Ohranjanje jezikovne zahtevnosti besedil pri prevajanju testov PISA

Špela Arhar Holdt,*♦ Iztok Kosem*♦

* Zavod za uporabno slovenistiko Trojina (CUJT),
Trg republike 3, 1000 Ljubljana
♦ Filozofska fakulteta Univerze v Ljubljani,
Aškerčeva 2, 1000 Ljubljana
spela.arhar@trojina.si, iztok.kosem@trojina.si

Povzetek

Eden glavnih izzivov mednarodnih testiranjih je zagotoviti ohranjanje jezikovne zahtevnosti besedil, še zlasti v primerih, ko testiranja preverjajo jezikovno kompetenco testirancev, kot velja za raziskavo bralne pismenosti PISA. V prispevku argumentirava, da je ohranjanje jezikovne zahtevnosti testov mogoče evalvirati in izboljšati z uporabo korpusnih podatkov o tipičnosti jezikovnih pojavov v realni jezikovni rabi. Z medjezikovno korpusno primerjavo pokaževa, da so slovenski testi spričo redkega besedišča in atipičnih kolokacij zahtevnejši od angleških, in predlagava postopek, s katerim bi bilo mogoče identificirano stanje v prihodnje izboljšati.

Preserving Text Difficulty in Translations of PISA Tests

One of the main challenges of international assessments is to maintain the same level of difficulty of texts across different languages, which is particularly relevant for (reading) literacy assessments such as PISA which test language abilities of participants. This paper argues that the level of text difficulty in different languages can be evaluated and maintained by using corpora, which contain information on (a)typicality of words, phrases etc. in real language use. Using an interlanguage corpus-based comparison we show that Slovene versions of PISA texts are more demanding than their English counterparts considering the amount of rare vocabulary and atypical collocations. Finally, we propose a procedure that could be used to address such problems and ensure that the level of text difficulty is maintained in different languages.

1 Uvod

PISA (*The Programme for International Student Assessment*)[1] je mednarodna raziskava, ki primerja uspešnost 15-letnih učencev na področju matematike, naravoslovja in branja. Raziskava, ki jo je leta 2000 lansirala OECD, je bila v Sloveniji prvič izvedena leta 2006, od takrat pa se ponavlja vsaka tri leta.

Ker gre za mednarodno raziskavo, se med dejavniki vpliva omenjajo – poleg denimo ekonomske razvitosti države oz. lokalnega okolja, vsebine učnih načrtov, motivacije in izvežbanosti šol ter učencev za testiranje – tudi prevodi testov v nacionalne jezike (npr. Grisay et al., 2007; Arffman, 2012; Solano-Flores et al., 2013), pri čemer je med glavnimi izzivi potreba po ohranjanju jezikovne zahtevnosti besedil. Ta potreba je v smernicah za prevajanje in prilagoditev testov izpostavljena s splošnimi vodili (OECD 2010: 11–13), vendar ob tem ni opredeljena metodologija, ki bi omogočila objektivno zagotavljanje in preverjanje zahtevanega stanja.

Namen raziskave, ki jo predstavlja pričujoči prispevek, je zato oceniti, ali se v slovenskih prevodih testov PISA pojavlja sprememba jezikovne zahtevnosti, in razviti postopek, s katerim bi bilo tovrstna težavna mesta prevoda mogoče sistematično detektirati.

2 Ohranjanje jezikovne zahtevnosti testov

Ohranjanje jezikovne zahtevnosti testov v tem prispevku povezujemo s tipičnostjo jezikovnih pojavov v realni rabi. Če se zgodi, da prevajalec pri svojem delu med več jezikovnimi variantami izbere tisto, ki je v primerjavi z izvirnikom v rabi manj tipična – ali iz dveh besed, ki sta vsaka zase sicer v rabi pogosti, tvori besedno zvezo, ki je v rabi redka – je posledično jezik prevoda zahtevnejši od jezika izvirnika. Podobno velja, če prevajalec izbere skladenjsko strukturo, ki je v jeziku prevoda manj tipična za izbrani jezikovni kontekst. Primer, h kateremu se vračamo v razdelku 4.1, je prevod zveze *mobile phone.* Slednja je v angleški jezikovni rabi relativno pogosta in mogoče je predvideti, da z dekodiranjem pomena testirani po večini ne bodo imeli težav. V slovenščini je na voljo več poimenovalnih možnosti, med katerimi je prevajalec izbral *prenosni telefon.* Ker je ta zveza v realni jezikovni rabi redka (precej redkejša od npr. *mobilni telefon*), izbira pomeni manjšo verjetnost, da testirani zvezo pozna, kar lahko povzroči počasnejše in tudi manj uspešno dekodiranje besedila. Na enak način je seveda mogoče predvideti tudi obraten vpliv, tj. da z določeno jezikovno izbiro prevajalec zahtevnost testa zniža.

