Merjenje berljivosti japonščine kot tujega jezika na korpusu učbeniških besedil

Kristina Hmeljak Sangawa

Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana, Slovenija
kristina.hmeljak@ff.uni-lj.si

Povzetek

Samodejno merjenje jezikovne težavnosti je koristen postopek, ko pišemo, prirejamo ali izbiramo besedila za manj sposobne bralce tujega jezika. Po vzoru postopkov, ki so bili razviti za merjenje berljivosti angleških besedil, sta v zadnjih letih nastali dve orodji za merjenje berljivosti japonskih besedil za materne govorce. V preizkusu teh orodij na korpusu besedil iz učbenikov japonščine kot tujega jezika je bilo ugotovljeno, da sta ti orodji le delno uporabni za tuje bralce japonščine. Na istem korpusu sta bila preverjena dva kazatelja berljivosti: ugotovljeno je bilo, da je povprečna dolžina povedi preprost, a učinkovit pokazatelj besedilne težavnosti, razmerje med različnicami in pojavnicami znakov pa nekoliko manj zanesljiv pokazatelj.

Assessment of the readability of Japanese as a foreign language on a textbook corpus

Automatic measurement of readability is a useful procedure when writing, editing or selecting texts for weak readers. Drawing on research on English readability, some formulas and tools have been developed recently for the assessment of Japanese text readability for native speakers of Japanese. Three such tools were applied to a corpus of textbooks of Japanese as a foreign language, and found to be only partially useful in predicting readability for foreign learners. Two indicators of readability were tested on the same corpus; average sentence length was found to strongly correlate with nominal text difficulty in the textbook corpus, while character type-to-token ratio was found to have a weaker correlation.

1. Uvod

Branje je ena od osnovnih dejavnosti, na katerih sloni učenje tujega jezika, in lahko najbolj učinkovito pripomore k razvoju bralčevih jezikovnih sposobnosti takrat, ko je berilo ne samo zanimivo, ampak tudi razumljivo, torej primerne težavnosti za bralca, ki se jezika šele uči.

Bralno gradivo primerne težavnosti za začetne bralce tujega jezika učitelji običajno pripravijo tako, da sestavijo ali priredijo besedila in pri tem uporabijo samo jezikovne strukture in besedišče, ki ga ciljni bralci obvladajo, gradivo za bolj izkušene bralce pa pogosto tako, da izberejo primerno težka besedila na osnovi objektivnih ali subjektivnih kriterijev. Oba postopka se že od nekdaj izvajata ročno, k obema pa lahko veliko pripomorejo jezikovne tehnologije.

Pri pisanju in prirejanju besedil so koristna orodja za besedilno analizo, ki izpostavijo jezikovne strukture in besedišče nad določeno stopnjo težavnosti (npr. vse besede ali slovnične vzorce, ki niso vključeni v dani seznam že predelane učne snovi). Pri izbiri besedil za določeno stopnjo jezikovne sposobnosti pa je koristna splošna ocena težavnosti besedila, na osnovi katere množico razpoložljivih besedil razvrstimo po težavnosti in iz nje izberemo besedilo želene stopnje.

Glede na ogromno količino besedil, ki so na spletu dostopna v digitalni obliki, lahko namreč zlahka pridobimo veliko število besedil na izbrano temo, izbor primerno težkega besedila pa je zamudno opravilo, pri katerem je samodejna ocena berljivosti (težavnosti) besedila lahko v veliko pomoč tako učitelju kot tudi učencu tujega jezika.

Zato da lahko besedila razvrstimo po stopnji težavnosti, potrebujemo lestvico težavnosti in metodo, s katero besedila razvrstimo po tej lestvici. Za merjenje berljivosti angleških besedil obstaja že veliko postopkov, podprtih z obsežnimi raziskavami. Tudi za ocenjevanje stopnje berljivosti japonščine kot maternega jezika obstaja nekaj formul in orodij, toda o berljivosti japonščine kot tujega jezika je bilo doslej le malo raziskav.

Kot prvi korak k ocenjevanju berljivosti japonščine kot tujega jezika smo zato zgradili korpus besedil iz učbenikov japonščine kot tujega jezika in na njem preizkusili obstoječa orodja in dva od najbolj pogosto uporabljanih kazateljev berljivosti.

