

Merjenje berljivosti strojnih prevodov s sledilcem očesnih gibov

Kristijan Armeni,* Grega Repovš,† Špela Vintar‡

* Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
kristijan.armeni@gmail.com

† Oddelek za psihologijo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
grega.repovs@psy.ff.uni-lj.si

‡ Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
spela.vintar@ff.uni-lj.si

Povzetek

V prispevku predstavimo študijo, s katero smo želeli preveriti uporabnost sledilca očesnih gibov kot dopolnilno metodo za evalvacijo strojnih prevodov iz angleščine v slovenščino. Najprej predstavimo priložnostno kategorizacijo napak v strojnih prevodih, na podlagi katere smo pripravili tri vrste prevodov (popravljen, delno popravljen in nepopravljen). Med tem ko so udeleženci brali besedila za razumevanje, smo posneli gibanje oči. Po branju je vsak udeleženec podal še subjektivne ocene prebranega besedila. Udeleženci so sicer vse tri tipe prevodov ocenili kot kvalitativno različne, rezultati sledilca pa pokažejo razlike le med branjem povsem popravljenih in nepopravljenih prevodov, ne pokažejo pa razlik med branjem delno popravljenih in nepopravljenih prevodov.

Abstract: Eye tracking measures as indicators of readability for machine-translated texts

In the present paper, we present a study where we tested the newly proposed methodology for evaluating machine-translated texts (English-Slovene language pair). We first outline our ad hoc error categorization scheme, which served as the basis for the preparation of three types of translations (raw MT output, partially corrected, and completely corrected). Participants were instructed to read the texts for comprehension while their eye movements were recorded. After reading, they were asked to give their subjective evaluations as well. While the three types of translations were given significantly different scores by the participants, eye-tracking measures show difference between completely corrected and non-corrected MT output, but no difference between partially and non-corrected MT output.

1. Uvod

Namen študije, predstavljene v tem prispevku, je bil test sledilca očesnih gibov kot metodologije za evalvacijo strojnih prevodov (jezikovni par angleščina-slovenščina).¹ Sledilec očesnih gibov je v psiholingvisti standardna metodologija za raziskovanje kognitivnih procesov med branjem, z razvojem tehnologije pa se uporaba prenaša na sorodna področja. Množica empiričnih dognanj raziskav tekom zadnjih nekaj desetletij kaže, da je gibanje oči med branjem tesno povezano s sočasnimi kognitivnimi procesi pri bralcih (ang. *online cognitive processes*, Rayner, 1998; Just et al., 1980).

Osnovna hipoteza te študije je, da predstavljajo manj berljiva besedila, ki vsebujejo veliko napak (npr. zelo slab strojni prevod), za bralca večje kognitivno breme kot bolj berljiva besedila brez napak. Če obstaja povezava med kognitivnimi procesi ter gibanjem oči med branjem, potem bi morale specifične mere sledilca očesnih gibov (npr. število fiksacij, čas branja, povprečno trajanje fiksacije) odražati razlike v berljivosti besedil. Ker na ta način posnamemo branje strojnih prevodov v realnem času, bi takšne mere lahko predstavljale zanimivo, dodatno informacijo o kvaliteti besedil.

2. Evalvacija strojnih prevodov

Na področju strojnih prevajalnikov je evalvacija sestavni del razvojnega cikla posameznega sistema.

Obstaja več metod evalvacije strojnih prevodov, pri čemer ima vsaka svoje prednosti in pomanjkljivosti. Metode za evalvacijo strojnih prevodov se v grobem delijo v dve skupini: ročne in avtomatske.

Metoda ročne evalvacije, tj. mnenje človeških ocenjevalcev oz. končnih uporabnikov, je tista, ki velja za referenčno in ki najbolje odseva kvaliteto strojnega prevoda. Zelo pogosto omenjeni pomanjkljivosti ročnih evalvacij sta cena ter zamudnost (Papineni et al., 2002; Callison-Burch et al., 2007; Koehn, 2010; Verdonik in Sepesy Maučec, 2013); obsežne evalvacijske delavnice z veliko udeleženci namreč predstavljajo nezanemarljive stroške. V zadnjem času se zato evalvacije opravlja tudi z množičnim zbiranjem podatkov preko spletnih platform (Graham et al., 2013b).