Če zaradi medjezikovnih razlik ni mogoče zagotavljati povsem primerljive pogostnosti na ravni posameznih besed oz. zvez, pa je uravnoteženost vsekakor treba zagotoviti na ravni besedila kot celote. Navedena naloga v praksi ni enostavno rešljiva, saj je pogostnost jezikovnih elementov v rabi pri pripravi in redakciji prevoda težje sistematično spremljati in medjezikovno primerjati.[2] Kot možno rešitev v tem prispevku predlagava dopolnitev postopka prevajanja testov s statistično primerjavo jezikovnih pojavov, kakor se kažejo v referenčnih korpusih jezika izvirnika in prevoda. V raziskavi po naročilu Pedagoškega inštituta sva predlagano metodo preizkusila na testih bralne pismenosti PISA v letih 2009 in 2012. V nadaljevanju predstavljava metodo in rezultate raziskave, v razpravi pa predlagano metodo oceniva z vidika uporabljenih korpusnih virov.

[1] http://www.oecd.org/pisa/home/

[2] Kot tudi ni mogoče zgolj z jezikovno intuicijo sintetično spremljati drugih jezikovnih značilnosti, ki lahko vplivajo na zahtevnost prevoda, npr. pomenskih specifičnosti izbranih prevodnih ustreznic ali kompleksnosti skladenjskih struktur. Vendar so vsaj na ravni posamezne pojavitve ti izzivi opaznejši, morda tudi zato, ker so bolj pričakovani, obenem pa so tudi lažje rešljivi s pomočjo obstoječih jezikovnih priročnikov.

3 Kvantitativna analiza

3.1 Metoda primerjave korpusnih podatkov

Za namene raziskave sta bila uporabljena referenčna korpusa, ki predstavljata reprezentativna vzorca sodobne slovenščine oz. angleščine: 1,2-milijardni korpus Gigafida (Logar Berginc et al., 2012) za slovenski jezik in 2,1-milijardni Oxford English Corpus (OEC)[3] za angleški jezik. Primerjalna analiza rabe in frekvenc besedišča je bila izvedena z orodjem Sketch Engine (Kilgarriff et al., 2004).[4]

V prvem koraku so bile besede in besedne zveze, ki se pojavljajo v besedilih analiziranih testov, urejene v primerjalne tabele, skupaj z (ustrezno relativiziranimi) podatki o pogostnosti iz obeh korpusov. V analizo so bile zajete polnopomenske besede, saj so funkcijske med najpogostejšimi v obeh jezikih in kot take manj relevantne za primerjavo. Analiza besed in besednih zvez je bila nato dopolnjena s širšimi podatki o kolokacijah, ki zajemajo večji nabor skladenjskih vzorcev in omogočajo primerjavo kombinacij besed, ki se ne pojavljajo neposredno skupaj, npr. glagolov in samostalnikov, glagolov in predlogov ipd.

V tabelah so bili nato označeni primeri, kjer prihaja do največjih odstopanj na ravni pogostnosti rabe: v Tabelah 1 in 2, ki prikazujeta izsek podatkov za test z naslovom *Varnost prenosnih telefonov*, je besedišče, ki se v enem jeziku pojavlja več kot 50 % redkeje (na milijon besed) kot v drugem, natisnjeno okrepljeno.

Tabela 3 prinaša izsek podatkov kolokacijske analize, kjer so bili kot relevantni za nadaljnjo analizo označeni primeri, kjer je statistična moč kolokacije za več kot 25 % nižja od moči kolokacijskega para v drugem jeziku. Kolokacije smo opazovali v razponu od -5 do +5, upoštevana statistika pa je MI.log_f (v orodju Sketch Engine predhodno poznana kot *salience*).[5]

slovenska beseda	bes. vrsta	pogostnost Gigafida	pogostnost na milijon	angleška beseda	bes. vrsta	pogostnost OEC	pogostnost na milijon
mladi	**sam.**	**1430**	**1,2**	(the) young	sam.	78121	37,7
možgani	sam.	54190	45,7	brain	sam.	145120	70,0
nakazati	**gl.**	**28049**	**23,6**	suggest	gl.	562670	271,4
namen	sam.	241383	203,4	purpose	sam.	253924	122,5
napačen	**prid.**	**56106**	**47,3**	wrong	prid.	323375	156,0
napravica	sam.	4734	4,0	gadget	sam.	12190	5,9
navodila	sam.	54336	45,8	**instructions**	**sam.**	**49273**	**23,8**