V prispevku so v drugem razdelku predstavljeni dosedanji pristopi k strojnemu ocenjevanju berljivosti, vključno z merjenjem berljivosti japonščine. V tretjem razdelku je predstavljen poskus uporabe orodij, ki so bila razvita za merjenje berljivosti japonščine kot maternega jezika, na korpusu učbenikov japonščine kot tujega jezika. V četrtem razdelku so prikazani rezultati meritev dveh preprostih, a robustnih kazateljev berljivosti na istem korpusu, v zadnjem razdelku pa so prikazane možnosti za nadaljnje delo.

2. Postopki za merjenje berljivosti

Branje je kompleksen proces, na razumevanje branega besedila pa vpliva veliko dejavnikov, ki jih moramo upoštevati, ko ocenjujemo berljivost določenega besedila: samo besedilo in njegove sestavine (znaki, besede, struktura, pomen), bralec in njegovo znanje (jezikovno, kulturno in širše) ter cilj oziroma namen branja.

2.1. Pristopi k merjenju berljivosti

Pri merjenju berljivosti oziroma težavnosti besedil v dosedanjih raziskavah lahko ločimo dva glavna pristopa.

Prvi pristop je statistična jezikovna analiza, pri kateri se različni merljivi dejavniki v besedilu (npr. povprečna dolžina povedi ali besed) izmerijo na korpusu besedil z vnaprej znanimi oz. določenimi stopnjami težavnosti, nato pa se tisti dejavniki, ki se izkažejo za učinkovite kazatelje berljivosti, na osnovi regresijske analize sestavijo v formulo, v kateri ima vsak dejavnik primerno utežitev.

Drugi pristop je jezikovno modeliranje, pri katerem se - ravno tako na osnovi korpusa, ki je sestavljen iz podkorpusov različnih vnaprej določenih težavnostnih stopenj - izoblikuje serija jezikovnih modelov z rastočo stopnjo

težavnosti; pri oceni berljivosti poljubnega besedila se potem to besedilo primerja z vsemi modeli s pomočjo klasifikatorjev in oceni s stopnjo težavnosti modela, ki je besedilu najbližji.

2.2. Referenčne lestvice za merjenje berljivosti

Pri obeh pristopih je prvi korak določitev težavnostne lestvice in izoblikovanje referenčnega korpusa, na osnovi katerega se nato ali razvije formula, ki vsebuje različne dejavnike berljivosti, ali zgradijo jezikovni modeli, ki nato služijo za kategoriziranje besedil.

Referenčna lestvica, ki se najbolj pogosto zasledi v literaturi o merjenju berljivosti za bralce maternega jezika, je zelo intuitivna lestvica: to so šolski razredi oziroma število let šolanja v določenem jeziku (npr. od 1. razreda osnovne šole do zadnjega razreda gimnazije ipd.), saj naj bi ti odgovarjali jezikovni in splošneje kognitivni razvojni stopnji bralcev. Referenčni korpus pri raziskavah, ki uporabljajo to lestvico, je pogosto zbirka besedil iz učbenikov za vsak razred šolanja. Ta izbor sloni na predpostavki, da so strokovnjaki, ki so izdelali učbenike, na podlagi svojih izkušenj uporabljali jezik, ki je primeren za določeno starostno stopnjo in šolski razred. Pri merjenju težavnosti za tuje bralce smo zasledili samo en podoben pristop k določanju težavnostne lestvice (François 2009), ki uporablja Skupni evropski referenčni okvir za jezike (CEFR).

Druga možna referenčna lestvica je odstotek populacije, ki je sposobna razumeti določeno besedilo. Pri tem pristopu se najprej pripravi zbirka različnih besedil, s testiranjem na reprezentativnem vzorcu populacije govorcev določenega jezika se nato ugotovi, kolikšen odstotek bralcev besedilo razume, na osnovi tega pa se besedila razporedijo v poljubno število stopenj.

Tretji, zelo redek pristop (Sato 2011), je uporaba reprezentativnega korpusa, pri čemer se vsa besedila v korpusu, ki naj bi bil reprezentativna slika vseh besedil v določenem jeziku, s pomočjo obstoječega orodja razporedijo po težavnosti in nato razdelijo v izbrano število stopenj (podkorpusov), ki služijo kot referenčna lestvica.