Druga velika težava človeških ocen je neskladnost oziroma t. i. stopnja (ne)strinjanja med posameznimi ocenjevalci (ang. *inter-annotator agreement*). Do neskladnosti pride tudi pri vsakem posamezniku, zato se dodatno preverja še ocenjevalčevo konsistentnost (ang. *intra-annotator agreement*). Pri ocenjevanju se najpogosteje uporabljajo pet in sedem stopenjske lestvice (Callison-Burch et al., 2007), pri čemer se v praksi izkaže, da imamo ljudje zelo različne kriterije in smo lahko različno strogi. Pojavi se torej dilema, ali ne povedo takšni rezultati več o samih ocenjevalcih kot pa o besedilih. Vrednost tako dobljenih rezultatov je lahko vprašljiva, posledično pa tudi sami zaključki, do katerih naj bi vodili (Koehn, 2010; Graham et al. 2013a).

Glavni prednosti avtomatske metode sta hitrost in cenovna dostopnost, vendar pa se avtomatske metrike se še vedno smatrajo le za nepopolno nadomestilo človeške

¹ Vsebina tega prispevka povzema ideje in rezultate, predstavljene v magistrski nalogi (2014) z istim naslovom.

evalvacije in bodo namreč označene za učinkovite le, če bodo podale rezultate, ki visoko korelirajo z rezultati človeške evalvacije (Callison-Burch et al., 2007; Verdonik in Sepesy Maučec, 2013; Graham et al., 2013b).

Avtomatske metrike imajo še druge pomanjkljivosti, med drugim predvsem to, da je končni rezultat takšne evalvacije zgolj specifična številčna vrednost, ki odseva več različnih parametrov. Ker je v postopke računanja končne vrednosti vključenih več faktorjev, pomen posamezne številke ni jasen (Koehn, 2010). Poleg tega je pri takšnih avtomatskih metrikih zelo pomembno, kakšen je referenčni prevod (ali več prevodov), ki ga vključimo v evalvacijsko gradivo (Verdonik in Sepesy Maučec, 2013).

V znanstveni literaturi o evalvaciji tako novosti in predlogi za izboljšave obstoječih tehnik niso redkost. Nedavno je v raziskovalni skupnosti prišlo do ideje, da bi kot vir podatkov o berljivosti strojnih prevodov in tako posredno tudi kot orodje za evalvacijo lahko uporabili sledilec očesnih gibov.

3. Predhodne študije

Za izhodišče smo vzeli dve nedavni študiji, kjer so pokazali, da bi sledilec očesnih gibov lahko uporabili kot dopolnilo standardnim tehnikam za evalvacijo strojnih prevodov, kot so npr. avtomatske metrike in človeške ocene. Dohertyja in soavtorje (2010) je zanimalo osnovno vprašanje, ali bo na podlagi mer sledilca očesnih gibov možno razlikovati med slabimi strojnimi prevodi in dobrimi strojnimi prevodi. Udeleženci v eksperimentu so brali predhodno ocenjene strojno prevedene stavke, ki so bili označeni kot »dobri« oziroma »slabi«. Rezultati so pokazali, da je bilo število fiksacij višje in povprečen čas branja daljši za slabo ocenjene kot za dobro ocenjene strojne prevode, medtem ko pri povprečnem trajanju fiksacije in razširjenju zenice ni prišlo do razlik. Kvaliteta posameznih stavkov je bila zmerno negativno korelirana z rezultati sledilca očesnih gibov.

Stymne in soavtorji (2012) so idejo nadgradili in pripravili nekoliko drugačen eksperiment. Zanimalo jih je, ali bi lahko sledilec očesnih gibov uporabili za analizo napak v strojnih prevodih. Udeleženci so brali cela besedila, ki so bila strojno prevedena s tremi različnimi statističnimi prevajalniki. V vsakem besedilu so označili dele z napakami in dele besedila, kjer napak ni bilo. Rezultati so pokazali razlike v številu fiksacij in povprečnem času pogleda (ang. *average gaze time*) za dele besedila, kjer so bile napake in kjer jih ni bilo. Do sistematičnih razlik (povprečen čas pogleda) je prišlo tudi med napakami pri posameznih strojnih prevajalnikih, prav tako tudi pri posameznih tipih napak, pri čemer je branje najbolj otežilo napačen vrstni red besed.

Naša študija gradi na obeh opisanih raziskavah, a se od obeh tudi razlikuje. Podobno kot je pojasnjeno v obeh študijah, tudi našo v osnovi razumemo kot metodološko. V prvi študiji (Doherty et al., 2010) so udeleženci brali strojno prevedene stavke v dveh skupinah, medtem ko smo v našem primeru uporabili cela besedila in dodaten tip besedila med »dobrim« in »slabim«. S tega vidika je naša raziskava bolj primerljiva z drugo študijo (Stymne et al., 2012), kjer so uporabili cela besedila in primerjali napake v prevodih treh strojnih prevajalnikov. Po branju so podobno zbrali ocene udeležencev, a niso uporabili kontinuiranih lestvic kot v našem primeru. Rezultate

sledilca so nato primerjali z rezultati avtomatskih metrik, česar sami nismo vključili v zasnovo.