Tabela 1: Primerjava odstopanj na ravni pogostnosti rabe v korpusih Gigafida in OEC: besede.

slovenska besedna zveza	pogostnost Gigafida	pogostnost na milijon	angleška besedna zveza	pogostnost OEC	pogostnost na milijon
prenosni telefon	**4922**	**4,1**	mobile phone	47752	23,0
protislovno poročilo	**9**	**0,0**	conflicting report	903	0,4
radijski val	2569	2,2	**radio waves**	**1687**	**0,8**
radiofrekvenčno valovanje	1	0,0	radio frequency waves	22	0,0
rak pri otrocih	**121**	**0,1**	childhood cancer	719	0,3
sodobni življenjski slog	153	0,1	modern lifestyles	240	0,1
stanje pripravljenosti	1479	1,2	on standby	2595	1,3

Tabela 2: Primerjava odstopanj na ravni pogostnosti rabe v korpusih Gigafida in OEC: zveze.

slovenska kolokacija	pogostnost Gigafida	pog. na milijon	moč kolokacij	mesto na seznamu	angleška kolokacija	pogostnost OEC	pog. na milijon	moč kolokacij	mesto na seznamu
kupiti / telefon	3513	3,0	52,645	190	**buy / phone**	**1808**	**0,9**	**36,591**	**265**
opazovati / v	11874	10,0	26,98	270	observe / under	1279	0,6	26,479	743
povezan / z	195453	164,7	75,838	5	linked / to	58607	28,3	66,741	3
sevanje / iz	**702**	**0,6**	**23,859**	**380**	radiation / from	3589	1,7	38,282	133
energija / povezava	**290**	**0,2**	**13,962**	**> 1000**	power / communicate	249	0,1	19,684	586
močen / sevanje	455	0,4	41,202	90	**high / emission**	**433**	**0,2**	**28,77**	**252**
razprava / o	55026	46,4	75,163	5	**debate / about**	**964**	**0,5**	**26,547**	**197**

Tabela 3: Primerjava kolokacijske moči v korpusih Gigafida in OEC.[6]

[3] http://www.oxforddictionaries.com/words/the-oxford-english-corpus, uporabljena verzija korpusa je iz februarja 2012.

[4] Korpus Gigafida smo analizirali v lokalni inštalaciji podjetja Amebis, d. o. o., Kamnik, korpus OEC pa v inštalaciji podjetja Lexical Computing Ltd., pri čemer smo za uporabo korpusa OEC pridobili dovoljenje založbe Oxford University Press.

[5] Enačba je navedena v: https://trac.sketchengine.co.uk/raw-attachment/wiki/SkE/DocsIndex/ske-stat.pdf

[6] Podatek o pogostnosti besed oz. mesto kolokatorja na seznamu kolokatorjev ponuja dopolnilo podatku o kolokacijski moči, kar pride prav v primerih tipa *velikopotezna zamisel – ambitious idea*, ki sta po moči primerljivi, vendar pa sta v rabi besedi *velikopotezen* in *zamisel* precej redkejši od *ambitious* in *idea*.

