazalo je da moderni prevoditeljski alati doprinose dosljednosti i kvaliteti u prijevodnom procesu te da tehnologije unatoč svojim ograničenjima pridonose značajnom napretku u prevoditeljskoj praksi, osobito u području repetitivnih pravnih, ekonomskih, tehničkih i znanstvenih tekstova (Kučiš i Seljan 2014: 320). Potvrdilo je to i istraživanje provedeno na skupini od 51 ispitanika gdje je uvođenje dodatnih računalno potpomognutih alata pozitivno utjecalo na kvalitetu i dosljednost prijevoda. Naime, oni ispitanici koji su se služili *online* korpusima i višejezičnim terminološkim bazama postigli su daleko bolje rezultate uzevši u obzir broj leksičkih, pravopisnih i interpunkcijskih pogrešaka (Kučiš i dr. 2009: 349). Stoga ne čudi što je i usavršavanje prevoditeljskih alata danas u punom jeku, a usto se javljaju i mnoga nova znanstveno-istraživačka pitanja.

Usporedno s razvitkom alata za strojno prevođenje razvile su se i brojne metode vrednovanja strojnoga prevođenja. Iako danas već postoje i automatske metrike koje svakako imaju svoje posebno mjesto u vrednovanju strojnoga prevođenja, one su i dalje fokusirane na imitiranje ljudskih procjena kvalitete (Han 2018: 9). Ljudska evaluacija, naime, predstavlja „zlatni standard" u vrednovanju strojnoga prevođenja (Brkić i dr. 2011: 93), no kao najveći nedostatak te metode izdvaja se subjektivnost. Subjektivno se ocjenjivanje kvalitete strojnih prijevoda u načelu provodi tako da odabrani vrednovatelji izlaznim podacima dodjeljuju bodove ili ocjenu na unaprijed definiranoj ljestvici. Međutim, zbog svojih stavova, iskustava, načina razmišljanja ili drugih razloga svaki vrednovatelj naprosto ima različito mišljenje o prijevodu koji se nalazi pred njim. To utječe na replikabilnost, ali i pouzdanost metode subjektivnog vrednovanja jer „nije moguće predvidjeti kakav stav i očekivanja ispitanici imaju kada pristupaju strojnome prijevodu" odnosno „ako je ispitanik skeptičan prema jezičnim tehnologijama ili ima averziju prema računalima uopće, vjerojatno će se to nepovoljno odraziti na njegovu ocjenu" (Simeon 2008: 141). Još i gore posljedice mogu imati pretjerani optimizam i preuveličavanje kvalitete izlaznih podataka jer vrednovanje „u javnu raspravu mora unijeti realan pogled na to što sustavi za strojno prevođenje mogu, a što ne mogu učiniti" te mora pomoći predvidjeti „što bi bili u stanju učiniti u budućnosti" (Yusof i dr. 2017: 253). Prije provedbe vrednovanja ključno je usuglasiti se o profilu ispitanika. Zlatni standard u načelu podrazumijeva da neka „bilingvalna osoba usporedi izlazne podatke prevedene s izvornog jezika na ciljni i oformi svoje mišljenje o kvaliteti prijevoda" (Sanders i dr. 2011: 750), ali u praksi provoditelji vrednovanja često žele vrednovati izlazne podatke s raznih izvornih jezika, a na raspolaganju nemaju dvojezičnih

ispitanika za sve te jezične kombinacije. S vremenom se pokazalo da za određene aspekte vrednovanja, npr. za vrednovanje tečnosti, tj. gramatičke ispravnosti i idiomatskog odabira riječi, jednojezični ispitanici ne samo da imaju dovoljne kompetencije, već su ponekad čak i poželjniji. Problemu procjene točnosti prenesenih informacija, odnosno očuvanja izvornog značenja rečenice, doskočilo se uporabom kvalitetnih referentnih prijevoda koje prije istraživanja napravi profesionalni prevoditelj kako bi odabrani vrednovatelji u njima mogli iščitati pravo značenje teksta.

Kako je zasad nejasno razlikuju li se procjene jednojezičnih i višejezičnih vrednovatelja i koliko drastično te postoji li viša razina slaganja među jednima ili drugima, cilj je ovog rada utvrditi utjecaj višejezičnosti na procjenu kvalitete strojnih prijevoda. Ovo pilot-istraživanje usmjereno je na istraživačka pitanja kojima je svrha dati uvid u osnovne probleme postupka vrednovanja i kvalifikacija vrednovatelja strojnih prijevoda. Međutim, takav postupak vrednovanja dugotrajan je i kognitivno mukotrpan proces, pa je ključni nedostatak ovog istraživanja testni skup koji je u cilju olakšavanja procesa za ispitanike sveden na tek 21 segment, odnosno dvije standardne kartice teksta. Stoga su samo ograničeno obuhvaćeni mnogobrojni problemi koji se javljaju pri evaluaciji strojnoga prevođenja, no rezultati ipak pružaju određeni uvid u najznačajnije razlike između jednojezičnih i dvojezičnih vrednovatelja. Važno je napomenuti i što *nije* bio cilj ovog istraživanja. Iako je ovo istraživanje provedeno na konkretnim i stvarnim izlaznim podacima jednog sustava za strojno prevođenje i iako su upravo izlazni podaci najčešći predmet ovakvih istraživanja, ovdje nije bio cilj ustvrditi njihovu kvalitetu. U ovom istraživanju izlazne podatke treba smatrati tek instrumentom s pomoću kojega se aktivira ponašanje vrednovatelja koje želimo promotriti odnosno zadatkom koji je stavljen pred ispitanike kako bi se uvidio i usporedio njihov način donošenja odluka.

Ovaj se rad sastoji od šest cjelina. U drugom se dijelu (Teorijska pozadina) donosi pregled najrelevantnijih istraživanja koja su u posljednje vrijeme provedena u ovoj znanstvenoj disciplini, s naglaskom na istraživanja provedena s hrvatskim jezikom. U trećem je dijelu (Metodologija) opisano provedeno istraživanje te metodologija. Četvrta je cjelina (Rezultati) podijeljena na dva podnaslova u kojima se prvo prikazuju rezultati jednojezične, a zatim dvojezične skupine ispitanika. U petoj cjelini (Diskusija) svaki se od četiri podnaslova bavi jednim istraživačkim pitanjem, a u šestom su dijelu (Zaključci) sažeti najvažniji zaključci ovog istraživanja.

2. Teorijska pozadina

Na temu vrednovanja strojnog prevođenja provedeno je mnogo istraživanja koja uključuju hrvatski jezik, pa već postoje uvidi u to što pri vrednovanju predstavlja najveće probleme. Subjektivna metoda ubraja se u tzv. *black box* metode kod kojih se pri vrednovanju u obzir uzima samo krajnji rezultat, odnosno prijevod koji neki sustav ponudi, za razliku od tzv. *glass box* metoda kod kojih se kvaliteta sustava za strojno prevođenje procjenjuje ispitivanjem njegovih unutarnjih komponenti (Dorr i dr. 2011: 745). Pogled kroz „staklenu kutiju" u načelu je informativniji za programere i tvorce sustava za strojno prevođenje, dok „crna kutija" ne može pružiti objektivne podatke o tome zašto se sustav ponaša na određeni način, ali itekako mnogo govori o njegovoj funkcionalnosti i uspješnosti. *Black box* metode često se temelje na usporedbi izlaznih podataka s ljudskim prijevodom: vrednovatelji procjenjuju kvalitetu u odnosu na referentni prijevod fokusirajući se na određene karakteristike prijevoda, najčešće točnost i tečnost (Dorr i dr. 2011: 746).

Glavni su izazovi metode subjektivne procjene strojnih prijevoda: vrijeme potrebno za provođenje metode, skupocjenost te pouzdanost i konzistentnost (Lavie 2013: 5). Kao što je ranije spomenuto, niska korelacija među sudovima vrednovatelja rezultat je toga što ljudi različitih pozadina i iskustava sa strojnim prevođenjem drukčije procjenjuju težinu primjerice sintaktičkih pogrešaka ili pogrešaka u stilu (Nübel 1997., cit. prema Sanders i dr. 2011: 752). Prevoditeljska praksa također nas uči da ne postoji samo jedan ispravan prijevod kojemu treba težiti, već da se za formulacije u jednome jeziku mogu naći mnogobrojni legitimni ekvivalenti na ciljnome jeziku. To se jasno vidi u istraživanju englesko-hrvatskih tekstova iz pravne domene provedenom na testnom skupu od 200 rečenica (Seljan i dr. 2012). Vrednovanje je provedeno prema kriterijima tečnosti i točnosti, a zatim je provedena analiza vrsta pogrešaka u kategorijama: neprevedene, ispuštene i nepotrebno prevedene riječi, morfološke, leksičke i sintaktičke pogrešne te interpunkcija. Ljudski su vrednovatelji odabrane segmente analizirali uspoređujući ih s čak tri različita referentna ljudska prijevoda. Prvi je referentni prijevod predstavljao službeni, a zamijećeno je kako je svaki sljedeći referentni prijevod sadržavao sve manje riječi zbog uklanjanja redundancija koje su karakteristične za legislativni registar. Ipak, zadržano je puno značenje izvornih rečenica, kao i prikladan stil (Seljan i dr. 2012: 2144). Dodatno je u tom istraživanju proučen i rad automatske metrike BLEU u odnosu na veći broj referentnih prijevoda, pa je ustanovljeno da je BLEU bolje procijenio prvi referentni prijevod od druga dva, ali da je najbolji rezultat dao upravo kada je u obzir uzeo sva tri referentna prijevoda (Seljan i dr. 2012: 2147).