4. Kategorizacija napak

V prvi fazi smo opravili priložnostno kategorizacijo napak v strojnih prevodih za jezikovni par angleščina-slovenščina. V naš profil napak smo vključili naslednje kategorije: ujemalne napake, pomenske napake, stilistične napake, pravopisne napake, napačen besedni red, izpuščene in vrinjene besede ter neprevedene besede.

Največ primerov napak smo uvrstili v kategorijo t.i. ujemalnih napak, kar je do neke mere pričakovano, saj je slovenščina v primerjavi z angleščino morfološko precej bolj razčlenjena. Pri jezikovnih parih, kjer je eden od jezikov morfološko bogat, pride v primeru statističnih prevajalnikov do »problema redkih podatkov« (ang. *sparse data problem*, Koehn 2010). Tabela 1 povzema pojavnost posameznih tipov napak za tri besedila iz preizkušnje.

	besedilo 1		besedilo 2		besedilo 3	
ujemalne napake	18	45 %	22	51 %	21	40 %
pomen	3	7,5 %	4	9,3 %	5	9,6 %
izpust	4	10 %	3	7 %	1	1,9 %
vrinjena beseda	1	2,5 %	2	4,7 %	5	9,6 %
besedni red	4	10 %	2	4,7 %	7	13,5 %
stilistika	5	12,5 %	6	14 %	8	15,4 %
neprevedeno	0	0 %	0	0 %	0	0 %
pravopis	5	12,5 %	4	9,3 %	5	9,6 %
skupaj	40	100 %	43	100 %	52	100 %

Tabela 1: Pojavnost posameznih tipov napak

Za našo študijo je kategorija ujemalnih napak ključnega pomena. Če v besedilih popravimo samo ujemalne napake, popravimo glede na vrednosti v zgornji tabeli v povprečju približno polovico vseh napak. Predstavljene številke velja interpretirati s previdnostjo; zaradi majhnega vzorca, specifičnega tipa besedil in enega samega ocenjevalca na podlagi teh podatkov ni moč posploševati o kvaliteti prevajalnika. Menimo pa, da je kategorija ujemalnih napak ustrezen kriterij za sistematično pripravo besedil. Besedilo brez ujemalnih napak bi se moralo po berljivosti uvrstiti med nepopravljeno (najmanj berljivo) besedilo in popolnoma popravljeno (najbolj berljivo) besedilo.

5. Metoda

5.1. Udeleženci

V preizkušnji je sodelovalo 33 udeležencev, za končno analizo pa smo uporabili rezultate 31 udeležencev (26 žensk). Povprečna starost je bila 20,2 leti ($SD = 2,4$), najnižja 18 let, najvišja pa 27 let. Udeleženci so bili dodiplomski oziroma podiplomski študenti na Univerzi v Ljubljani in so bili materni govorci slovenščine. Vsi so imeli normalen vid ali pa so pri branju uporabljali korekcijo vida (leče ali očala).

5.2. Oprema

Celotna preizkušnja je bila vodena računalniško. Za snemanje smo uporabljali sistem EyeLink 1000 (frekvenca zbiranja podatkov: 1000 Hz) v psihološkem laboratoriju Filozofske fakultete Univerze v Ljubljani. Da bi dosegli čim boljšo kvaliteto podatkov, so imeli med branjem udeleženci glavo naslonjeno na posebnem stojalu, ki je bilo od zaslona oddaljeno približno 60 cm. V vseh primerih smo merili gibanje desnega očesa. Uporabljali smo 9-točkovni kalibracijski model.

5.3. Besedila

Povprečna dolžina vseh besedil, uporabljenih v preizkušnji je znašala 223 besed ($SD = 7,2$) oziroma 1200 znakov brez presledkov ($SD = 11,1$). Vsa, razen enega besedila v vaji, so bila preoblikovana v dva odstavka (pisava Calibri, velikost 12, 1,5 medvrstični razmik), naslov je bil vedno v odebeljenem tisku.

Vsa besedila, razen enega v vaji, smo dobili na spletni strani Evropske komisije. Glavni razlog za takšno odločitev ni vsebinske, pač pa v prvi vrsti predvsem pragmatične narave; vse angleške novice na spletni strani so namreč prevedene tudi v slovenščino. V primeru zagat pri pripravljanju oziroma popravljanju besedilnih različic za preizkušnjo smo imeli tako na voljo ustrezno referenco. Poleg tega se vse novice na strani Evropske komisije dotikajo Evropske unije, s čimer smo zagotovili vsebinsko primerljivost vsaj na najvišji besedilni ravni.