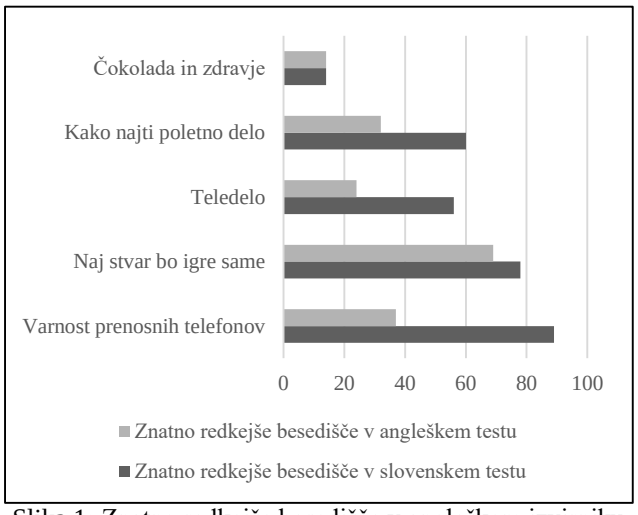
Pri interpretaciji podatkov je treba upoštevati različne vzroke, ki lahko vodijo do razlik v pogostnosti. V prvi vrsti so to razlike v pomenski členjenosti, kot npr. velja za par *napačen – wrong* iz Tabele 1. Za *napačen* referenčni slovarski viri navajajo en sam pomen, medtem ko ima *wrong* poleg slovenščini primerljivega *napačen* še pomen, ki ga v slovenščini izraža *narobe* ('Is anything wrong?' – 'Je kaj narobe?'). Upoštevati je treba tudi slovnične razlike med jezikoma, npr. kategorijo dovršnosti pri interpretaciji pogostnosti angleških in slovenskih glagolov (slovensko *nakazati* in *nakazovati,* angleško *to suggest*). Pomemben dejavnik je tudi struktura in označenost korpusov, npr. pri primeru *mladi* lahko pričakujemo napake pri pripisovanju besednovrstne oznake, in podobno. Nenazadnje je treba izpostaviti, da sta bili meji (50 % redkejša pogostnost besed in zvez ter 25 % redkejša kolokacijska moč) v tej fazi dela izbrani arbitrarno oz. intuitivno in da način primerjave podatkov med korpusoma daje natančnejše rezultate za pogoste primere, pri redkejših pa so prikazane razlike lahko do določene mere napihnjene. Zaradi naštetih razlogov lahko podatke razumemo zgolj kot izhodišče, ki lahko nakaže potencialne probleme, ki pa jih je treba v nadaljevanju natančneje kvalitativno raziskati.

Ker se podatki dotikajo izključno besediščne ravni jezika, uporabljava za predstavitev izsledkov raziskave namesto širšega pojma *jezikovna zahtevnost* poimenovanje *besediščna zahtevnost.*

3.2 Rezultati primerjave

Predstavljeni postopek je bil uporabljen na petih besedilih, ki so se uporabljala za testiranja bralne pismenosti v letih 2009 in 2012:

- *Varnost prenosnih telefonov* (2009) je polstrokovno besedilo na temo varnosti oz. nevarnosti, ki jih (morda) prinaša mobilna telefonija.
- *Naj stvar bo igre same* (2009) je dramski odlomek, v katerem se tri osebe pogovarjajo o tem, kako težko je začeti pisanje gledališke igre.
- *Teledelo* (2009) prinaša izjavi dveh oseb, ki pojasnjujeta svoji mnenji glede dela na daljavo.
- *Kako najti poletno delo* (2012) je letak, ki svetuje mladim, kako poiskati poletno zaposlitev.
- *Čokolada in zdravje* (2012) je kratko polstrokovno besedilo, ki navaja pozitivne učinke epikatehina.



Slika 1: Znatno redkejše besedišče v angleškem izvirniku in slovenskem prevodu testov PISA.

Slika 1 predstavlja rezultate kvantitativne analize, v katerih je združeno število besed in besednih zvez, ki se v enem od korpusov pojavljajo več kot 50 % redkeje kot v drugem, in kolokacij, pri katerih je statistična moč v enem od korpusov vsaj 25 % šibkejša kot v drugem.

Razmerja so od besedila do besedila različna. Besedilo *Čokolada in zdravje* se zdi glede na podatke najbolj uravnoteženo (je pa tudi daleč najkrajše), prav tako je primerljiva pogostnost besedišča v dramskem odlomku *Naj stvar bo igre same.* Pri preostalih treh besedilih je opaziti zelo visoka odstopanja, ki kljub predhodno opredeljenim metodološkim zadržkom dovolj jasno dokazujejo, da je besediščna zahtevnost slovenskih testov v splošnem višja kot pri angleških različicah. Ta ugotovitev sama na sebi seveda ne dokazuje tudi dejanskega vpliva na uspešnost pri testiranju, je pa močan indikator, da je v prihodnje vprašanju pogostnosti (in s tem jezikovne avtentičnosti) pri prevodih nujno posvetiti več pozornosti.

4 Kvalitativna analiza

Ker do razlik v pogostnosti lahko prihaja iz različnih razlogov, morajo biti primeri, ki glede na rezultate v slovenskem testu izstopajo po svoji redkosti, natančneje raziskani. V nadaljevanju predstavljava nekaj izbranih primerov tovrstne dodatne analize.