2.3. Merjenje berljivosti v angleščini

Merjenje berljivosti angleških besedil ima verjetno najdaljšo in najobsežnejšo tradicijo raziskovanja. Od leta 1923, ko sta Lively in Pressy predstavili prvo formulo za merjenje berljivosti, je bilo razvitih na stotine formul in postopkov za ocenjevanje berljivosti (DuBay 2004, 2006). Lastnosti besedil, ki se najbolj pogosto omenjajo v teh raziskavah, so povprečno število besed v povedi, povprečno število zlogov v besedah, povprečno število črk v besedah, odstotek besed iz seznamov osnovnega besedišča, idr.

Pri prvih formulah so se lastnosti besedila (število besed v povedi ipd.) štele ročno, zato so vsebovale samo lahko merljive dejavnike. S pojavom jezikovnih tehnologij pa so se pri merjenju berljivosti začele uporabljati tudi druge, bolj zahtevne analize, kot npr. bolj podrobna analiza besedišča na osnovi obsežnih seznamov pogostosti posameznih besed ali večbesednih enot v velikih korpusih (Anagnostou in Weir 2007), analiza skladenjske strukture s pomočjo samodejnih skladenjskih analizatorjev, merjenje razmerja med različnicami in pojavnicami, merjenje leksikalne kohezije, diskurzne analize ipd. (Graesser idr. 2004, Barzilay in Lapata 2008, Pitler in Nenkova 2008). Novejše raziskave, ki uporabljajo jezikovne modele n-gramov in klasifikatorje za ocenjevanje berljivosti besedil, so se izkazale za robustne in posebej primerne za spletna besedila z nepopolnimi povedmi (Collins-Thompson in Callan 2004, Schwarm in Ostendorf 2005, Heilmann idr. 2007, Feng idr. 2010)

2.4. Merjenje berljivosti v japonščini

Na Japonskem je Morioka leta 1952 objavil verjetno prvo raziskavo o berljivosti japonščine po vzoru formul za ugotavljanje berljivosti angleščine. Temu je sledilo nekaj drugih predlogov formul (Sakamoto 1962, Tateishi idr. 1988, Pichl in Narita 2007) oziroma dejavnikov berljivosti, kot npr. gostote besedišča (Sano in Maruyama 2008), toda nobena od predlaganih formul se ni širše uporabila za merjenje berljivosti japonščine. Šele v zadnjih letih sta dve skupini raziskovalcev razvili dve metodi za merjenje berljivosti besedil za bralce japonščine kot maternega jezika. Prvo metodo (RRL) je razvila Shibasaki s sodelavci (Shibasaki idr. 2008, 2010) z uporabo regresijske analize na korpusu učbenikov japonskega jezika za japonske otroke v dvanajstih razredih od prvega razreda osnovne šole do tretjega (zadnjega) razreda gimnazije. Drugo metodo je razvil Sato s sodelavci (Sato idr. 2008, Sato 2011), ki pa uporablja jezikovno modeliranje: v prvi metodi (Obi T13) ustvari jezikovne modele unigramov znakov v besedilih učbenikov za vse predmete v 12 razredih japonskih šol; v drugi metodi (Obi B9) pa so jezikovni modeli bigramov znakov oz. pismenk v besedilih osnovani na uravnoteženem korpusu japonskega pisanega jezika (BCCWJ, Maekawa idr. 2010). Pri določanju stopnje težavnosti oz. berljivosti določenega besedila, orodji Obi T13 in Obi B9 ocenita težavnostno stopnjo besedila tako, da ga primerjata s tako izdelanimi jezikovnimi modeli. Tako RRL kot Obi sta dostopni na strežnikih obeh raziskovalnih skupin.

Objavljena so tudi poročila o nekaterih raziskavah o berljivosti japonščine kot tujega jezika (Kawamura 1999a, Kitamura idr. 2009, Yamura-Takei idr. 2005, Mizushima idr. 2011, Nakamura idr. 2012, Yoshihashi idr. 2007, Mizuno idr. 2008, Hazelbeck in Saito 2009), toda razen metode, ki jo predstavljajo Kawamura, Kitamura in sodelavci in ki meri samo težavnost besedišča v določenem besedilu, ostale raziskave ne nudijo še dostopne in zaključene metode za celovito oceno berljivosti japonščine kot tujega jezika.