Potrebna kvaliteta strojnog prijevoda uvelike ovisi i o tome za što će se prijevod rabiti. Sve to značajno otežava vrednovanje i stoga je potreban niz nezavisnih vrednovatelja da bi prosjek njihovih ocjena bio relevantniji pokazatelj kvalitete. Međutim, i konzistentnost ljudskog suda ima dva aspekta: gotovo je jednako teško osigurati konzistentnost jednog jedinog vrednovatelja tijekom dugotrajnog i kognitivno napornog procesa vrednovanja (engl. *intra-coder agreement*) kao i konzistentnost među većim brojem vrednovatelja (engl. *inter-coder agreement*) (Lavie 2013: 5). U istraživanju evaluacije alata za strojno prevođenje iz 2011. Seljan i dr. analiziraju četiri domene tekstova (opis grada, legislativa, sport i tehnika) prevedenih s hrvatskog na engleski jezik primjenom četiriju prevodilačkih alata (Google Prevoditelj, Stars21, InterTran i Translation Guide) te s engleskog na hrvatski jezik primjenom alata Google Prevoditelj. Evaluaciju je provelo 48 ispitanika i utvrđeno je da se u različitim domenama javljaju različite vrste pogrešaka: u sportskoj je domeni zamijećen najveći broj neprevedenih riječi, a u tehničkom tekstu najveći broj morfoloških pogrešaka. U legislativnom je tekstu bilo najmanje pogrešaka općenito i ravnomjerno su raspoređene među kategorijama, dok je u opisu grada bilo najviše leksičkih pogrešaka (Seljan i dr. 2011: 340–341). Prijevodi s engleskog na hrvatski načelno su procijenjeni lošijima osim u domeni opisa grada, a zanimljivo je da je kod rangiranja alata koji su prevodili s hrvatskog na engleski jezik utvrđen vrlo visok stupanj međusobne konzistentnosti (Seljan i dr. 2011: 342).

S istim kriterijima tečnosti i točnosti te uz analizu pogrešaka prema šest kategorija (neprevedene i dodane riječi, morfološke, leksičke i sintaktičke pogreške te interpunkcija) Seljan i dr. (2015a) prikazuju rezultate za alate za strojno prevođenje Yandex i Google Prevoditelj. Ljudska je evaluacija provedena na 400 rečenica i utvrđeno je da se u njima javlja najviše morfoloških pogrešaka, iako je manje pogrešaka uočeno kod rusko-hrvatskog jezičnog para u odnosu na englesko-hrvatski. Vrednovatelji su lošije ocjene davali tečnosti, a bolje točnosti jer je zabilježen veći problem s formiranjem rečenica nego semantikom (Seljan i dr. 2015a: 1096). Moguće je, međutim, analizirati kvalitetu izlaznih podataka i prema drugim kriterijima. Seljan i Dunđer provode analizu prepoznavanja govora i strojnog prevođenja gdje su izlazni rezultati vrednovani prema kriteriju korisnosti koji obuhvaća razumljivost, točnost, tečnost i općenito zadovoljstvo vrednovatelja ekvivalentom (2014: 1982), a Seljan i dr. analiziraju kombiniranu upotrebu alata za sažimanje i alata za *online* prevođenje s engleskog, njemačkog i ruskog na hrvatski te provode ljudsku evaluaciju prema pitanjima: „tko, gdje, kada, kako i zašto“ na razini rečenica i na razini korpusa (2015b). Zamijećen je važan aspekt percepcije prenesene informacije kod korisnika jer su pojedine rečenice dobivale mnogo bolje ocjene u odnosu na evaluaciju teksta u cjelini (Seljan i dr. 2015b: 209).

Iako uobičajeni, kriteriji točnosti i tečnosti nikako nisu jedini kriteriji vrednovanja. Tijekom bogate povijesti vrednovanja strojnog prevođenja definirani su kriteriji razumljivosti, vjernosti i informativnosti (u kojima su donekle sadržani pojmovi točnosti i tečnosti), ali i potpuno nezavisni kriteriji poput prikladnosti (odgovaraju li izlazni podaci određenom kontekstu u kojemu će se sustav rabiti?), interoperabilnosti (funkcionira li sustav i na drugim softverskim ili hardverskim platformama?), pouzdanosti („pada" li sustav često i koliko mu je u tom slučaju potrebno da se ponovno pokrene?), koristivosti (rabi li se sučelje sustava lako i intuitivno?), učinkovitosti (koliko brzo prevodi sustav i koliko brzo ažurira promjene kada se prijevod revidira?), mogućnosti održavanja (može li se sustav prilagođavati i mijenjati u skladu s potrebama korisnika?), prenosivosti (može li nova verzija sustava lako zamijeniti staru kada se unaprijedi i popravi?) itd. (King i dr. 2003: 3–4). Kriteriji „vjernosti" i „točnosti" najjasnije se preklapaju pa često i izjednačavaju jer se oba odnose na očuvanje izvornog značenja rečenice (Simeon 2008: 29) i pružaju odgovor na pitanje jesu li izlazni podaci točno/vjerno prenijeli značenje koje sadržava izvorna rečenica. Kod tečnosti vrednovatelj pak postavlja pitanje jesu li izlazni podaci izraženi na „dobrom", tečnom jeziku, što uključuje i gramatičku ispravnost i idiomatski odabir riječi (Koehn 2009: poglavlje 8). U uputama za uporabu sučelja MQM (*Multidimensional Quality Metrics*[1]) navodi se da je tečnost povezana s „jednojezičnim kvalitetama izvornog ili ciljnog jezika" i da se procjenama tečnosti teksta može pristupiti i potpuno zanemarivši činjenicu da se radi o prijevodu. U uputama se također navode specifične pogreške koje negativno utječu na tečnost, a to su: pravopisne pogreške i zatipci (između ostalog u to uključuju i pisanje velikog i malog slova), tipografija (pretjerana ili pogrešna uporaba pravopisnih znakova) te gramatičke pogreške (sintaktičke pogreške, uporaba pogrešnog oblika ili vrste riječi, sročnost, pogrešno odabrano glagolsko vrijeme ili način, raspored riječi i neispravna uporaba funkcionalnih riječi poput prijedloga ili čestica).

Kao što je ranije spomenuto, kriteriji točnosti i tečnosti čine jednu cjelinu, ali istovremeno su i neovisni jedan o drugome. Prijevod može biti tečan čak i kada značenje izvornika nije vjerno preneseno, a značenje može biti vjerno preneseno čak i kada tekst nije u potpunosti tečan. No, da bi izlazni podaci bili prihvatljivi, oba kriterija moraju biti barem donekle zadovoljena. Stoga se kao ocjena kvalitete prijevoda često uzima aritmetička sredina ocjena pripisanih svakom od kriterija na nekoj bodovnoj ljestvici (Snover i dr. 2009: 259). Određene pogreške snažno utječu samo na jednu od kategorija; pogrešno odabrano glagolsko vrijeme, primjerice, ne

[1] Dostupno na: http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf. Datum posjeta stranici: 29. studenog 2018.

utječe na točnost[2] prijevoda, ali može smanjiti ocjenu pri procjenjivanju tečnosti, dok s druge strane izostavljanje neke riječi neće nužno utjecati na tečnost prijevoda, ali hoće na točnost (Snover i dr. 2009: 260). Kriteriji se prvo moraju ocijeniti zasebno, a aritmetička sredina objektivnije predstavlja kvalitetu prijevoda u cjelini.

Također valja definirati i sam pojam višejezičnosti koji se na prvu može činiti samorazumljivim, ali čak ni u psiholingvistici ne postoji jedna uvriježena definicija višejezičnosti. Najuža definicija polazi od toga da je osoba višejezična samo onda kada je od rođenja zbog različitih obiteljskih, socijalnih, kulturoloških ili društvenih uvjeta učila komunicirati na barem dva jezika (Höhle 2012: 175). Prema nešto široj definiciji višejezični su svi ljudi koji redovito komuniciraju na više od jednog jezika, što pretpostavlja da je njihovo jezično znanje na dovoljno visokoj razini da komunikacija protječe bez problema (Höhle 2012: 175). U ovom se istraživanju radi s njemačkim jezikom, pa se među višejezične odnosno dvojezične osobe ubrajaju oni ispitanici koji procjenjuju da im je znanje njemačkog jezika na razini C1/C2 prema europskom referentnom okviru za jezike[3] te oni koji se profesionalno bave prevođenjem u kombinaciji s njemačkim jezikom, predavanjem njemačkog i/ili koji su studirali njemački na fakultetu. Među jednojezične ispitanike u ovom se istraživanju ubrajaju svi koji ne govore njemački jezik ili ga poznaju na nižim razinama (A1–B2). Kako su i oni po zanimanju prevoditelji, profesori i lingvisti, vrlo je vjerojatno da su u zbilji i oni dvojezični, ali ih unutar okvira ovog istraživanja provedenog s njemačkim jezikom nazivamo jednojezičnima.

Naposljetku, u ovom je istraživanju posebna pažnja usmjerena upravo na dosad neistražene aspekte vrednovanja koji se tiču višejezičnosti. Kako bi se nadopunili uvidi u to koliko ta karakteristika vrednovatelja utječe na njihovu procjenu kvalitete izlaznih podataka, u ovom se radu nastoje pružiti odgovori na to (1) postoji li razlika u vremenskom trajanju vrednovanja strojnih prijevoda kod jednojezičnih i dvojezičnih vrednovatelja i koji su mogući razlozi za to, (2) koje su razlike u procjenjivanju kriterija točnosti kod jednojezičnih i dvojezičnih vrednovatelja, (3) kolika se odstupanja između ocjena točnosti i tečnosti javljaju u pojedinoj skupini vre-

[2] Autori se zapravo služe terminom prikladnost (engl. *adequacy*), ali navode da se tim kriterijem mjeri „prenosi li prijevod ispravno značenje“ (Snover i dr. 2009: 259). Dakle, „prikladnost“ definiraju jednako kao što je „točnost“ definirana u ovome radu. U teoriji vrednovanja postoji još i više naziva koji se odnose na isti koncept „ispravnog prenošenja značenja/informacija iz izvornika“, ali u ovom se radu pokušava izbjeći ta terminološka nedosljednost.