4.1 Prenosni vs. mobilni telefon

Kot je razvidno iz Tabele 2, se zveza *mobile phone* v OEC pojavi 23-krat na milijon zadetkov, medtem ko se izbrana prevodna ustreznica *prenosni telefon* v korpusu Gigafida pojavi le 4,1-krat na milijon. Pregled korpusnih podatkov pokaže, da sta alternativni možnosti *mobilni telefon* in *mobilnik* v rabi znatno pogostejši in zato morda za prevod ustreznejši: prva se pojavi 31,8-krat na milijon, druga pa 18,4-krat na milijon (Tabela 4). Dodatna potrditev, da je za slovenščino *mobilni telefon* bolj tipična izbira je dejstvo, da se slednja zveza pojavlja tudi kot del daljše zveze *varnost* oz. *nevarnost mobilnih telefonov* (čeprav z nizkim številom pojavitev). Nenazadnje korpusni podatki pokažejo, da se zveza *prenosni telefon* za razliko od *mobilni telefon* uporablja tudi v pomenu 'brezvrvični stacionarni telefon', kar je lahko za razumevanje besedila dodatno obremenjujoče.

Besedna zveza	Pogostnost v Gigafida	Pogostnost na milijon
prenosni telefon	4922	4,1
mobilni telefon	37720	31,8
mobilnik	21808	18,4
varnost prenosnih telefonov	0	0,0
varnost mobilnih telefonov	13	0,0
varnost mobilnikov	1	0,0
nevarnost prenosnih telefonov	0	0,0
nevarnost mobilnih telefonov	10	0,0
nevarnost mobilnikov	0	0,0

Tabela 4: *Prenosni telefon* in *mobilni telefon* v korpusu Gigafida.

4.2 Teledelo vs. delo na daljavo

V nadaljevanju navajamo del besedila *Teledelo,* v katerem so sivo obarvani primeri, ki se glede na opravljeno analizo pojavljajo v obravnavani ediciji testa znatno redkeje. Slika 2 prikazuje izsek v angleškem testu, Slika 3 pa slovenskega. V prikazu so podatki za redko besedišče in atipične kolokacije združeni, kar pomeni, da so primeri, ki se pojavljajo na ravni posamezne besede in obenem kot del zvez in kolokacij, označeni kot večbesedni problem.

TELECOMMUTING

The way of the future

Just imagine how wonderful it would be to "telecommute" to work on the electronic highway, with all your work done on a computer or by phone! No longer would you have to jam your body into crowded buses or trains or waste hours and hours travelling to and from work. You could work wherever you want to – just think of all the job opportunities this would open up!

Molly

Slika 2: Redkejše besedišče v odlomku *Telecommuting*.

TELEDELO

Delo prihodnosti

Predstavljajte si, kako krasno bi bilo imeti "teledelo" in potovati po elektronski avtocesti, pri čemer bi vse delo opravili na računalniku ali prek telefona! Ne bi se vam bilo treba več gnesti na natrpanih avtobusih ali vlakih ali zapravljati dolgih ur za vožnjo v službo in domov. Lahko bi delali, kjer bi hoteli - samo pomislite, koliko priložnosti za delo bi to odprlo!

Maja

Slika 3: Redkejše besedišče v odlomku *Teledelo*.

Oznake opozarjajo na mesta, ki zahtevajo dodaten prevajalski premislek. Pojav primerov *služba – job* in *pomislite – think* je mogoče pripisati razlikam v pomenski členjenosti besed (razdelek 3.1). Za preostale primere je mogoče s spletnim vmesnikom korpusa Gigafida relativno enostavno[7] poiskati v rabi pogostejše ustreznice:[8]

- *kako krasno* (90 pojavitev) – *kako čudovito* (418 pojavitev);
- *natrpan avtobus* (24 pojavitev) – *prepoln avtobus* (33 pojavitev);
- *gnesti/avtobusu* (8 pojavitev) – *drenjati/avtobusu* (12 pojavitev);
- *zapravljati dolge ure* (0 pojavitev) – *zapravljati ure in ure* (5 pojavitev) – *zapravljati čas* (1391 pojavitev);
- *hoteti* (517.056 pojavitev) – *želeti* (882.338 pojavitev);
- *priložnost za delo* (321) – *možnosti za zaposlitev* (1.326 pojavitev) – *delovno mesto* (111.160 pojavitev).