3. Korpus besedil iz učbenikov japonščine kot tujega jezika

Kot prvi korak k preverjanju uporabnosti obstoječih orodij, ki so bili razviti za materne govorce japonščine, tudi za merjenje berljivosti japonščine kot tujega jezika, smo sestavili korpus besedil iz petih različnih serij učbenikov japonščine kot tujega jezika. Medtem ko je pri določanju težavnostnih stopenj za materne govorce možna razporeditev besedil po starosti ali šolskem razredu bralcev, ki besedila razumejo, pa za učence japonščine kot tujega jezika taka lestvica ne obstaja. Za druge namene se sicer pogosto uporabljajo smernice standardnega izpita iz znanja japonščine kot tujega jezika (Japanese Language Proficiency Test), ki pa je od leta 2010 prešel s štirih na pet stopenj znanja in - po vzoru Skupnega evropskega referenčnega okvira za učenje, poučevanje in ocenjevanje tu-

jih jezikov (CEFR) - prešel z opisa stopenj s seznami besedišča in slovničnih struktur, ki so predvidene za vsako stopnjo, na bolj opisne ocene sposobnosti (npr. razumeti predavanja o strokovnih temah pri naravni hitrosti ipd.) , ki naj bi jih bili tuji učenci japonščine sposobni na vsaki stopnji, brez seznamov struktur in besedišča, tako da je težko uporaben kot referenčna lestvica za strojno merjenje jezikovnih dejavnikov.

Glede na to, da se v različnih jezikovnih programih, izobraževalnih ustanovah in učbeniških založbah uporabljajo različne lestvice težavnosti za japonščino, nismo sestavili korpusa s homogenimi podkorpusi npr. začetniške, nadaljevalne in izpopolnjevalne japonščine, pač pa smo za prvo preverjanje uporabnosti obstoječih orodij zbrali pet zbirk učbenikov za tuje učence japonščine, ki vsebujejo vsaka po nekaj knjig postopoma rastoče težavnosti, nato pa primerjali ocene berljivosti obstoječih orodij znotraj vsake posamezne zbirke, saj je zelo verjetno, da so se pri vsaki zbirki učbenikov dosledno uporabljala ista načela za razporejanje besedil po težavnosti. Zbrali smo besedila iz spodaj navedenih učbenikov.

1) *Japanese Graded Readers - Yomu yomu bunko* (2005-2009) založbe Ask, zbirka postopnih lahkih beril za začetnike in nadaljevalce na 4 nivojih (v nadaljevanju JGR1 do JGR4); v besedilu vsakega nivoja se uporabljajo samo besede in slovnični vzorci iz strogo določenega nabora besed in vzorcev, ki je osnovan na petih priljubljenih učbeniških serijah in delno sovpada s seznami JLPT.

2) Učbeniki zavoda Bunka Institute of Language *Shin bunka shokyû nihongo II* (2000), *Bunka chûkyû nihongo I, II* (1994, 1997) založbe Bonjinsha, v nadaljevanju BunA2, BunB1 in BunB2. To je serija učbenikov, ki vključuje vaje za branje, pisanje, poslušanje in govorjenje, zato smo za analizo uporabili le berila in dialoge za vajo iz branja.

3) Serija učbenikov Centra za tuje študente Univerze Sanno *Enjoyable Task Reading in Japanese - Nihongo o tanoshiku yomu hon: Pre-intermediate - Shochûkyû* (1996)*, Intermediate - Chûkyû* (1991) in *Pre-advanced - Chûjôkyû* (1993) založbe Univerze Sanno, v nadaljevanju Joy1, Joy2 in Joy3.

4) Serija učbenikov avtorice Nobuko Mizutani *Sôgô nihongo - Introduction to Japanese - An Integrated Course,* založbe Bonjinsha: *Introduction to Intermediate Japanese - Shokyû kara chûkyû e* (1990), *Intermediate Japanese - First Semester - Chûkyû zenki* (1989) ter *Intermediate Japanese - Chûkyû* (1987); druga in tretja knjiga se vsaka delita na dve seriji osnovnih in nadgrajevalnih beril, v nadaljevanju Mizu1, Mizu2a, Mizu2b, Mizu3a in Mizu3b.

5) Učbeniki Tokijske univerze za tuje jezike (Tokyo University of Foreign Studies) *Shokyû nihongo* (1994), *Chûkyû nihongo* (1994) in *Jôkyû nihongo* (1998), v nadaljevanju TUFS-Ad (dialogi), TUFS-Ar (berila), TUFS-B in TUFS-C.