5.4. Preizkusne različice

Za vsako od treh besedil, uporabljenih v preizkušnji, smo pripravili tri različice, ki so v našem primeru ustrezale trem eksperimentalnim pogojem. V izhodišču smo angleška besedila samo prevedli s pomočjo Googlevega prevajalnika. To različico smo označili z NP (nepopravljena različica). V naslednjem koraku smo v vseh strojnih prevodih popravili zgolj ujemanje napake, s čimer smo ustvarili delno popravljene različice, ki smo jih označili s P1 (popravljena različica 1). Nazadnje smo v vseh besedilih popravili še preostale napake in tako dobili popolnoma popravljeno besedilo. Te smo označili s P2 (popravljena različica 2). Na koncu smo dobili skupno 9 možnih kombinacij besedila (3) in tipa prevoda (3). Vsi predstavljeni rezultati za posamezen tip prevoda so nato združeni preko različnih besedil.

5.5. Potek preizkušnje

Vsem udeležencem je bilo podano enako navodilo, da naj berejo za razumevanje in na način, ki jim najbolj ustreza. Posebej je bilo poudarjeno, da si podatkov v besedilih ni potrebno dodatno zapomniti ter da se v nadaljevanju preizkušnje uporabljajo vprašalniki ne nanašajo na poznavanje vsebine prebranega. Ko so prebrali navodila, so udeleženci s pritiskom na odgovorno tablico sprožili prikaz besedila in začetek snemanja očesnih gibov. Konec branja so prav tako označili s pritiskom na gumb, s čimer se je zaključilo tudi snemanje. Pri branju ni bilo časovne omejitve. Posamezna preizkušnja, ki je zajemala branje petih besedil in odgovarjanje na vprašanja, je praviloma trajala med 20 in 30 minut.

5.6. Subjektivne ocene in verbalna poročila

Poleg mer sledilca očesnih gibov smo želeli med preizkušnjo pridobiti tudi subjektivno oceno branja vsakega posameznika. Po vsakem prebranem besedilu so udeleženci z uporabo kontinuirane lestvice (z miško so morali klikniti na ustrezni del prikazane lestvice oziroma skale) odgovorili na tri vprašanja o poteku branja, zahtevnosti razumevanja in vračanju na že prebrane dele besedila oziroma t. i. »regresijah«.

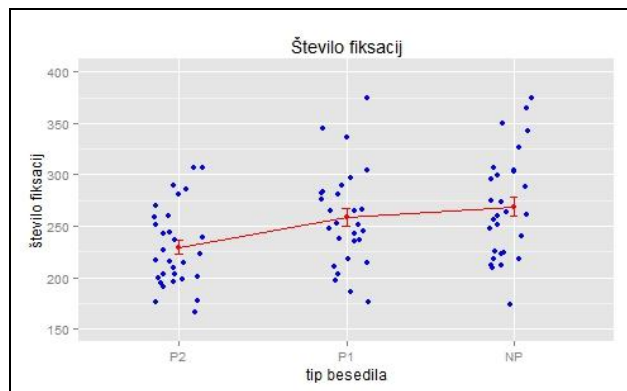
Po končanem branju smo z vsakim udeležencem opravili še krajši pogovor, da bi pridobili čim več dodatnih informacij, ki bi lahko koristile pri interpretaciji rezultatov sledilca očesnih gibov in subjektivnih ocen. Zastavili smo jim naslednja vprašanja: a) katero besedilo je bilo za branje najtežje in katero najlažje, b) ali bi bili sposobni povzeti vsebino prebranega, c) ali so med branjem spremenili bralno strategijo in d) ali obstaja kakšen tip napake, ki so si ga posebej zapomnili.

6. Rezultati

6.1. Sledilec očesnih gibov

6.1.1. Število fiksacij

Število fiksacij predstavlja število vseh fiksacij, narejenih med branjem posameznega besedila. Fiksacija je bila po vzoru predhodnih študij definirana kot vsak postanek, daljši od 100 ms.