[3] Prema zajedničkom europskom referentnom okviru za jezike osobe na stupnjevima A1 i A2 definiraju se kao temeljni korisnici, osobe na stupnjevima B1 i B2 kao samostalni korisnici, a osobe na stupnjevima C1 i C2 kao iskusni korisnici (v. https://europass.cedefop.europa.eu/sites/default/files/cefr-hr.pdf za detaljniju analizu stupnjeva).

dnovatelja i koji su mogući razlozi za to, te (4) postoji li unutar jedne od skupina veća razina konzistentnosti (*inter-coder agreement*)?

3. Metodologija

U ovom su istraživanju sudjelovala 24 ispitanika kojima je materinski jezik hrvatski. Ravnomjerno su bili podijeljeni u dvije skupine: skupinu A činilo je 12 ispitanika koji njemački jezik ne razumiju ili ga razumiju slabije, a skupinu B 12 bilingvalnih govornika hrvatskog i njemačkog jezika. U skupini A svi su ispitanici bili profesionalni prevoditelji ili profesori stranih jezika s iskustvom u prevođenju, dok je u skupini B takvog profila bilo sedam ispitanika, a preostalih pet ispitanika činili su studenti njemačkog jezika na završim godinama studija koji su već stekli prva iskustva s prevođenjem.

Obje skupine procjenjivale su kvalitetu izlaznih podataka segment po segment prema kriterijima točnosti i tečnosti na ljestvici s 5 bodova (tablica 1). Točnost i tečnost najčešće su rabljeni kriteriji za evaluaciju strojnih prijevoda: pritom točnost označava u kojoj su mjeri informacije iz izvornika prenesene u prijevod, a tečnost u kojoj su mjeri izlazni podaci prirodni, idiomatski i izraženi na dobrom hrvatskom (Seljan i dr. 2015a). Ispitanici su prije početka istraživanja upućeni na tablicu 1 kako bi se mogli upoznati s vrijednostima bodova. Teško je osigurati da pojedine vrijednosti isto znače različitim ispitanicima, pa je, u nadi da će to pripomoći u osiguravanju konzistentnosti, odabrana ljestvica s ocjenama od 1 do 5 koje hrvatski vrednovatelji prepoznaju iz obrazovnog sustava (usp. sličnu ljestvicu za vrednovanje točnosti predlažu Seljan i dr. 2012[4]).

Tablica 1. Bodovna ljestvica

Točnost		Tečnost	
1	ništa značenja nije preneseno	1	nije moguće pratiti
2	vrlo je malo značenja preneseno	2	ne čita se tečno, potrebno puno mozganja
3	dosta je značenja preneseno	3	zamjećuju se određene poteškoće
4	većina je značenja prenesena	4	dobro, potrebne tek manje preinake
5	značenje je u potpunosti preneseno	5	u potpunosti prirodno i razumljivo

[4] „Insufficient/inadequate/wrong information (1), Barely enough information (2), Intermediate level of information preserved (3), Very good but not complete (4), Complete information preserved (5)" (Seljan i dr. 2012: 2145).

Ispitanici su dobili upute da vrednuju strojne prijevode tako da svakom segmentu prvo pridruže ocjenu za kriterij točnosti, a zatim za kriterij tečnosti. Kao i u prijašnjim istraživanjima (usp. Brkić i dr. 2011) u obzir se uzela aritmetička sredina tih dviju ocjena. Ispitanici su imali priliku iskušati zadatak na dva primjera za vježbu, a potom su vrednovali ukupno 21 segment. Ispitanici su se tijekom ocjenjivanja mogli vraćati na prošle segmente i po željama mijenjati svoje odgovore. Prije početka vrednovanja ispitanici su zamoljeni da bilježe vrijeme potrebno za dovršavanje zadatka.

U istraživanju je provedena evaluacija za njemačko-hrvatski jezični par na tekstovima iz domene Europske unije, preuzetih s repozitorija clarin:el[5] (za izvorni testni skup v. Dodatak A). Kako su u prijevodu zamijećene određene pravopisne pogreške, a pri procjenama kvalitete od iznimne je važnosti da referentni prijevod bude kvalitetan, hrvatski je tekst prvo lektoriran (za referentni testni skup v. Dodatak B). Taj je uređeni referentni prijevod na raspolaganje stavljen skupini A, dok je skupina B dobila uvid u njemački izvornik. Ljudska je evaluacija provedena na 21 rečenici odnosno dvije kartice teksta od 1800 znakova s uključenim razmacima. Segmenti su evaluirani odvojeno, no u cjelini čine koherentan tekst. Izlazni podaci koje su ispitanici vrednovali generirani su 1. prosinca 2018. na neuronskoj verziji Google Prevoditelja (za strojno prevedeni testni skup v. Dodatak C). Tekstovi su pisani općim jezikom i nisu sadržavali stručne termine. Najkraća rečenica u izvorniku imala je devet, a najduža 46 riječi. Prosjek duljine rečenica iznosio je 22 riječi.

Premda su rezultati ovog istraživanja kvantitativno izraženi u obliku prosječnih ocjena na ljestvici od 1–5, zbog relativno malog korpusa istraživanja (s 24 ispitanika te 21 evaluiranim segmentom) nije provedena statistička provjera utvrđenih razlika te su rezultati prvenstveno analizirani kvalitativno. Ovi se rezultati smatraju preliminarnima i tek daju uvid u moguće odgovore na postavljena istraživačka pitanja.

4. Rezultati

U sljedeća se dva podnaslova predstavljaju rezultati skupine A s jednojezičnim ispitanicima odnosno skupine B s dvojezičnim ispitanicima.

[5] Dostupno na:
https://athena.clarin.gr/resources/download/0e666350523a11e59319aa3fc8d33ad828bf2b7f786b47fc8ffd138c862607aa/ Datum posjeta stranici: 6. siječnja 2019.

4.1. *Rezultati skupine A*

Prosječna ocjena točnosti za skupinu jednojezičnih ispitanika iznosi 3,2, dok prosječna ocjena dodijeljena tečnosti segmenata iznosi 3. Ukupno su ispitanici izlaznim podacima dali 3,1 bod. Što se točnosti tiče, segment najlošije kvalitete dobio je 2,4 boda, a najboljem je segmentu pripisano 4,3 boda. Isti je segment za kriterij tečnosti ocijenjen s 4,2 boda, a najmanje tečan segment dobio je 2,3 boda. Najveća razlika između segmenata iznosi, dakle, gotovo 2 boda za oba kriterija (v. tablicu 2 za prikaz ovih rezultata u suodnosu s rezultatima skupine B). No kada pogledamo sve bodove dodijeljene pojedinim segmentima, a ne samo prosječne ocjene, vidimo da se sudovi o kvaliteti ovisno o ispitanicima razlikuju još i više. Čak su tri segmenta u obje kategorije ocijenjena svim mogućim bodovima od 1 do 5, a samo kod dva segmenta u obje kategorije raspon se bodova razlikuje samo za 1 (npr. svi su ispitanici segmentu dali 2 ili 3 boda). Zabilježena su i dva segmenta kod kojih je jedan vrednovatelj smatrao da ništa značenja nije preneseno i da nije moguće pratiti segment, a dva su smatrala da je značenje u potpunosti preneseno i da je segment u potpunosti tečan. Takav je slučaj primjer (1):

(1) *Procjenjuje se da 40 milijuna ljudi u EU govori materinski jezik različit od službenih jezika svoje zemlje podrijetla.*

Referentni prijevod: „*Procjenjuje se da više od 40 milijuna ljudi u EU govori materinskim jezikom koji nije službeni jezik njihove zemlje podrijetla.*“

Izvornik: „*Schätzungsweise 40 Millionen Menschen in der EU sprechen eine Muttersprache, die nicht die offizielle Sprache ihres Herkunftslandes ist.*“

Naposljetku, postupak vrednovanja 21 segmenta vrednovateljima je u prosjeku oduzeo 10 minuta, varirajući od najmanje 5 do najviše 15 minuta.

Tablica 2. Rezultati skupina A i B

	Skupina A	Skupina B
Prosječna ocjena točnosti	3,2	2,9
Prosječna ocjena tečnosti	3	2,8
Ukupna prosječna ocjena kvalitete	3,1	2,85
Maksimalan raspon u ocjenama točnosti odnosno tečnosti	1,9 / 1,9	2,8 / 3
Prosječno vrijeme potrebno za provođenje istraživanja (u minutama)	10	13

4.2. *Rezultati skupine B*

Kao što je vidljivo iz tablice 2, prosječna ocjena točnosti za skupinu dvojezičnih ispitanika iznosi 2,9, a tečnosti 2,8. Izlazni su podaci ukupno dobili 2,85 bodova, što je nešto niže od ocjene jednojezičnih ispitanika. Najlošije je ocijenjeni segment u kategoriji točnosti dobio 1,8 bodova, a u kategoriji tečnosti 1,5. Najbolje je ocijenjeni segment u kategoriji točnosti dobio 4,6 bodova, a u kategoriji tečnosti 4,5. To znači da rasponi u broju bodova iznose 2,8 odnosno 3, što je osjetno više variranje nego što je to bio slučaj u skupini A. Sudovi se vrednovatelja još više razlikuju: pet je segmenata ocijenjeno svim mogućim bodovima od 1 do 5 u kategoriji točnosti, a u kategoriji tečnosti ponovno su tri segmenta ocijenjena svim bodovima. U kategoriji točnosti zabilježena su dva segmenta s rasponom bodova od samo 1, a u kategoriji tečnosti samo je jedan takav primjer (u kojemu su svi ispitanici segmentu dali 4 ili 5 bodova). Ponovno je zabilježeno da su za dva segmenta ispitanici smatrali da ih nije moguće pratiti i da ništa značenja nije preneseno, ali im je jedan ispitanik dao najviše bodove za oba kriterija smatrajući da je značenje u potpunosti preneseno i da je segment točan. Takav je primjer (2):

(2) *Europski socijalni fond, na primjer, uz potporu talijanske Arturo Toscanini Foundation, koji drži tečajeve za nezaposlene glazbenika od sredine 90-ih godina.*

Referentni prijevod: „*Europski socijalni fond, primjerice, podupire talijansku Zakladu 'Arturo Toscanini' koja od sredine 90-ih godina održava tečajeve za glazbenike bez posla*."