Vse učbenike oziroma berila smo preslikali, obdelali s programom za razpoznavanje znakov, ročno preverili in popravili, normalizirali in vsako besedilo vsakega učbenika shranili v ločenih datotekah.

Obseg korpusa, število vsebovanih besedil in povprečna dolžina besedil v vsakem učbeniku (po številu znakov, kot je običajno pri merjenju dolžine japonskih besedil) so predstavljeni v Tabeli 1.

	Učbenik	št. besedil	št. znakov	povprečno št. znakov/besedilo
1)	JGR1	17	23569	1386
	JGR2	15	44039	2936
	JGR3	12	52413	4368
	JGR4	11	83967	7633
2)	BunA2	19	40540	2134
	BunB1	21	17327	825
	BunB2	15	18697	1246
3)	Joy1	17	18794	1106
	Joy2	9	10541	1171
	Joy3	9	16764	1863
4)	Mizu1	15	8925	595
	Mizu2a	12	12795	1066
	Mizu2b	12	5332	444
	Mizu3a	12	11927	994
	Mizu3b	12	5000	417
5)	TufsAd	28	25471	910
	TufsAr	10	7393	739
	TufsB	21	26662	1270
	TufsC	10	28140	2814

Tabela 1: Obseg korpusa učbenikov japonščine kot tujega jezika

4. Aplikacija obstoječih orodij za merjenje berljivosti na korpusu

Na zgoraj opisanem korpusu učbenikov smo preverjali učinkovitost obstoječih metod za ocenjevanje težavnosti japonščine kot maternega jezika, ki smo jih omenili v prejšnjem razdelku:

- Obi T13, ki oceni besedilo na lestvici 13 razredov od osnovne šole do prvega leta univerze (Sato idr. 2008);
- Obi B9, ki oceni besedilo na lestvici devetih stopenj, določenih na osnovi porazdelitve besedil v reprezentativnem korpusu japonskega jezika BCCWJ (Sato 2011),
- metodo RRL, ki podobno kot T13 ocenjuje besedila na lestvici šolskih razredov, toda samo na lestvici prvih devet razredov obveznega šolanja (Shibasaki idr. 2010).

Z vsemi tremi orodji smo izmerili predvideno stopnjo berljivosti besedil v zgoraj opisanem korpusu učbenikov japonščine kot tujega jezika.

4.1. Rezultati in analiza meritev

Primerjava ocen berljivosti besedil iz različnih serij učbenikov, kot smo omenili v razdelku 3, ni smiselna, zato smo primerjali nominalni rang učbenikov znotraj iste serije, kot ga predvidevajo avtorji učbenika sami, z rangi ocen treh uporabljenih orodij. S pomočjo spletnega orodja (Wessa 2010) smo izračunali Spearmanovo korelacijo rangov ρ (Kendall 1970), ki nakazuje ujemanje med razporeditvijo besedil po težavnostni stopnji v seriji sami in razporeditvijo besedil, kot jo ocenjujejo tri orodja. Podatek o korelciji rangov je podan samo kot vodilo pri oceni primernosti, saj gre za majhne razpone rangov.

Kot je razvidno iz tabele 2, metodi Obi T13 in Obi B9 pravilno razporedita večino serij, se pa ne ujemata z lestvico težavnosti zbirke Mizutani. Metoda RRL se ravno tako ujema pri nekaterih serijah, ponuja pa posebej presenetljiv rezultat pri zbirki JGR, ki vsebuje najlažja besedila, kjer je korelacija nakazana celo kot negativna.

Učbenik	Obi T13	ρ	Obi B9	ρ	RRL	ρ
JGR1	5	0.8	1	0.8	6.5	-0.65
JGR2	6		1		6.5	
JGR3	6		1		6	
JGR4	6		3		6.4	
BunA2	5	1	2	1	6.5	1
BunB1	6		4		6.8	
BunB2	9		5		6.9	
Joy1	6	1	1	1	5.9	1
Joy2	8		4		6.6	
Joy3	9		5		7	
Mizu1	9	0.5	5	0.5	6.5	0.9
Mizu2a	9		5		7	
Mizu2b	9		5		7.5	
Mizu3a	9		5		7.3	
Mizu3b	9		5		8.7	
TufsAd	3	1	1	0.95	4.2	0.9
TufsAr	5		1		4.2	
TufsB	9		5		7.2	
TufsC	10		7		7.2	

Tabela 2: Obijeva in RRL-jeva ocena berljivosti besedil v korpusu učbenikov japonščine kot tujega jezika

Kot možen razlog za to neujemanje se ponuja dejstvo, da lahko tako na obe metodi Obi, ki sta osnovani na modelu unigramov oz. bigramov znakov v besedilih, kot tudi na metodo RRL, ki ravno tako preverja uporabljane pismenke, zelo vpliva izbor znakov pri zapisu japonščine.