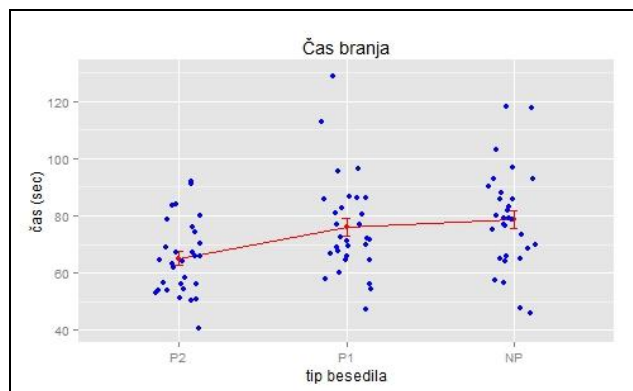


Slika 1: Število fiksacij

Enosmerna analiza variance (ANOVA) pokaže, da so razlike statistično pomembne glede na tip prevoda, $F(2, 60) = 16,9$, $p < 0,05$, in v manjši meri tudi glede na posamezno besedilo, $F(2, 60) = 3,4$, $p < 0,05$. Ločeni t-test pokaže statistično pomembne razlike med popolnoma popravljeno in nepopravljeno vrsto prevoda, $t(30) = 5,3$, $p < 0,05$. Med delno popravljeno in nepopravljeno različico ni bilo statistično pomembnih razlik, $t(30) = 1,2$, $p > 0,05$.

6.1.2. Čas branja

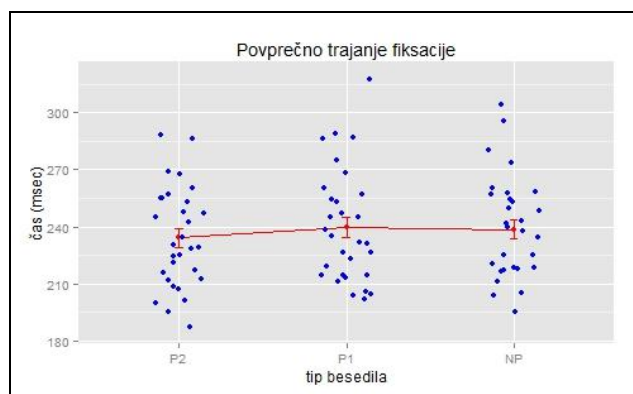
Čas branja je seštevek trajanja vseh fiksacij, narejenih med branjem besedil.



Slika 2: Čas branja

Enosmerna analiza variance (ANOVA) pokaže, da so razlike statistično pomembne glede na tip prevoda, $F(2, 60) = 18,4$, $p < 0,05$, in v manjši meri tudi glede na posamezno besedilo, $F(2, 60) = 3,1$, $p < 0,05$. Ločeni t-test pokaže statistično pomembne razlike med popolnoma popravljeno in nepopravljeno vrsto prevoda, $t(30) = 6$, $p < 0,05$. Med delno popravljeno in nepopravljeno različico ni bilo statistično pomembnih razlik, $t(30) = 1,2$, $p > 0,05$.

6.1.3. Povprečno trajanje fiksacije



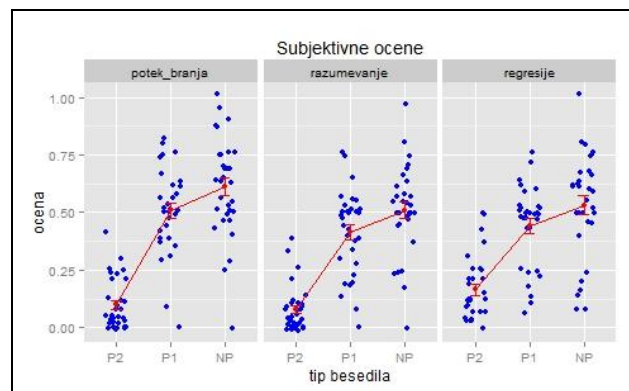
Slika 3: Povprečno trajanje fiksacije

Enosmerna analiza variance (ANOVA) pokaže, da so razlike statistično pomembne glede na tip prevoda, $F(2, 60) = 4,7$, $p < 0,05$. Glede na posamezno besedilo razlike niso bile statistično pomembne, $F(2, 60) = 0,6$, $p > 0,05$. Ločeni t-test pokaže statistično pomembne razlike med popolnoma popravljeno in nepopravljeno vrsto prevoda, $t(30) = 2,6$, $p < 0,05$. Med delno popravljeno in nepopravljeno različico ni bilo statistično pomembnih razlik, $t(30) = 0,7$, $p > 0,05$.

6.2. Subjektivne ocene

Odgovori, zbrani s kontinuiranimi lestvicami, so predstavljeni kot proporci lestvice, pri čemer nižje vrednosti (bližje 0) predstavljajo pozitivno vrednotenje (na primer: »branje je potekalo povsem tekoče«), medtem

ko višje vrednosti proti 1 predstavljajo negativno vrednotenje (na primer »razumevanje je bilo zelo oteženo«).