Izvornik: „*So unterstützt der Europäische Sozialfonds beispielsweise die italienische Arturo-Toscanini-Stiftung, die seit Mitte der 90er Jahre Kurse für arbeitslose Musiker abhält*."

U ovom je segmentu već iz strojnog prijevoda jasno da dio informacija nedostaje, a uvidom u referentni prijevod ili izvornik vrednovatelj shvaća da su neke informacije čak i potpuno pogrešne: nije „Arturo Toscanini Foundation" ta koja podupire „europski socijalni fond koji drži tečajeve za nezaposlene glazbenike", što bi se iz izlaznih podataka dalo zaključiti, već je u stvarnosti upravo obrnuto. Stoga je jasno kako je došlo do nižih ocjena, ali ne i zašto je jedan ispitanik ipak smatrao da segment zaslužuje najviše ocjene. Nedostatak *black box* metoda poput ove leži upravo u tome što o uspješnosti sustava možemo suditi samo po krajnjim ocjenama vrednovatelja, a jednako kao što ništa ne saznajemo o tome koje su komponente sustava dovele do upravo takvih izlaznih podataka, tako ništa ne saznajemo ni o čimbenicima koji su doveli do upravo ovih odluka vrednovatelja.

I vrijeme potrebno za rješavanje zadatka u skupini B variralo je mnogo više nego u skupini A. Ispitanicima je bilo potrebno najmanje pet, ali najviše čak 30 minuta, što je dvostruko nadmašilo predviđeno trajanje istraživanja, ali i vrijeme koje je bilo potrebno „najsporijem“ ispitaniku skupine A. Istraživanje je vrednovateljima skupine B u prosjeku oduzelo 13 minuta.

5. Diskusija

Analiza rezultata ukazala je na zanimljive razlike između jednojezičnih i dvojezičnih ispitanika. Na primjerima iz korpusa u sljedećim će se podnaslovima pokazati što se može zaključiti o postupku vrednovanja kvalitete strojnoga prevođenja s obzirom na višejezičnost vrednovatelja. Iako se u okviru ovog istraživanja ne mogu dati definitivni odgovori na sva postavljena istraživačka pitanja, jasno se vide tendencije koje bi u sljedećoj fazi valjalo provjeriti na većem uzorku.

5.1. *Istraživačko pitanje I*

Prvo je istraživačko pitanje glasilo postoji li razlika u trajanju vrednovanja strojnih prijevoda kod jednojezičnih i dvojezičnih vrednovatelja i koji su mogući razlozi za to. Istraživanje je pokazalo da je ispitanicima skupine B u prosjeku trebalo tri minute dulje da dovrše vrednovanje nego što je trebalo ispitanicima skupine A. U skupini B veći se broj ispitanika približio predviđenom trajanju istraživanja od 15 minuta, jednom je ispitaniku trebalo 19, a drugome čak 30 minuta. U skupini A ispitanicima je istraživanje većinski oduzelo deset minuta, a tek je jednom ispitaniku bilo potrebno 15 minuta. U obzir treba uzeti da se ovdje radi o kraćem istraživanju zbog čega se ove razlike mogu činiti zanemarivima, ali pri vrednovanju duljih tekstova te bi se razlike mogle proporcionalno povećavati. Svakako je jasno uočena tendencija da je dvojezičnim ispitanicima potrebno više vremena za vrednovanje nego što je potrebno jednojezičnima, a ostaje pitanje zašto je to tako.

Javljaju se razna objašnjenja, a vjerojatno je da ulogu igra tzv. *code-switching*, odnosno činjenica da se dvojezični ispitanici u glavi neprestano tijekom vrednovanja moraju „prebacivati“ s jednog jezika na drugi, dok se jednojezičnim ispitanicima cjelokupno vrednovanje odvija na istome jeziku. Drugo je moguće objašnjenje da postoji korelacija između vremena potrebnog za vrednovanje i dosadašnjeg iskustva s vrednovanjem strojnog prevođenja. Iako su brojna istraživanja pokazala kako *online* alati pozitivno utječu na kvalitetu prijevoda (v. Kučiš i dr. 2009, Kučiš i Seljan 2014), tek su dva ispitanika iz skupine A izjavila da je vrednovanje strojnih prijevoda dio njihove poslovne svakodnevice. Međutim, zamijećeno je da su upra-

vo oni vrednovanje obavili u osjetno kraćem roku, što je utjecalo na kraći prosjek za skupinu A. Ispitanicima koji se ranije nisu susreli s vrednovanjem strojnog prevođenja u pravilu je trebalo nešto više vremena, uz tek dvije iznimke ispitanika bez dosadašnjeg iskustva s vrednovanjem koji su vrednovanje obavili u kraće od deset minuta. Ove informacije ne govore nam samo o duljini istraživanja, već svjedoče i o jednom drugom fenomenu važnom za subjektivno vrednovanje, a to je da svaki vrednovatelj, ovisno o svojim dosadašnjim iskustvima, različito pristupa zadatku vrednovanja. Nažalost, metodom se subjektivnog vrednovanja ne mogu ustvrditi korelacije između iskustava i stavova ispitanika i njihove procjene kvalitete prijevoda niti se oni mogu neutralizirati kako bi vrednovanje bilo objektivno, što je i jedan od razloga zašto su se u povijesti vrednovanja smišljale druge metode koje bi doskočile tom problemu.

5.2. *Istraživačko pitanje II*

Drugo se istraživačko pitanje odnosilo na to postoje li razlike u procjenjivanju kriterija točnosti kod jednojezičnih i dvojezičnih vrednovatelja. Pretpostavka je da su jednojezični ispitanici svjesni da pred sobom nemaju izvornik koji je doista jedini „izvorni" nositelj informacija, pa se doslovnije pridržavaju referentnog prijevoda umjesto da, kao što i sami zasigurno rade dok prevode, dopuštaju više mogućih prijevodnih ekvivalenata i rješenja koja izražavaju osnovnu misao. I ovdje se vidi jasna tendencija: skupina A izlaznim je podacima prosječno dala 3,2 boda, a skupina B tek 2,9, a zapravo se isti trend zamjećuje i za kriterij tečnosti (skupina A = 3 boda, skupina B = 2,8 bodova). U čak 15 segmenata točnost je bolje ocijenjena kod skupine A, a samo preostalih šest kvalitetnijima je procijenila skupina B, pri čemu su razlike među ocjenama obiju skupina neznatne. Izdvaja se samo jedan primjer izlaznih podataka koji je skupina B ocijenila boljim za čak 0,6 bodova:

(3) *Svi trošimo u prosjeku do tri sata gledajući vijesti, sport, filmove i druge programe.*

Referentni prijevod: „*Svatko od nas u prosjeku provodi do tri sata dnevno gledajući vijesti, sport, filmove i druge sadržaje.*"

Izvornik: „*Wir alle verbringen durchschnittlich bis zu drei Stunden damit, Nachrichten, Sport, Filme und andere Programme anzuschauen.*"

Četiri su ispitanika skupine A smatrali da je značenje „u potpunosti" preneseno (5 bodova), tri ispitanika da je „većina" značenja preneseno (4 boda), četiri da je „dosta značenja preneseno" (3 boda), a jedan je smatrao da je tek „vrlo malo" značenja preneseno (2 boda). Razloge za taj nesrazmjer možemo pokušati pronaći u uspo-

redbi izlaznih podataka s referentnim prijevodom. Razlikuju ih tri glavne točke:

1) U hrvatskom glagol „trošiti“ ima negativan prizvuk za razliku od neutralnog „provoditi“. Jedan bi vrednovatelj mogao procijeniti da se tu radi o stilističkoj pogrešci koja tek blago utječe na značenje teksta, dok bi drugi mogao argumentirati da se radi o leksičkoj pogrešci koja grubo narušava kvalitetu teksta, pa bi sukladno tome dodijelili svoje bodove.

2) U izlaznim je podacima izostavljen prilog „dnevno“ koji precizira vremenski okvir gledanja televizije. Neki će vrednovatelji smatrati logičnim da se televizija gleda „do tri sata *dnevno*“ i neće zamjeriti što to nije precizirano, dok drugima to nikako neće biti samo po sebi razumljivo i pitat će se odnose li se ta tri sata na jedan dan, jedan tjedan, mjesec ili neki drugi, više ili manje vjerojatni, vremenski okvir. U ovakvim primjerima, međutim, jasno vidimo kako pogreške izostavljanja riječi utječu samo na točnost, ali ne i na tečnost segmenta (v. Snover i dr. 2009).

3) Dio će vrednovatelja cijeniti precizniji ekvivalent „sadržaji“ u referentnom prijevodu, za razliku od ponešto nejasne riječi „programi“ u izlaznim podacima. Drugi se pak neće zamarati takvim nijansama. Sve to doprinosi niskoj razini slaganja i većim razlikama u ocjenama.

Proučimo sada skupinu B koja točnost izlaznih podataka provjerava u izvorniku. Čak šest ispitanika smatra da je značenje „u potpunosti preneseno“ (5 bodova), a pet ih smatra da je „većina“ značenja prenesena (4 boda). Tek jedan ispitanik smatra da je „dosta“ značenja preneseno (3 boda). Osim što je kvaliteta procijenjena boljom, postignuta je i veća razina slaganja među vrednovateljima. Kako je do toga došlo? Ponovno ćemo provjeriti iste tri značajke teksta:

1) Hrvatski glagol „trošiti“ ima negativan prizvuk u usporedbi s neutralnim njemačkim glagolom „verbringen“ iz izvornika. Nema velikih razlika u odnosu na usporedbu s referentnim prijevodom.

2) Međutim, priloga „dnevno“ nema ni u njemačkom izvorniku. Postaje jasno da je prevoditelj sam odlučio precizirati vremenski okvir ili je ubacivanjem priloga rečenicu htio učiniti tečnijom. Njegova prevoditeljska odluka nije upitna u smislu kvalitete prijevoda, ali pri vrednovanju izazva određenu pomutnju: ispitanici skupine A smatraju da je došlo do izostavljanja riječi i mogu to smatrati pogreškom, a ispitanici skupine B na to neće ni obratiti pažnju.