Za zapis japonščine se namreč uporabljajo trije nabori znakov: dve zlogovnici (hiragana in katakana), pri katerih vsak od 46 znakov zapisuje en zlog, ter približno 2000 kitajskih pismenk, pri katerih vsak znak zapisuje eno besedo ali morfem. V standardnem zapisu japonskih besedil se za skoraj vse polnopomenske besede uporabljajo kitajske pismenke, zlogovnici pa samo za obrazila, funkcijske besede in tujke, toda v učbenikih in besedilih za otroke, ki se šele učijo branja, se v vsakem razredu uporabljajo samo točno določene pismenke, ki se jih po veljavnih smernicah japonskega ministrstva za šolstvo otroci naučijo v vsakem razredu. V besedilih za 1. razred osnovne šole se tako uporabljata zlogovnici ter prvih 80 pismenk, v besedilih za 2. razred znaki iz 1. razreda in dodatno še 160 pismenk iz standardnega seznama itd. Vse besede, ki se v standardnem pisanju zapišejo s kitajskimi pismenkami, ki jih otroci v določenem razredu še ne spoznajo, se v tako prirejenih besedilih zapišejo fonetično, z zlogovnico hiragana, brez uporabe pismenk, ki jih otroci ne poznajo.

Pri besedilih, ki so pisana za japonske otroke, je torej že iz nabora znakov, ki jih besedilo uporablja, dokaj jasno razvidno, za kateri razred so napisana. Zato je verjetno, da vse tri metode ocenjevanja berljivosti, ki slonijo na merjenju vsebovanih znakov, uspešno ugotovijo, za kateri razred je besedilo napisano, če je napisano po standardnih navodilih za japonska šolska besedila.

V korpusu učbenikov japonščine kot tujega jezika se v seriji TUFS uporablja podoben sistem zapisovanja, po katerem so besedila v učbeniku za začetnike zapisana z omejenim naborom kitajskih pismenk, učbenik za srednjo stopnjo s postopno več pismenkami, učbenik za višjo stopnjo pa v standardnem zapisu. To je možen razlog za uspešno razvrstitev besedil po metodah Obi in RRL. Po drugi strani pa so v zbirki JGR besedila zapisana v standardnem zapisu, ob vseh pismenkah pa je z zlogovnico še dodan glasovni zapis (furigana). Pri strojni obdelavi besedil smo dodatne glasovne zapise odstranili, ker drugače obdelava ni mogoča, zato so se v besedilih pojavile pismenke, ki bi bile verjetno težko berljive za začetnike, ki ne poznajo kitajskih pismenk, ko ne bi bilo zraven glasovnega zapisa, dejansko pa so besedila izredno berljiva in razumljiva tudi začetnikom, kot smo lahko ugotovili med lastnim poukom japonščine za začetnike.

To hipotezo smo preverili tako, da smo ista besedila iz zbirke JGR prepisali po kriterijih za zapis besedil za prvih šest razredov osnovne šole in ugotovili, da obe orodji ocenjujeta isto besedilo kot bistveno bolj berljivo, če je zapisano v zlogovnici, kot pa če je zapisano s pismenkami.

Glede na to, da obstaja več spletnih orodij, s katerimi lahko poljubnemu besedilu v japonščini samodejno dodamo zapis glasovnih vrednosti v zlogovnici, ali samodejno obdelamo besedilo tako, da se izpiše v zlogovnici, je potem pri branju besedil v elektronski obliki uporaba pismenk namesto zlogovnice ali obratno pravzaprav nepomembna. Iz tega lahko sklepamo, da so obstoječa orodja za ocenjevanje berljivosti japonščine kot maternega jezika samo delno uporabna pri ocenjevanju japonščine kot tujega jezika.