Če k temu dodamo še primerjavo med besedo *teledelo* (222 pojavitev) in zvezo *delo na daljavo* (354 pojavitev), ki omogoča nekoliko bolj tekoč prevod prve povedi besedila, je mogoče na podlagi opravljene analize pripraviti novo različico besedila (Slika 4).[9] S prenosom predstavljene statistične analize na skladenjski nivo bi bilo mogoče zagotoviti, da bi bil prevod še bližje značilnostim slovenskega jezika, vendar že izboljšave na ravni besedišča pripomorejo k berljivosti in vtisu avtentičnosti besedila.

DELO NA DALJAVO

Delo prihodnosti

Predstavljajte si, kako čudovito bi bilo delati na daljavo in se peljati v službo po elektronski avtocesti, pri čemer bi vse delo opravili na računalniku ali prek telefona! Ne bi se vam bilo treba več drenjati na prepolnih avtobusih ali vlakih ali zapravljati časa za vožnjo v službo in domov. Lahko bi delali, kjer bi želeli - samo pomislite, koliko možnosti za zaposlitev bi to odprlo!

Maja

Slika 4: Ponovni prevod odlomka *Delo na daljavo*.

5 Diskusija

Čeprav opravljena raziskava nakazuje, da podatki o razmerjih v pogostnosti jezikovnih pojavov v rabi lahko pripomorejo k ohranjanju jezikovne zahtevnosti prevoda, se ob predlaganem postopku odpira nekaj pomembnih poudarkov. Prvi je, da pogostnosti pojavov v referenčnem korpusu seveda ni mogoče razumeti kot indikator, kateri jezikovni elementi so del dejanske jezikovne rabe mladih. Na eni strani zato, ker referenčni korpusi tipično vsebujejo besedila, s katerimi se srečujejo predvsem odrasli govorci,[10] na drugi strani zato, ker v nobenem primeru na osnovi podatkov o jezikovni recepciji ni mogoče sklepati o jezikovni produkciji. Slednjo je za slovensko šolajočo se populacijo mogoče raziskovati s korpusom Šolar (Rozman et al., 2012; Kosem et al., 2012; Arhar Holdt et al., 2016), vendar le določen del, tj. pisno produkcijo, ki poteka v okviru šolskega pouka.

Za dano nalogo je uporaba referenčnega korpusa utemeljena zato, ker testi PISA prinašajo besedila, povsem

[7] www.gigafida.net, sinonime je mogoče poiskati s seznama kolokatorjev v zavihku Okolica, nato pa preveriti pogostnost in kontekst rabe v konkordančnih nizih, ki so rezultat enostavnega ali naprednega iskanja.

[8] V prispevku navajava nekaj izbranih alternativnih možnosti za vse identificirane pare in iz navedenih primerov je razvidno, da niso vse razlike enako relevantne. Za namene prikaza upoštevava v prevodu na Sliki 4 tudi manj relevantne rezultate (kjer so frekvence obeh različic bodisi zelo nizke bodisi zelo visoke), pri uporabi korpusa za izboljšavo prevoda dejanskih testov pa je treba upoštevati tako razmerja kot frekvence dobljenih rezultatov.

[9] Prevod je provizoričen in služi ponazoritvi možnosti, ki jih prinaša predlagani postopek za avtomatsko identifikacijo besediščno problematičnih mest.

[10] Ob čemer je nujen poudarek, da je slabo raziskano, katera besedila mladi dejansko berejo in katerih ne. Vsekakor je mogoče predvideti, da so med njimi dela s seznamov šolskega branja in mladinska besedila, ki jih je najti na seznamih najbolje prodajanih in izposojanih v knjižnicah. Načrti za nadgradnjo slovenskega referenčnega korpusa že predvidevajo nadgradnjo s tovrstnim gradivom (Krek et al., 2016). O samostojnem branju neumetnostnih besedil je podatkov manj.

primerljiva tistim, zajetim v referenčne korpuse (strokovna besedila, različne publicistične zvrsti, odlomke iz leposlovja, ipd.). V tem smislu je referenčni korpus ustrezen vir za primerjavo pogostnosti oz. redkosti alternativnih ubeseditvenih možnosti ali tipičnosti oz. atipičnosti kolokacij v ciljnem jeziku prevoda. Pri tem je ključna medjezikovna primerjava, ki razkrije najbolj problematična mesta, torej tista, kjer je določen jezikovni pojav v izhodiščni ediciji testa pogost, v ciljnem jeziku pa redek.