3) Kako se u izvorniku javlja njemačka imenica „Programme“, i u ovom je elementu vjerojatno da su vrednovatelji skupine B automatski smatrali da su „programi“ dobar prijevodni ekvivalent koji nisu dalje propitivali.

Primjer (3) ogledno prikazuje kako i najmanji detalji mogu utjecati na percepciju prenesene informacije kod korisnika (usp. Seljan i dr. 2015b) te kako čak i referentni prijevodi mogu predstavljati određen problem za vrednovanje. Ipak, ocjene ispitanika nikako se ne smiju promatrati samo kao suhoparan zbroj ispravnih i neispravnih ekvivalenata jer bi u tom slučaju među vrednovateljima postojala daleko viša razina slaganja. Kvaliteta samih prijevodnih ekvivalenata prvenstveno se utvrđuje metodama poput analize pogrešaka, no one nam ne daju informacije o tome koliko je neki strojno prevedeni tekst ljudima koristan i *smatraju* li ga kvalitetnim, što je daleko nepredvidljivije.

Primjer (3) pokazuje kako materijalne značajke strojnog prijevoda, referentnog prijevoda ili izvornika mogu usmjeriti dio vrednovatelja na određene sudove. No primjer (4) oglednije pokazuje kako jednojezični ispitanici ipak tendiraju dati bolje ocjene od dvojezičnih. Skupina B primjeru (4) dala je 2,5 bodova, a skupina A čak 3,3 boda. Šest je ispitanika u skupini A i pet ispitanika u skupini B smatralo da je „dosta" značenja preneseno (3 boda), no do razlike u prosječnoj ocjeni od 0,8 boda dovele su ocjene ostalih vrednovatelja koje su pošle u posve različitim smjerovima. Dvoje vrednovatelja skupine A smatralo da je značenje „u potpunosti" preneseno (5 bodova), a niti jedan nije smatrao da „ništa" značenja nije preneseno (1 bod), dok se u skupini B dogodilo upravo suprotno: niti jedan vrednovatelj nije smatrao da je značenje „u potpunosti preneseno", a dva su smatrala da „ništa" značenja nije preneseno.

(4) *Svake godine jedan ili dva grada su odabrani kao europski kulturni prijestolnici, koji se kvalificiraju za financijsku potporu.*

Referentni prijevod: „*Svake se godine jedan ili dva grada proglase kulturnim prijestolnicama Europe, čime se kvalificiraju za financijsku potporu.*"

Izvornik: „*Jedes Jahr werden eine oder zwei Städte als europäische Kulturhauptstädte ausgewählt, die sich dadurch für finanzielle Unterstützung qualifizieren*."

Zašto se ovi izlazni podaci čine boljima u usporedbi s referentnim prijevodom, a toliko lošijima u usporedbi s izvornikom? Pokušali smo odgovore potražiti analizom elemenata i zamijetili smo tek da su izlazni podaci prilično slični izvorniku. Glavna je rečenica, primjerice, i u izlaznim podacima kao i u izvorniku izražena pasivnim oblikom, što je u skladu s pravilima njemačkog jezika, ali u hrvatskom jeziku nije uobičajeno u istoj mjeri. U radnoj praksi mnogi su prevoditelji s njemačkog jezika (ispitanici skupine B) automatski naučili izbjegavati takve konstrukcije, pa je moguće da su stoga tu pogrešku smatrali ozbiljnijom. Odluke ispitanika

mogle bi biti jasnije kada bi se sa svakim ispitanikom ponaosob proveo intervju u kojemu bi oni argumentirali svoje odluke, ali nije sigurno da bi se time išta promijenilo u razini slaganja među vrednovateljima jer se i težina svakog argumenta može relativizirati[6].

5.3. *Istraživačko pitanje III*

Treće se istraživačko pitanje odnosilo na odstupanja između ocjena točnosti i tečnosti u pojedinim skupinama vrednovatelja. Pretpostavljalo se da će dvojezični ispitanici lakše odvojiti kategorije točnost i tečnosti jer prvo uspoređuju točnost prenesenih informacija s izvornikom na njemačkom jeziku, a zatim se potpuno neovisno o njemu usredotoče na tečnost samo na hrvatskom jeziku. Jednojezični bi ispitanici prema tome svoju procjenu o točnosti i tečnosti formirali istovremeno pa bi i vjerojatnost da tim dvama kriterijima daju iste ili slične ocjene bila viša. Za ovo se istraživačko pitanje, međutim, ne može jasno procijeniti odgovor.

U objema se skupinama našao po jedan vrednovatelj koji je u svim segmentima jednako bodovao točnost i tečnost. U obje skupine postoje ispitanici koji su smatrali da u više od pola segmenata postoji jasna razlika između točnost i tečnosti. No prosjek je da vrednovatelji otprilike polovici segmenata daju jednake bodove za kriterije točnosti i tečnosti, a da im se u drugoj polovici procjene razlikuju tek za 1 bod. Unatoč svim definicijama točnosti i tečnosti, jasno se vidi da ispitanicima nije uvijek lako odvojiti te kategorije, a i njihova se kvaliteta uvijek donekle preklapa. Iz skupine B od prosjeka odskače ispitanik koji je samo šest segmenata dao istu ocjenu za točnost i tečnost, a u tri se segmenta njegove procjene razlikuju čak za 3 boda. Taj si je ispitanik za vrednovanje uzeo osjetno dulje vremena pa se može smatrati da je osobito revno pristupio zadatku pred sobom i trudio se donositi odluke prema nekoj određenoj unutarnjoj logici. U primjeru (5) njegova se procjena kvalitete točnosti i tečnosti osobito snažno razlikovala:

(5) *Razina tehnološkog napretka dovela je do konvergencije emitiranja i telekomunikacija.*

Izvornik: „*Das Tempo des technologischen Fortschritts hat zur Konvergenz von Rundfunk und Telekommunikation geführt*."

[6] U povijesti se, uostalom, pokušavala provesti metoda *skupnog* vrednovanja gdje se više vrednovatelja kroz argumentirane rasprave moralo usuglasiti o kvaliteti izlaznih podataka. Zbog nepraktičnosti metode od nje se odustalo „jer zahtijeva da veći broj stručnjaka bude prisutan na jednome mjestu dulje vrijeme te da postignu konsenzus" (Simeon 2008: 94).

Ovaj je vrednovatelj točnosti segmenta dodijelio 2, a tečnosti 5 bodova. Pogreška koja je zaslužna za tako nisku ocjenu točnosti zasigurno je pogrešan leksički odabir za njemačku riječ „Rundfunk“ koja označava „radioteleviziju“, a ne apstraktno „emitiranje“. Pogrešno je prevedena i riječ „Tempo“ koja znači „brzina“, a ne „razina“. Točno se značenje segmenta može zaključiti samo konzultirajući se s izvornikom, stoga je jasno zašto vrednovatelj smatra da je tek „vrlo malo“ značenja preneseno. Ali što je s ocjenom tečnosti? Sveukupno nije mnogo ispitanika tečnosti dalo visoku ocjenu, već su procijenili da je tečnost niže kvalitete kao i točnost. Samo ovaj vrednovatelj smatra da je rečenica „tečna“, a ona zbilja ni ne sadrži niti pravopisne, niti gramatičke niti ikakve druge pogreške za koje se inače smatra da utječu na tečnost. Određeno „zapinjanje“ pri čitanju rezultat je isključivo pogrešnog leksičkog odabira za riječ „Rundfunk“. Ovaj je ispitanik svoje ocjene dodijelio sukladno tome. Ovime se još jednom ističe kako svaki ispitanik vrednovanju pristupa na svoj način i kako se katkad ti pristupi jako razlikuju. S druge strane, primjer (6) pokazuje da nekada pristupi mogu biti i vrlo slični:

(6) *Televizija je naš glavni izvor informacija i zabave.*

Referentni prijevod: „*Televizija je naš primarni izvor informacija i zabave*.“

Izvornik: „*Fernsehen ist unsere Hauptquelle von Information und Unterhaltung*.“

Primjer (6) najkraći je segment i u njemu nema osobitih zamki. Većina se vrednovatelja složila da jednako dobro ispunjava oba kriterija i najčešće su ocjene 5 i 4. Kako ne možemo ustvrditi nikakve pogreške u izlaznim podacima, vjerujemo da se već i s 4 boda izražava određena nespremnost vrednovatelja u to da povjeruju da je sustav zbilja postigao u potpunosti točan i tečan prijevod. Ali, kako objasniti što su dva ispitanika skupine A prijevod smatrala daleko neuspješnijim? Jedan je vrednovatelj segmentu i u kriteriju točnosti i tečnosti dao tek 3 boda, a drugi je točnosti dao 2 boda, a tečnosti samo 1 smatrajući da ovaj segment „nije moguće pratiti“. Izlazni se podaci od referentnog prijevoda razlikuju samo u jednoj riječi („glavni“, a ne „primarni“ izvor), a i tu se radi o bliskoznačnicama, pa je zbilja teško pronaći razloge i opravdati tako niske ocjene. Ovdje se moramo prisjetiti činjenice da je glavni i nepremostivi nedostatak ove metode subjektivnost koja dopušta čak i najrazličitija mišljenja o baš svakom prijevodu.

5.4. *Istraživačko pitanje IV*

Posljednje istraživačko pitanje temeljno je za ovu metodu, a odnosi se na to postoji li ili može li se u kontroliranim uvjetima s dobro odabranim materijalima i ispitani-

cima postići veći sklad među mišljenjima vrednovatelja odnosno veća vjerodostojnost njihove skupne procjene kvalitete. Međutim, u skladu s onime što se već djelomično naslućivalo u literaturi (o poteškoćama u osiguravanju konzistentnosti v. Lavie 2013), a ovim istraživanjem dodatno naglasilo, nije utvrđeno da se unutar jedne ili druge skupine ispitanika zamjećuje veća razina konzistencije. U slučaju da se ispitanike odabirom materijala može potaknuti na određene odgovore, značilo bi to da za određene izlazne podatke postoje i „ispravne" (pr)ocjene, a metoda subjektivne procjene kvalitete tako ne funkcionira. Ona dopušta vrednovateljima da sami prosuđuju i izražavaju svoja stajališta o segmentima, i pritom se ne vrednuju tijekovi misli koji dovode do njihovih prosudbi. Upravo u tome leži vrijednost te metode zbog koje se često zanemaruju brojni njezini nedostaci.