5. Merjenje dejavnikov berljivosti v učbenikih japonščine kot tujega jezika

Kot prvi korak k razvoju formule za merjenje berljivosti japonskih besedil za bralce japonščine kot tujega jezika smo na istem korpusu učbenikov preverili dve izmed lastnosti besedil, ki so se doslej uporabile v merjenju berljivosti japonščine.

5.1. Jezikovni dejavniki berljivosti japonščine

V dosedanjih raziskavah o berljivosti japonščine so se kot kazatelji berljivosti merile naslednje merljive lastnosti japonskih besedil:

- pri zapisu: razmerje med številom znakov različnih naborov, t.j. hiragane, katakane, kitajskih pismenk in latinice (Morioka 1952, Sakamoto 1964, Tateisi idr. 1988, Pichl in Narita 2007); delež kitajskih pismenk različnih stopenj po izbranih referenčnih seznamih (Pichl in Narita 2007 uporabljata seznam pismenk japonskega Ministrstva za šolstvo za pouk v osnovni in srednji šoli, Kawamura 1998 uporablja seznam JLPT); delež pismenk z več kot 15 potezami (Pichl in Narita 2007); dolžina sosledij znakov iz istega nabora (Tateisi idr. 1988); razmerje med številom ločil in drugih znakov (Morioka 1952); razmerje med številom pik in vejic kot kazatelj skladenjske zapletenosti

(Tateisi idr. 1988); delež besed z dodanim glasovnim zapisom kot kazatelj težavnosti besedišča (Morioka 1952);

- pri besedišču: delež besed iz različnih stopenj referenčnih seznamov, kot npr. seznam pogostosti (Sakamoto 1962, Kawamura idr. 2008), tf*idf glede na časopisni korpus (Kitamura idr. 2009), seznam besedišča za JLPT (Kawamura 1998), domačnost (Kawamura 2008); delež funkcijskih besed iz seznama JLPT (Mizuno idr. 2008); delež besed japonskega, kitajskega in drugega izvora, delež okrajšav, delež lastnih imen, delež abstraktnih besed v vlogi osebka (Morioka 1952); besediščna gostota po Hallidayevi formuli, t.j. razmerje med številom polnopomenskih besed in številom povedkov (Sano in Maruyama 2008);

- pri skladnji: dolžina povedi (Morioka 1952, Sakamoto 1962, Tateisi idr. 1988, Oono in Inazumi 2007, Shibasaki 2008); dolžina stavkov (Morioka 1952); število povedkov na poved (Shibasaki 2008); delež prisamostalniških odvisnikov (Mizushima idr. 2011); delež neizraženih udeleženskih vlog (Yamura-Takei 2008) ali neizraženih osebkov (Nakamura idr. 2012).

- na ravni besedila: dolžina besedila (Morikawa idr. (2010) omenjajo dolžino 10000 znakov kot najdaljšo sprejemljivo za nadaljevalne učence); delež premega govora (Morioka 1952);

- na ravni sloga: delež povedkov v neformalnem in formalnem slogu (Morioka 1952); delež pogovornih členkov in polnil (Morioka 1952).

Izmed teh smo izbrali dve najbolj osnovni in zanesljivo merljivi: povprečno dolžino povedi in razmerje med različnicami in pojavnicami (TTR) znakov. Povprečna dolžina povedi se je izkazala kot najbolj zanesljiv med merljivimi kazatelji težavnosti v angleških besedilih (Feng idr. 2010), kar se tudi ujema z raziskavami o pomnjenju povedi (Goetz idr. 1981), TTR znakov pa je lahko dober kazatelj pestrosti in posledično težavnosti besedišča.

Povprečno dolžino povedi smo izmerili tako, da smo v besedilih prelomili vrstice po vseh ločilih, ki označujejo konec povedi (。 ？ ！」 』) in ročno preverili prelom vrstic pri oklepajskih znakih 「 in 『, ki se lahko uporabljata tudi sredi povedi. Nato smo prešteli število vrstic v vsakem besedilu (vključno z vrsticami naslovov, ki niso vsebovale ločil, a so bile že v prvotnih besedilih ločene od ostalega besedila) ter delili število znakov v besedilu s številom vrstic. Dolžino smo merili po številu znakov in ne besed, ker je tokenizacija (deljenje na besede) pri japonščini, ki se zapisuje brez presledkov med besedami, lahko vir napak pri štetju, še posebej pri sestavljenih besedah, ki jih različna orodja za tokenizacijo različno obravnavajo.