Slika 4: Subjektivne ocene

Enosmerna analiza variance (ANOVA) za kriterij berljivosti pokaže, da so razlike statistično pomembne glede na tip prevoda, $F(2, 60) = 112$, $p < 0,05$. Ločeni Friedmanov test pokaže statistično pomembne razlike med popolnoma popravljeno in nepopravljeno vrsto prevoda, $\chi^2 = 30$, $p < 0,05$. Razlike v ocenah so bile statistično pomembne tudi med delno popravljenimi in nepopravljenimi prevodi, $\chi^2 = 5,1$, $p < 0,05$.

Ne glede na kriterij ocenjevanja, po katerem se je spraševalo (potek branja, razumevanje ali regresije), opazamo razlike v povprečni oceni za posamezen tip besedila in enak trend naraščanja povprečnih vrednosti. Povsem popravljena besedila (P2) so bila po vseh treh kriterijih ocenjena z nižjimi vrednostmi kot delno popravljena besedila (P1) in nepopravljena besedila (NP).

6.3. Korelacije

Zanimalo nas je tudi, ali morda obstaja povezanost med subjektivnimi ocenami udeležencev in merami sledilca očesnih gibov. V tabeli 2 so predstavljeni koeficienti korelacije med subjektivnimi ocenami berljivosti in tremi merami sledilca. Vrednosti v odebeljenem tisku so statistično pomembne pri kritični meji $p = 0,05$.

	ocena ~ št. fiksacij	ocena ~ čas branja	ocena ~ pov. čas fiksacije
besedilo 1	0,26	0,40	0,46
besedilo 2	0,43	0,41	0,15
besedilo 3	0,51	0,52	0,16

Tabela 2: Koeficienti korelacije (r) pri posameznih besedilih

6.4. Verbalna poročila

Tabela 3 povzema število odgovorov za posamezen tip besedila. Udeleženci so morali povedati, katero od treh prebranih besedil je bilo najtežje in katero najlažje, pri čemer je bilo možno ostati neodločen.

V tabeli 3 so zajeti le odgovori 22 udeležencev, ker preostalim nismo zastavili povsem identičnih vprašanj.

Nekateri so morali namreč povedati le, katero besedilo je bilo najtežje. Med temi so trije kot najtežje izbrali besedilo iz pogoja NP, eden P2, eden pa je ostal neodločen. Druge odgovore smo izključili, ker smo po naknadnem preverjanju opazili, da so udeleženci izbirali besedila iz vaje in je šlo torej za napačno razumevanje vprašanja, ki se je nanašalo le na zadnja 3 besedila.

	NP	P1	P2	nobeno
najtežje branje	17	3	0	2
najlažje branje	0	1	19	2

Tabela 3: Verbalno poročanje o berljivosti (N = 22)

Na vprašanje, ali bi bili sposobni po branju povzeti najosnovnejše informacije o prebranih besedilih, je 6 udeležencev odgovorilo z »ne«, 25 pa z »da«. Na vprašanje, ali so v primeru, ko so v besedilih naleteli na napake, na kakršenkoli način spremenili način branja oziroma bralno strategijo, je 22 udeležencev odgovorilo pritrdilno, 9 pa je bilo mnenja, da načina branja niso spreminjali. Pri tem so udeleženci omenjali obe možni spremembi branja – nekateri so trdili, da so začeli brati počasneje z vračanjem na že prebrane dele besedila, medtem ko je kar nekaj udeležencev potrdilo, da se niso posebej posvečali napakam v besedilu in so posledično brali celo malce hitreje kot običajno. Na vprašanje, ali so si kakšen tip napake posebej zapomnili, eden od udeležencev ni izpostavil ničesar, medtem ko so vsi ostali lahko naštel vsaj en tip napake. Najpogostejše so navajali napačen besedni red, neujemanje (»napačen sklon«, »napačen spol«, »napačne oblike besed«, »nedoločnik, kjer ni potrebno«), napačno postavljene vejice ter vrnjene besede (»ponavljanje besed«).

7. Diskusija

7.1. Primerjava rezultatov

V povprečju so bila nepopravljena besedila fiksirana bolj pogosto, čas branja pa je bil daljši kot pri delno popravljenih in povsem popravljenih različicah, kar je v skladu z začetno hipotezo. Do statistično pomembnih razlik je kljub temu prišlo le med obema skrajnima pogojem, ne pa tudi med delno popravljenimi in nepopravljenimi prevodi.