Visoka razina slaganja unutar jedne skupine vrednovatelja, dakle, nije postignuta, ali više od polovine vrednovatelja skupine A jednako je procijenilo kvalitetu 11 segmenata (za obje kategorije), a više od polovine vrednovatelja skupine B dali su istu ocjenu osam segmenata (za točnost) odnosno deset segmenata (za tečnost). Vidimo da je i to djelomično slaganje zapravo niže kod dvojezičnih ispitanika, a ne kod jednojezičnih. Kod jednojezičnih ispitanika također su zabilježena dva segmenta kod kojih je deset, odnosno 11 vrednovatelja razinu točnosti procijenilo jednako, kao u primjeru (7). Deset je vrednovatelja točnosti dalo 4 boda smatrajući da je „većina" značenja prenesena, a po jedan od ostala dva vrednovatelja smatrao je da je preneseno „dosta" odnosno „vrlo malo" značenja. Zanimljivo je da se tu radi o jednom od dužih segmenata, pa je bilo i više prilika za pogreške. No segment se sastoji od nabrajanja i vrednovatelji su uglavnom smatrali da je „većina" ključnih komponenti prisutna u izlaznim podacima (čak iako bi se o točnosti nekih moglo raspravljati).

(7) *Mnoga područja politike EU imaju kulturnu dimenziju, kao što su obrazovanje (uključujući i učenje jezika), znanost i istraživanje, promicanje novih tehnologija i informacijskog društva i socijalnog i regionalnog razvoja.*

Referentni prijevod: „*I brojna druga područja politike EU-a imaju svoju kulturnu dimenziju, kao primjerice obrazovanje (uključujući i učenje stranih jezika), znanstvena istraživanja, razvijanje novih tehnologija i informatičkog društva te socijalni i regionalni razvitak*."

Izvornik: „*Viele Politikbereiche der Union weisen außerdem auch eine kulturelle Dimension auf, beispielsweise Bildung (einschließlich Fremdsprachenerwerb), Wissenschaft und Forschung, Förderung neuer Technologien und der Informationsgesellschaft sowie soziale und regionale Entwicklung*."

Naposljetku, kada se usporede prosječne ocjene za pojedine segmente svakog ispitanika skupina A i B, postaje jasno da velikih razlika među skupinama zapravo nema. Zamijećena je tek tendencija da dvojezični vrednovatelji općenito daju nešto niže ocjene. Samo u šest su segmenata jednojezični vrednovatelji kvalitetu procijenili niže od dvojezičnih. Bodovi se u tim segmentima ne razlikuju mnogo, a samo u jednom segmentu razlika je u gotovo cijelom jednom bodu. Primjeru (8) skupina A u prosjeku je dala 2,55 bodova, a skupina B tek 1,6. Segmentu u glavnoj rečenici nedostaje predikat što stvara probleme u razumljivosti, pa se većina vrednovatelja obiju skupina odlučila izlaznim podacima dati 2 boda i po kriteriju točnost i po kriteriju tečnosti. Iako su istraživanja pokazala da su vrednovatelji skloni davati lošije ocjene tečnosti, a bolje točnosti jer u izlaznim podacima semantika često predstavlja manji problem od formatiranja rečenica (v. Seljan i dr. 2015a), ispitanici su pokazali da im je ova pogreška podjednako utjecala i na točnost i tečnost, pa su obama kriterijima dali niske ocjene. Ostali su vrednovatelji skupine A tendirali izlaznim podacima dati više bodova (3 ili 4), a u skupini B uglavnom su se odlučivali za najmanji mogući broj bodova (1).

(8) *Medijski programi s ciljem da europski audiovizualni sektor budu dinamičniji i konkurentniji od 1990. godine.*

Referentni prijevod: „*Medijski programi, čiji je cilj učiniti europski audiovizualni sektor dinamičnijim i konkurentnijim, provode se od 1990. godine*."

Izvornik: „*Medienprogramme, die darauf abzielen, den audiovisuellen Sektor Europas dynamischer und wettbewerbsfähiger zu machen, laufen seit 1990*."

U ovom se istraživanju prvenstveno pokazalo kako višejezičnost mijenja materijalne uvjete u pripremi istraživanja. Više od same višejezičnosti na prosudbu vrednovatelja utječu čimbenici koji su skriveni i istraživačima, a često i samim vrednovateljima. Kvalitativnom se metodom u ovom istraživanju odgovore na istraživačka pitanja pokušalo pronaći u konkretnim značajkama referentnog prijevoda ili izvornika koje su mogle utjecati na procjenu vrednovatelja. Nekoliko je ispitanika u razgovoru nakon istraživanja izjavilo da imaju zamjerke i na referentni prijevod odnosno nesvjesno su vrednovali i rad ljudskog prevoditelja. Stav vrednovatelja o referentnom prijevodu može se dvojako odraziti na procjenu strojnoga prijevoda: možda će i stroju lakše oprostiti pogreške ako uvide da se one mogu događati i ljudima, a možda im cjelokupno značenje i zbog izlaznih podataka i zbog referentnog prijevoda bude nejasno, pa upravo stoga odluče sustav ocijeniti lošijim. A kao što smo vidjeli u primjeru (3), i najsitnije prevoditeljske odluke mogu promijeniti percepciju o kvalitetnim izlaznim podacima. No, kao što se navodi i u drugim istraživanjima (v. Nübel 1997; Simeon 2008; Yusof i dr. 2017), ti čimbenici mogu biti i

drugi. Utjecati mogu: dosadašnja iskustva vrednovatelja sa strojnim prevođenjem, stavovi o strojnome prevođenju općenito, motivacija za sudjelovanje u istraživanjima ove vrste koja znaju biti zamorna i mukotrpna, pa čak i raspoloženje vrednovatelja u trenutku provođenja istraživanja. Osim niske razine konzistentnosti među različitim vrednovateljima još je veći nedostatak subjektivne metode vrednovanja strojnih prijevoda to što ispitanici ne mogu garantirati ni konzistentnost sa samim sobom: segment koji su jedan dan ocijenili na jedan način, drugi bi dan možda ocijenili drukčije ili bi zbog iste ili slične vrste pogreške lošijom ocjenom ocijenili neki kasniji nego raniji segment i sl.

6. Zaključci

Ovim je istraživanjem preliminarno pokazano kako višejezičnost utječe na provedbu i rezultate istraživanja kvalitete strojnih prijevoda. Prvenstveno je zamijećeno da je dvojezičnim ispitanicima potrebno više vremena za vrednovanje nego jednojezičnim ispitanicima, a da jednojezični ispitanici u prosjeku daju bolje ocjene izlaznim podacima od dvojezičnih ispitanika. Kvalitativnom je analizom vrednovanih segmenata utvrđeno da na percepciju kvalitete strojnog prijevoda kod jednojezičnih vrednovatelja snažno utječe referentni ljudski prijevod. Unutar skupina vrednovatelja nije, međutim, zamijećen veći stupanj konzistencije, što je u skladu s dosadašnjim saznanjima o subjektivnosti kao najvećem nedostatku ovakve metode vrednovanja. Ključ postizanja konsenzusa u istraživanja ovakvog tipa leži u što većem broju ispitanika, a nakon ovog preliminarnog istraživanja valjalo bi donesene zaključke ponoviti na većem testnom skupu i uzorku uz primjenu primjerenih statističkih metoda.

Ako nas kao proizvođača ili korisnika sustava zanima kvaliteta nekog sustava za strojno prevođenje, pri odabiru vrednovatelja uvijek u obzir treba uzeti njihove prednosti i mane. Glavne su prednosti jednojezičnih ispitanika to što se u praksi lakše može okupiti veći broj jednojezičnih ispitanika i što oni mogu vrednovati sve izlazne podatke jednako, ne oviseći o izvornome jeziku pod uvjetom da im je osiguran referentni prijevod. Referentni je prijevod pak glavni nedostatak istraživanja s jednojezičnim ispitanicima jer je njegova kvaliteta uvijek diskutabilna i to otežava provođenje ispitivanja. Glavna prednost dvojezičnih ispitanika leži upravo u tome što nikakva priprema za istraživanje nije potrebna, već oni samo s izvornikom i izlaznim podacima mogu izvršiti traženi zadatak. Glavni nedostatak dvojezičnih ispitanika je to što je teže okupiti toliki broj stručnjaka koji govore određenim jezikom i koji su spremni sudjelovati u vrednovanju.

Zaključno o metodi subjektivne procjene kvalitete može se reći da onome tko traži *stav* nekog stručnjaka o izlaznim podacima, metoda pruža jednako dobar uvid u to neovisno o tome odaberu li se jednojezični ili dvojezični ispitanici. No onaj tko traži *objektivan odgovor* na pitanje jesu li neki izlazni podaci kvalitetni i u kojoj mjeri, možda će ipak morati uvesti neka dodatna ograničenja što se tiče kvalifikacije ispitanika i dosadašnjeg iskustva prevoditelja ili pak posegnuti za nekom posve različitom metodom vrednovanja kvalitete. U kombinaciji sa subjektivnom procjenom mogla bi se provesti i analiza pogrešaka ili se pak ispitanicima može dati test razumijevanja. U budućim istraživanjima bilo bi korisno nadopuniti ovu subjektivnu procjenu objektivnijim testovima razumijevanjima i provjeriti jesu li ove ocjene u skladu s tim rezultatima.

Literatura

Primarni izvori

Repozitorij clarin:el. Dostupno na: https://athena.clarin.gr/resources/download/0e666350523a11e59319aa3fc8d33ad828bf2b7f786b47fc8ffd138c862607aa/. Datum posjeta stranici: 6. siječnja 2019.