Korpusne podatke bi bilo sicer mogoče uporabiti tudi na druge načine, npr. za označevanje besed, ki so v jeziku zelo pogoste, srednje pogoste in redke (West, 1953; Xue & Nation, 1984; Coxhead, 2000). Po Nation in Waring (1997) naj bi imel jezik štiri skupine besed: splošno besedišče, akademsko besedišče, terminološke besede in zelo redke besede. Splošno besedišče naj bi obsegalo približno 2000 besed, ki jih pri komunikaciji redno uporabljamo in naj bi jih poznali vsi materni (ali spoznali vsi nematerni) govorci nekega jezika. Glede na to opredelitev bi bilo od 15-letnika, ki piše test PISA, mogoče pričakovati, da bo poznal večino 2000 najpogostejših besed, medtem ko bo zahtevnost besed izven te skupine naraščala s padanjem pogostnosti rabe. Takšen pregled testov sicer ne bi izpostavil prevodnih težav enako neposredno kot medjezikovna primerjava, zelo uporaben pa bi bil za preverjanje besediščne zahtevnosti nacionalnih preverjanj znanj ter raznovrstnih učnih gradiv oz. za spremljanje, usmerjanje in vrednotenje razvoja pisne produkcije posameznikov, vključenih v šolski proces.

V vsakem primeru se zdi vključitev podatkov o realni jezikovni rabi v prevajanje testov nujna. Raziskava razkriva težave predvsem na kolokacijski ravni, kjer najdemo več primerov, pri katerih je slovenska kolokacija statistično šibkejša kot angleška, kot primerov, pri katerih je angleška kolokacija redkejša oziroma statistično šibkejša od slovenske.[11] Rezultat so besedila, ki na prvi pogled delujejo neproblematična, vendar so zaradi vrste atipičnih kombinacij manj tekoča in »naravna«, kot to velja za angleško edicijo testa. Zveze, ki jih v korpusu Gigafida ni moč najti, se pojavljajo tudi v delu testa z vprašanji oziroma nalogami za učence, kar ima lahko še bolj neposreden vpliv na uspešnost reševanja.

Rezultati raziskave vodijo k ugibanju, ali na besediščno zahtevnost ne vpliva tudi želja prevajalca izbrati najbolj nesporno standardno oz. knjižno ustreznico (kot nakazuje že večkrat omenjeni primer izbire *prenosni telefon* vs. *mobilni telefon*). Na tej točki je razpravo potrebno povezati s problemom pomanjkljivega slovarskega opisa za slovenščino (Gorjanc et. al., 2015), predvsem na ravni predstavitve kolokacij in sinonimije. Kot kažejo rezultati, bi bila v obeh primerih zelo zaželena vključitev podatkov o pogostnosti jezikovnih pojavov v rabi. Čeprav je postopek, ki ga predlagava v prispevku, lahko v vsakem primeru za prevajalca dobrodošla povratna informacija, bi seveda v temelju kazalo poskrbeti za to, da bo imel slednji že v prvem koraku na voljo podatke, ki jih potrebuje. Seveda pa je pri razpravi o težavah in možnih rešitvah potrebna dobršna stopnja previdnosti, saj jih je – tudi zaradi kompleksnega prevodnega postopka, v katerega je vključenih več oseb in vsebuje številna usklajevanja – težko enoznačno določiti.

[11] Nekaj primerov iz testa *Varnost prenosnih telefonov: izražanje genov*, *laboratorijske razmere, neodvisno testirati, protislovno poročilo, poškodba zaradi toplote, zaščitne napravice* in *zmanjšana zbranost.*