Takšne rezultate lahko razumemo kot delno skladne z ugotovitvami Dohertyja in soavtorjev (2010), kjer so udeleženci brali zelo dobro prevedene in zelo slabo prevedene stavke, učinek pa je bil razviden v številu fiksacij in času branja. Po drugi strani pa so rezultati za povprečno trajanje fiksacije nepričakovani in jih z našo hipotezo težko učinkovito pojasnimo. Na prvi pogled so bile razlike med tremi pogoji majhne, čeprav analiza variance razkrije sistematično razliko med posameznimi tipi prevodov. Glede na to, da se ti rezultati ne skladajo s predhodno študijo, iščemo morebitno razlago v dejstvu, da smo sami uporabljali cela besedila, medtem ko so v študiji Dohertyja in soavtorjev (2010) uporabljali le posamezne stavke.

Naslonili bi se lahko namreč na fenomen »sinteze« (ang. *wrap up effects*). V več študijah se je pokazalo, da je trajanje fiksacije na zadnjih besedah oziroma ob koncu stavka daljše kot v predhodnih delih. Raziskovalci so si ta pojav razlagali s hipotetičnim procesom sinteze prebrane vsebine v smiselno celoto, ki poteče na koncu prebrane enote (Staub et al., 2007). Možna razlaga naših rezultatov bi tako bila, da pride med branjem besedil večkrat do integracije prebranega dela v celoto. Ker je v slabih, manj berljivih besedilih takšna sinteza bolj potrebna, lahko do razlike v povprečnem trajanju fiksacij pride le na ravni besedil, ne pa tudi na ravni posameznih stavkov.

Bolj kot rezultati specifičnih mer sledilca pa so za naš namen zanimive primerjave s človeškimi ocenami in verbalnimi poročili. Ne glede na kriterij, po katerem smo spraševali, je bil trend subjektivnih ocen enak: popravljene različice so v povprečju dobivale višje ocene kot delno popravljene, te pa so bile ocenjene boljše kot nepopravljeni prevodi. Ocene med drugim jasno kažejo, da so udeleženci vse tri tipe prevodov, ne le najskrajnejše, dojemali kot kvalitativno različne. Podobno je moč sklepati tudi iz rezultatov verbalnih poročil, ki kažejo na to, da udeleženci niso (kvalitativno) enačili delno popravljenih in nepopravljenih prevodov. Na splošno so bile ocene bolj enotne za povsem popravljene različice kot za delno in nepopravljene različice, kjer so ocene bolj razpršene. Podobno kot pri Stymne in soavtorjih (2012) tudi naši rezultati sicer pokažejo zmerno korelacijo z ocenami udeležencev, pri čemer je bila korelacija najbolj konsistentna za čas branja, najmanj pa za povprečno trajanje fiksacije.

7.2. Sledilec in evalvacija strojnih prevodov

Neskladje med različnimi viri podatkov za vrednotenje berljivosti je zanimivo. Glede na splošno sprejeto prakso pri evalvaciji strojnih prevodov vzamemo rezultate človeških ocen kot standard za ovrednotenje rezultatov sledilca, te pa pokažejo, da so udeleženci razlikovali med vsemi tipi strojnih prevodov, uporabljenih v preizkušnji.

Menimo, da je inherentna omejitev uporabe sledilca za predlagani namen ta, da rezultati (kjer je variabilnost med posamezniki precejšnja) ne odsevajo zgolj kvalitete besedila, temveč nujno zajemajo tudi druge vplive na kognitivno procesiranje: posameznikovo bralno kompetenco, slog branja, odziv na navodila v nalogi, tip besedila ipd. Posameznik tako lahko prepozna določeno besedilo kot manj kvalitetno, vendar to zaradi množice potencialnih dejavnikov na branje morda ne bo nujno razvidno iz opazovanega vedenja, tj. gibanja oči med branjem.

Če na podlagi mer sledilca res ni možno razlikovati med takšnimi besedili, ki ne vsebujejo precejšnjega dela napak, in tistimi besedili, ki vsebujejo prav vse napake, potem smo lahko skeptični do smotnosti uporabe sledilca v evalvacijske namene. Predvidevamo namreč, da so v realnih okoliščinah, tj. pri razvoju novih prevajalnikov, razlike v prevodih nekega sistema in v prevodih nekoliko izboljšane različice tega istega sistema lahko precej manj očitne, kot so bile razlike v besedilih iz naše preizkušnje. Vprašanje, ali je besedilo brez napak bolj berljivo od takega z veliko napakami pa smatramo za trivialno in kot tako ne zahteva preverjanja s tako zahtevno metodologijo. Ključno vprašanje je namreč, ali lahko na podlagi

rezultatov sledilca dobimo informacijo, ki nam je druge metode ne omogočajo.