Sekundarni izvori

Brkić, Marija; Seljan, Sanja; Matetić, Maja. 2011. Machine translation evaluation for Croatian-English and English-Croatian language pairs. U Sharp, Bernadette; Zock, Michael; Carl, Michael; Jakobsen, Arnt Lykke (ur.), *Proceedings of the 8th International NLPCS Workshop: Human-machine interaction in translation*, 93–104. Kopenhagen: Copenhagen Business School.

Dorr, Bonnie; Snover, Matthew; Madnani, Nitin. 2011. Machine translation evaluation and optimization: Introduction. U Olive, Joseph; Christianson, Caitlin; McCary, John (ur.), *Handbook of natural language processing and machine translation*, 745–747. New York: Springer. http://dx.doi.org/10.1007/978-1-4419-7713-7_5

Han, Aaron Li Feng. 2018. *Machine translation evaluation: A survey.* Dostupno na: https://arxiv.org/pdf/1605.04515.pdf. Datum posjeta stranici: 9. siječnja 2019.

Höhle, Barbara. 2012. *Psycholinguistik*. Berlin: Akademie Verlag GmbH. http://dx.doi.org/10.1524/9783050060149

King, Margaret; Popescu-Belis, Andrei; Hovy. Eduard. 2003. *FEMTI: Creating and using a framework for MT evaluation*. Dostupno na: http://www.mt-archive.info/MTS-2003-King.pdf. Datum posjeta stranici: 9 siječnja 2019.

Koehn, Philipp. 2009. *Statistical machine translation*. Cambridge: Cambridge University Press.

Kučiš, Vlasta; Seljan, Sanja; Klasnić, Ksenija. 2009. Evaluation of electronic translation tools through quality parameters. U Stančić, Hrvoje; Seljan, Sanja; Bawden, David; Lasić-Lazić, Jadranka; Slavić, Aida (ur.), *INFuture2009: Digital resources and knowledge sharing - Proceedings*, 341–351. Zagreb: Odsjek za informacijske znanosti, Filozofski fakultet, Sveučilište u Zagrebu.

Kučiš, Vlasta; Seljan, Sanja. 2014. The role of online translation tools in language education. *Babel* 60(3). 303–324. https://doi.org/10.1075/babel.60.3.03kuc

Lavie, Alon. 2013. *MT evaluation: Human measures and assessment method*s. Dostupno na: http://demo.clab.cs.cmu.edu/sp2013-11731/slides/09.eval1.pdf. Datum posjeta stranici: 30. studenog 2018.

MQM usage guidelines. Dostupno na: http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf. Datum posjeta stranici: 29. studenog 2018.

Nübel, Rita. 1997. End-to-end evaluation in VERBMOBIL I. U Teller, Virginia; Sundheim, Beth (ur.), *Proceedings of MT Summit VI – Machine translation: past, present, future*, 232–239. San Diego: Association for Machine Translation in the Americas. Dostupno na: http://www.mt-archive.info/MTS-1997-Nubel.pdf. Datum posjeta stranici: 9. siječnja 2019.

Olive, Joseph; Christianson, Caitlin; McCary, John (ur). 2011. *Handbook of natural language processing and machine translation*. New York: Springer. http://dx.doi.org/10.1007/978-1-4419-7713-7

Sanders, Gregory; Przybocki, Mark; Madnani, Nitin; Snover, Matt. 2011. Machine translation evaluation and optimization: Human subjective judgments. U Olive, Joseph; Christianson, Caitlin; McCary, John (ur.), *Handbook of natural language processing and machine translation*, 750–758. New York: Springer. https://doi.org/10.1007/978-1-4419-7713-7_5

Seljan, Sanja; Brkić, Marija; Kučiš, Vlasta. 2011. Evaluation of Free Online Machine Translations for Croatian-English and English-Croatian Language Pairs. U Billenness, Clive; Hemera, Annette; Mateljan, Vladimir; Banek Zorica, Mihaela; Stančić, Hrvoje; Seljan, Sanja (ur.), *INFuture2011: The future of information sciences – information sciences and e-society*, 331–344. Zagreb: Odsjek za informacijske znanosti, Filozofski fakultet, Sveučilište u Zagrebu.

Seljan, Sanja; Vičić, Tomislav; Brkić, Marija. 2012. BLEU evaluation of machine-translated English-Croatian legislation. U Calzolari, Nicoletta; Choukri, Khalid; Declerck, Thierry; Uğur Doğan, Mehmet; Maegaard, Bente; Mariani, Joseph; Moreno, Asuncion; Odijk, Jan; Piperidis, Stelios (ur.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2143–2148. Istanbul: ELRA. Dostupno na: https://www.bib.irb.hr/582714. Datum posjeta stranici: 9. siječnja 2019.

Seljan, Sanja; Dunđer, Ivan. 2014. Combined automatic speech recognition and machine translation in business correspondence domain for English-Croatian. *International*

Journal of Computer, Information, Systems and Control Engineering 8(11). 1069–1075.

Seljan, Sanja; Tucaković, Marko; Dunđer, Ivan. 2015a. Human evaluation of online machine translation services for English/Russian-Croatian. U Rocha, Álvaro; Correia, Ana Maria; Costanzo, Sandra; Reis, Luís Paulo (ur.), *Advances in intelligent systems and computing – New contributions in information systems and technologies*, 1089–1098. Cham: Springer International Publishing Switzerland. http://dx.doi.org/10.1007/978-3-319-16486-1_108

Seljan, Sanja; Klasnić, Ksenija; Stojanac, Mara; Pešorda, Barbara; Mikelić Preradović, Nives. 2015b. Information transfer through online summarizing and translation technology. U Anderson, Karen; Duranti, Luciana; Jaworski, Rafał; Stančić, Hrvoje; Seljan, Sanja; Mateljan, Vladimir (ur.), *INFuture2015: E-institutions – openness, accessibility, and preservation.* 197–210. Zagreb: Odsjek za informacijske znanosti, Filozofski fakultet, Sveučilište u Zagrebu. http://dx.doi.org/10.17234/INFUTURE.2015.24

Simeon, Ivana. 2008. *Vrednovanje strojnoga prevođenja*. Neobjavljena doktorska disertacija. Zagreb: Sveučilište u Zagrebu.

Snover, Matt; Madnani, Nitin; Dorr, Bonnie; Schwartz, Richard. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. U Callison-Burch, Chris; Koehn, Philipp; Monz, Christof; Schroeder, Josh (ur.), *Proceedings of the Fourth Workshop on Statistical Machine Translation*, 259–268. Atena: Association for Computational Linguistics. https://doi.org/10.3115/1626431.1626480

Yusof, Norwati MD; Darus, Saadiyah; Aziz Mohd Juzaiddin AB. 2017. Evaluating intelligibility in human translation and machine translation. *3L: The Southeast Asian Journal of English Language Studies* 23(4). 251–264. http://doi.org/10.17576/3L-2017-2304-19

Zajednički europski referentni okvir za jezike. Dostupno na: https://europass.cedefop.europa.eu/sites/default/files/cefr-hr.pdf. Datum posjeta stranici: 8. siječnja 2019.

Dodatak A

Testni skup na izvornom jeziku

(1) Sprache, Literatur, bildende Kunst, Malerei, Architektur, Handwerk, Film und Fernsehen tragen alle zu Europas kultureller Vielfalt bei.

(2) Das Ziel der EU ist ein doppeltes: diese Vielfalt zu erhalten und zu fördern und dazu beizutragen, sie anderen zugänglich zu machen.

(3) Die Kulturindustrien in der EU – Film und audiovisuelle Medien, Verlagswesen, Musik und Kunsthandwerk – sind wichtige Quellen für Einnahmen und Arbeitsplätze, beschäftigen sie doch etwa sieben Millionen Menschen.

(4) Die Union hat diesem Sektor gegenüber wirtschaftliche Verantwortung und sie ist bestrebt, die richtigen Bedingungen zu schaffen, damit europäische Industrien international wettbewerbsfähig sind.

(5) So unterhält die EU Unterstützungsprogramme für bestimmte Kulturindustrien und ermutigt sie, die Gelegenheiten zu ergreifen, die der Binnenmarkt und digitale Technologien bieten.

(6) Sie ist auch bestrebt, ein dynamisches Umfeld für diese Industrien zu schaffen, indem sie für Bürokratieabbau sorgt, den Zugang zu Finanzierungen erleichtert, bei Forschungsprojekten behilflich ist und zu mehr Zusammenarbeit zwischen Partnern innerhalb und außerhalb der Union ermutigt.

(7) Viele Politikbereiche der Union weisen außerdem auch eine kulturelle Dimension auf, beispielsweise Bildung (einschließlich Fremdsprachenerwerb), Wissenschaft und Forschung, Förderung neuer Technologien und der Informationsgesellschaft sowie soziale und regionale Entwicklung.

(8) So unterstützt der Europäische Sozialfonds beispielsweise die italienische Arturo-Toscanini-Stiftung, die seit Mitte der 90er Jahre Kurse für arbeitslose Musiker abhält.

(9) Die konkreten Ziele des derzeitigen Kulturprogramms sind Sensibilisierung für kulturelle Güter, die für Europa von Bedeutung sind, sowie deren Erhalt, die Förderung der grenzüberschreitenden Mobilität der Kulturschaffenden, die Unterstützung der Verbreitung von künstlerischen und kulturellen Werken und Erzeugnissen sowie die Förderung des interkulturellen Dialogs.

(10) Medienprogramme, die darauf abzielen, den audiovisuellen Sektor Europas dynamischer und wettbewerbsfähiger zu machen, laufen seit 1990.

(11) Jedes Jahr werden eine oder zwei Städte als europäische Kulturhauptstädte ausgewählt, die sich dadurch für finanzielle Unterstützung qualifizieren.

(12) Sprachliche Vielfalt ist ein kultureller und demokratischer Eckpfeiler der Europäischen Union

(13) Schätzungsweise 40 Millionen Menschen in der EU sprechen eine Muttersprache, die nicht die offizielle Sprache ihres Herkunftslandes ist.