6 Sklep

Za korektno pridobivanje in primerjavo rezultatov mednarodnih testiranj morajo biti prevodi testov v smislu jezikovne zahtevnosti primerljivi z izvirnikom. Predlagani postopek primerjalne rabe referenčnih korpusov za identifikacijo spremembe besediščne in kolokacijske zahtevnosti (in avtentičnosti) pri prevajanju testov PISA se je izkazal kot koristen korak pri doseganju tega cilja: s postopkom je mogoče identificirati potencialno problematična mesta pri izbiri prevodnih ustreznic in omogočiti prevajalcu širši, sintetični pogled na zahtevnost izvirnika in besedila. Rezultati raziskave so pokazali, da so slovenski testi v primerjavi z angleškimi na besediščni ravni zahtevnejši in posledično težji za razumevanje. Vključitev predlaganega koraka v postopek priprave testov se torej kaže kot smiselna in potrebna. Čeprav se postopek osredotoča zgolj na enega od številnih dejavnikov vpliva, namreč omogoča izboljšanje stanja z relativno nizkim izhodiščnim finančno-časovnim vložkom, ki pa bi ga bilo z nadaljnjimi prilagoditvami metodologije mogoče še dodatno optimizirati. Pomemben korak za prihodnje delo pa je ugotoviti, v kolikšni meri izboljšave prevoda dejansko vplivajo na rezultate testiranj bralne pismenosti PISA.

7 Zahvala

Analiza testov PISA je bila delno financirana s strani Pedagoškega inštituta, delno pa s strani ARRS v okviru infrastrukturnega programa *Center za uporabno jezikoslovje* pri zavodu Trojina (šifra I0-0051). Zahvala gre tudi strokovnjakom na seminarju Šolskega polja Centra za študij edukacijskih politik za dragocene povratne informacije glede interpretacije pridobljenih rezultatov ter anonimnima recenzentoma prispevka za koristna dopolnila.

8 Literatura

Inga Arffman. 2012. Unwanted Literal Translation: An Underdiscussed Problem in International Achievement Studies. *Education Research International*, 2016 (ID 503824): 1–13.

Špela Arhar Holdt, Iztok Kosem in Polona Gantar. 2016. Corpus-Based Resources for L1 Teaching: The Case of Slovene. V: A. Marcus-Quinn in T. Hourigan, ur. *Handbook on Digital Learning for K-12 Schools*. Springer, v tisku.

Averil Coxhead. 2000. A New Academic Word List. *TESOL Quarterly* 34(2)**:** 213–238.

Vojko Gorjanc, Polona Gantar, Iztok Kosem in Simon Krek, ur. 2015. *Slovar sodobne slovenščine: problemi in rešitve*. Znanstvena založba Filozofske fakultete UL.

Aletta Grisay, John H.A.L. de Jong, Eveline Gebhardt, Alla Berezner in Beatrice Halleux-Monseur. 2007. Translation equivalence across PISA countries. *Journal of Applied Measurement*, 8(3): 249–266.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrz in David Tugwell. 2004: The Sketch Engine. V.: G. Williams in S. Vessier, ur. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004 Lorient, France July 6–10, 2004*, str. 105–116, Lorient. Universite de Bretagne - sud.

Iztok Kosem, Mojca Stritar Kučuk, Sara Može, Ana Zwitter Vitez, Špela Arhar Holdt in Tadeja Rozman. 2012.

Analiza jezikovnih težav učencev: korpusni pristop. Trojina, zavod za uporabno slovenistiko.

Simon Krek, Polona Gantar, Špela Arhar Holdt in Vojko Gorjanc. 2016. Nadgradnja korpusov Gigafida, Kres, ccGigafida in ccKres. *Konferenca Jezikovne tehnologije in digitalna humanistika 2016*, v pripravi.

Nataša Logar Berginc, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt in Simon Krek. 2012. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba.* Ljubljana, Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Paul Nation, Robert Waring. 1997. Vocabulary size, text coverage and word lists. V N. Schmitt, M. McCarthy (ur.) *Vocabulary: Description, Acquisition and Pedagogy, str.* 6–19. Cambridge: Cambridge University Press.

OECD. 2010. *Translation and adaption guidelines for PISA 2012*. Dostop 5. 3. 2016: http://www.oecd.org/pisa/pisaproducts/49273486.pdf

Tadeja Rozman, Irena Krapš Vodopivec, Mojca Stritar Kučuk in Iztok Kosem. 2012. *Empirični pogled na pouk slovenskega jezika*. Trojina, zavod za uporabno slovenistiko.

Guillermo Solano-Flores, Luis Ángel Contreras-Niño in Eduardo Backhoff. 2013. The measurement of translation error in PISA-2006 items: An application of the theory of test translation error. V: M. Prenzel, M. Kobarg, K. Schöps in S. Rönnebeck , ur. *Research on PISA*, str. 71–85. Springer Netherlands.

Michael West. 1953. *A General Service List of English Words*. London: Longman, Green and Co.

Gun-yi Xue, Paul Nation. 1984. A University Word List. *Language Learning and Communication* 3**:** 215–229.