Glede na to, da, kolikor nam je znano, sistematično zbrani podatki sledilca očesnih gibov med branjem za slovenščino še niso na voljo (prim. Ferbežar, 2012), bi se vzdržali prestrogega vrednotenja metodologije. Vsakršno nadaljnje delo na tem področju bi pripomoglo tudi k boljši kakovosti podatkov, ki v našem primeru ni bila vedno zadovoljiva.

V morebitnih nadaljnjih študijah bi bilo zanimivo opraviti analizo podatkov na ravni posameznih besed, kot so to storili Stymne in soavtorji (2012), kjer se je prav ta pristop izkazal za bolj informativnega od uveljavljenih tehnik. V naši študiji zaradi tehničnih težav na koncu ta možnost ni bila izvedljiva. Poleg analize na ravni besed, bi lahko v morebitnih prihodnjih študijah tako testirali še druge, bolj specifične mere sledilca, kot je npr. število regresij (sakade na že prebrane dele besedila), ki se pogosto povezujejo s težavami razumevanja med branjem.

Nadaljnja raziskovalna vprašanja se odpirajo tudi pri uporabi kontinuiranih lestvic za ocenjevanje, namesto uveljavljenih 5- ali 7-stopenjskih (prim. Graham et al., 2013a).

8. Zaključek

Z rezultati sledilca očesnih gibov smo lahko le delno potrdili uvodno hipotezo. Mere sicer kažejo pričakovan trend naraščanja, a se ne izkažejo kot zelo učinkovit pokazatelj razlik med posameznimi tipi prevodov. Na podlagi rezultatov te študije tako menimo, da sledilec očesnih gibov našega testa ni prestopil: da bi ga lahko označili kot uporabno orodje za evalvacijo, bi morali rezultati pokazati sliko, skladno s predvidevanjem na podlagi kategorizacije napak in ocenami udeležencev.

Kljub sicer nekoliko bolj skeptičnemu zaključku smo mnenja, da gre za obetaven pristop, ki sicer zahteva kar nekaj metodološke pozornosti in natančnosti, a hkrati ponuja drugačen vpogled v obravnavano tematiko. V prihodnje bi sledilec očesnih gibov ob ustrezno izboljšani kakovosti podatkov veljalo izkoristiti za bolj podrobno analizo branja ter tudi za bazične raziskave branja.

9. Literatura

- Callison-Burch, C., F. Cameron, P. Koehn, C. Monz in J. Schroeder, 2007. (Meta-) Evaluation of Machine Translation. *Proceedings of ACL-2007 Workshop on Statistical Machine Translation*.
- Doherty, S., S. O'Brien in M. Carl, 2010. Eye-tracking as MT evaluation technique. *Machine Translation*, 1: 1–13.
- Ferbežar, I., 2012. *Razumevanje in razumljivost besedil*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani.
- Graham, Y., T. Baldwin, A. Moffat in J. Zobel, 2013a. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Linguistic Annotation Workshop and Interoperability with Discourse*, 33–41. 8. in 9. avgust, Sofija, Bolgarija.
- Graham, Y., T. Baldwin, A. Moffat in J. Zobel, 2013b. Crowd-Sourcing of Human Judgments of Machine Translation Fluency. *Proceedings of the Australasian Language Technology Workshop*, 16–24. 4.–6. december, Dunedin, Nova Zelandija.
- Just, M.A. in P.A. Carpenter, 1980. A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review*. 4:329–354.
- Koehn, P., 2010. *Statistical Machine Translation*. Cambridge: Cambridge University Press.
- Papineni, K., S. Roukos, T. Ward in W.-J. Zhu, 2002: BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. Julij 2002, Philadelphia, ZDA.
- Rayner, K., 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, 3: 372–422.
- Staub, A. in K. Rayner, 2007. Eye Movements and on-line comprehension processes. Gaskell, G. (ur.), *The Oxford Handbook of Psycholinguistics*, Oxford: Oxford University Press.
- Stymne, S., H. Danielsson, S. Bremin, H. Hu, J. Karlsson, A.P. Lillkull in M. Wester, 2012. Eye Tracking as a Tool for Machine Translation Error Analysis. *Zbornik konference LREC 2012*, 1121–1126.
- Verdonik, D. in M. Sepesy Maučec, 2013. O avtomatski evalvaciji strojnega prevajanja. *Slovenščina 2.0*. 1/1. 111–113.