(14) Das Ziel der gemeinsamen Landwirtschaftspolitik ist es, Bauern einen vernünftigen Lebensstandard und Verbrauchern Qualitätslebensmittel zu fairen Preisen zu ermöglichen, sowie unser bäuerliches Erbe zu bewahren.

(15) Fernsehen ist unsere Hauptquelle von Information und Unterhaltung.

(16) Wir alle verbringen durchschnittlich bis zu drei Stunden damit, Nachrichten, Sport, Filme und andere Programme anzuschauen.

(17) Jede nationale Regierung betreibt ihre eigene Audiovisionspolitik, während die Union Regeln und Richtlinien setzt, soweit gemeinsame Interessen berührt sind, wie offene EU-Grenzen und fairer Wettbewerb.

(18) Um ihre eigene kulturelle Vielfalt zu schützen und lokale Produktionen zu fördern, hat die EU sich mit Erfolg bei der Welthandelsorganisation um die sogenannte ‚kulturelle Ausnahmeregelung' bemüht.

(19) Das Tempo des technologischen Fortschritts hat zur Konvergenz von Rundfunk und Telekommunikation geführt.

(20) Offene Grenzen und erschwingliches Reisen haben den Europäern bisher einmaligen Grad an persönlicher Mobilität beschert.

(21) Güter werden schnell und effizient von der Fabrik zum Kunden – oft in verschiedenen Ländern – gesendet.

Dodatak B

Referentni prijevod testnog skupa

(1) Jezik, književnost, likovna umjetnost, slikarstvo, arhitektura, ručni radovi, kinematografija te radio i televizija pridonose europskoj kulturnoj raznolikosti.

(2) Cilj je Europske unije dvojak: očuvati i poduprijeti tu raznolikost te pridonijeti da ona bude dostupna i drugima.

(3) Kulturne industrije u EU – kinematografija i audiovizualni mediji, nakladništvo, glazba i umjetnički zanati – važni su izvori prihoda i radnih mjesta jer ipak zapošljavaju oko sedam milijuna ljudi.

(4) Europska unija ima gospodarsku odgovornost prema ovome sektoru te nastoji osigurati prikladne uvjete kako bi europske industrije mogle biti i međunarodno konkurentne.

(5) Stoga EU provodi programe za potporu određenih kulturnih industrija i potiče ih da iskoriste prilike koje im nude jedinstveno tržište i digitalne tehnologije.

(6) Također nastoji stvoriti dinamično okruženje za ove industrije smanjujući birokraciju, omogućujući lakši pristup sredstvima financiranja, pomažući u istraživačkim projektima te potičući veću suradnju s partnerima unutar i izvan Unije.

(7) I brojna druga područja politike EU-a imaju svoju kulturnu dimenziju, kao primjerice obrazovanje (uključujući i učenje stranih jezika), znanstvena istraživanja, razvijanje novih tehnologija i informatičkog društva te socijalni i regionalni razvitak.

(8) Europski socijalni fond, primjerice, podupire talijansku Zakladu 'Arturo Toscanini' koja od sredine 90-ih godina održava tečajeve za glazbenike bez posla.

(9) Konkretni su ciljevi trenutnoga kulturnog programa promicanje svjesnosti o kulturnim dobrima od europskoga značaja i njihovo očuvanje, promicanje nadnacionalne mobilnosti djelatnika u kulturi, poticanje širenja umjetničkih i kulturnih djela i tekovina te poticanje međukulturnoga dijaloga.

(10) Medijski programi, čiji je cilj učiniti europski audiovizualni sektor dinamičnijim i konkurentnijim, provode se od 1990. godine.

(11) Svake se godine jedan ili dva grada proglase kulturnim prijestolnicama Europe, čime se kvalificiraju za financijsku potporu.

(12) Jezična raznolikost kulturni je i demokratski kamen temeljac Europske unije.

(13) Procjenjuje se da više od 40 milijuna ljudi u EU govori materinskim jezikom koji nije službeni jezik njihove zemlje podrijetla.

(14) Cilj zajedničke poljoprivredne politike jest poljoprivrednicima osigurati pristojan životni standard, potrošačima osigurati kvalitetne prehrambene proizvode po pravednim cijenama i očuvati naše ruralno naslijeđe.

(15) Televizija je naš primarni izvor informacija i zabave.

(16) Svatko od nas u prosjeku provodi do tri sata dnevno gledajući vijesti, sport, filmove i druge sadržaje.

(17) Svaka nacionalna vlada provodi svoju audiovizualnu politiku, a Unija donosi pravila i smjernice kada su u pitanju zajednički interesi poput otvorenih granica EU-a i pravedne konkurencije.

(18) Kako bi zaštitila vlastitu kulturnu raznolikosti i promicala lokalnu proizvodnju, EU je od Svjetske trgovinske organizacije zatražila i dobila takozvanu 'kulturnu iznimku'.

(19) Tempo tehnološkog napretka doveo je do konvergencije radiotelevizije i telekomunikacija.

(20) Otvorene granice i cjenovno prihvatljiva putovanja omogućili su Europljanima dosad neviđenu razinu osobne mobilnosti.

(21) Roba se brzo i učinkovito prevozi iz tvornice do potrošača – koji su često u drugim zemljama.

Dodatak C

Strojni prijevod testnog skupa

(1) Jezik književnosti, likovne umjetnosti, slikanje, arhitektura, obrti, film i televiziju sve doprinose europske kulturne raznolikosti.

(2) Cilj Europske unije je dvojak: očuvanje i promicanje te raznolikosti te olakšavanje pristupa ostalima.

(3) Kulturne industrije u Europskoj uniji - Film i audiovizualnih, izdavaštvo, glazba i obrti - važan su izvor prihoda i radnih mjesta, oni su i dalje zapošljavaju oko sedam milijuna ljudi.

(4) Unija ima ekonomsku odgovornost za ovaj sektor i nastoji stvoriti prave uvjete za europske industrije da se natječu na međunarodnoj razini.

(5) EU održava programe podrške za određene kulturne industrije, potičući ih da iskoriste mogućnosti koje nudi unutarnje tržište i digitalne tehnologije.

(6) Također se zalaže za stvaranje dinamično okruženje za te industrije pruža rezanja birokracije, lakši pristup financijama, pomoći u istraživačkim projektima i potiče veću suradnju između partnera unutar i izvan Unije.

(7) Mnoga područja politike EU imaju kulturnu dimenziju, kao što su obrazovanje (uključujući i učenje jezika), znanost i istraživanje, promicanje novih tehnologija i informacijskog društva i socijalnog i regionalnog razvoja.

(8) Europski socijalni fond, na primjer, uz potporu talijanske Arturo Toscanini Foundation,

koji drži tečajeve za nezaposlene glazbenika od sredine 90-ih godina.

(9) Specifični ciljevi trenutnog programa Kultura su podizanje svijesti o kulturnim dobrima koja su od europske važnosti i njihovo očuvanje, promicanje prekogranične mobilnosti kulturnih djelatnika, potpora cirkulaciju kulturnih umjetničkih djela i proizvoda te promicanje međukulturnog dijaloga.

(10) Medijski programi s ciljem da europski audiovizualni sektor budu dinamičniji i konkurentniji od 1990. godine.

(11) Svake godine jedan ili dva grada su odabrani kao europski kulturni prijestolnici, koji se kvalificiraju za financijsku potporu.

(12) Jezična raznolikost kulturni i demokratski temelj Europske unije

(13) Procjenjuje se da 40 milijuna ljudi u EU govori materinski jezik različit od službenih jezika svoje zemlje podrijetla.

(14) Cilj zajedničke poljoprivredne politike je omogućiti poljoprivrednicima razuman životni standard i kvalitetnog potrošači hrane po vrlo povoljnim cijenama i očuvati naš rustikalni baštine.

(15) Televizija je naš glavni izvor informacija i zabave.

(16) Svi trošimo u prosjeku do tri sata gledajući vijesti, sport, filmove i druge programe.

(17) Svaka nacionalna vlada ima svoju vlastitu audiovizualnu politiku, dok Unija određuje pravila i smjernice prilikom zajednički interesi koji su uključeni, kao i otvorene granice EU-a i poštenog natjecanja.

(18) Da bi zaštitili svoju kulturnu različitost i promicanje lokalne produkcije, EU je pokušao s uspjehom u Svjetskoj trgovinskoj organizaciji o takozvanoj „kulturne iznimke".

(19) Razina tehnološkog napretka dovela je do konvergencije emitiranja i telekomunikacija.

(20) Otvorene granice i pristupačna putovanja dali su Europljane neviđenu razinu osobne mobilnosti.

(21) Roba se šalje brzo i učinkovito od tvornice do kupca, često u različitim zemljama.

Adresa autorice:

Sveučilište u Zadru
Ulica Mihovila Pavlinovića, 2300 Zadar
E-mail: sljubas@student.unizd.hr

JEZIKOSLOVLJE
21.2 (2020): 207-235

THE IMPACT OF RESPONDENTS' MULTILINGUALISM ON HUMAN EVALUATION OF MACHINE TRANSLATION QUALITY

This paper presents a study of the impact of multilingualism on the subjective method of evaluating machine translation quality. The subjectivity of this method is usually manifested in the low level of inter-coder agreement. In this preliminary research, two groups of human judges, the first comprised of monolingual and the second of bilingual respondents, evaluated the accuracy and fluency of the same set of machine-translated text segments. The segments have been translated with Google Translate. The monolingual respondents compared the MT-generated output with a human translation and the bilingual respondents with the original text. The aim of this study was to determine how the discrepancies between monolingual and bilingual respondents shape evaluation patterns with respect to the length of the evaluation process, potential deviations in the median evaluation score and the analysis of causes influencing these evaluation discrepancies. The qualitative analysis has shown that bilingual respondents in general give lower scores to output data, but need more time to complete the evaluation process than monolinguals. However, no tendency towards a higher level of inter-coder agreement has been noted in either group of human judges.

Key words: evaluation; machine translation; multilingualism; subjective evaluation.