Izmerili smo tudi razmerje med različnicami in pojavnicami znakov (TTR). Tudi tu smo zaradi možnih napak pri tokenizaciji merili TTR znakov raje kot TTR besed. Ker na to razmerje močno vpliva dolžina besedila, smo dolžino besedil normalizirali na dolžino najkrajšega besedila v korpusu, ki je štelo 270 znakov, tako da smo črpali začetnih 270 znakov vsakega besedila. Nato smo na vsakem vzorcu izmerili število različnic in pojavnic, izračunali razmerje med njima ter izračunali povprečje teh razmerij v vsakem učbeniku.

Rezultati analize so povzeti v tabeli 3. Tudi tu je Spearmanova korelacija rangov ρ izračunana samo informativno. Kot je razvidno iz tabele, povprečna dolžina povedi v vseh serijah učbenikov raste vzporedno s stopnjo težavnosti učbenika. Tudi pri besedilih v učbenikih japonščine kot tujega jezika se je torej dolžina povedi izkazala za preprosto, a obenem učinkovito mero besedilne težavnosti.

Po drugi strani TTR manj učinkovito ocenjuje stopnje težavnosti v učbenikih. Razlog za to je lahko neprimerni način meritve (prekratki vzorci besedil), lahko izbor besedil samih - torej neprimernost učbeniške serije kot referenčne lestvice za merjenje težavnosti besedil, ali pa neprimernost te mere za razvrščanje besedil v japonščini kot tujem jeziku.

	Učbenik	povprečno št. znakov/poved	ρ	TTR znakov	ρ
1)	JGR1	14.8	1.0	0.28	1.0
	JGR2	19.9		0.31	
	JGR3	21.9		0.33	
	JGR4	25.9		0.38	
2)	BunA2	20.6	1.0	0.38	0.87
	BunB1	30.1		0.41	
	BunB2	37.3		0.41	
3)	Joy1	25.8	1.0	0.35	1.0
	Joy2	28.5		0.37	
	Joy3	36.6		0.43	
4)	Mizu1	32.9	1.0	0.37	0.82
	Mizu2a	35		0.42	
	Mizu2b	39.4		0.48	
	Mizu3a	39.7		0.42	
	Mizu3b	33		0.51	
5)	TufsAd	22.4	1.0	0.28	1.0
	TufsAr	30.8		0.31	
	TufsB	37.1		0.39	
	TufsC	44.8		0.42	

Tabela 3: Kazatelji berljivosti v korpusu učbenikov japonščine kot tujega jezika

6. Zaključek

V prispevku smo predstavili korpus besedil iz učbenikov japonščine kot tujega jezika, ki lahko služi kot osnova za izdelavo formule za merjenje berljivosti japonskih besedil za tuje bralce. Ugotovljeno je bilo, da so obstoječe mere berljivosti za rojene govorce japonščine le delno uporabne za ugotavljanje berljivosti japonščine kot tujega jezika. Na korpusu je bila preverjena koristnost dveh preprostih kazateljev merljivosti, od katerih se je povprečna dolžina povedi izkazala za koristno mero. Analiza drugih besedilnih lastnosti, ki lahko služijo pri samodejnem ocenjevanju besedilne težavnosti, bo predmet naših prihodnjih raziskav.

7. Literatura

Anagnostou, N. in Weir, G., 2007. Average Collocation Frequency as an Indicator of Semantic Complexity. V *Proceedings of ICTATLL 2007*.

Barzilay, R. in Lapata, M., 2008. Modeling local coherence: An entity-based approach, *Computational Linguistics* 34:1, 1-34.

Collins-Thompson, K. in Callan, J., 2004. A Language Modeling Approach to Predicting Reading Difficulty. V *HLT-NAACL 2004 Main Proceedings*, Boston: ACL. 193--200.

DuBay, W., 2004. *The principles of readability*. Costa Mesa: Impact Information.

DuBay, W., 2006. *The Classic Readability Studies*. Costa Mesa: Impact Information.

Feng, L., Jansche, M., Huenerfauth, M. in Elhadad, N., 2010. A Comparison of Features for Automatic Readability Assessment. V *23rd International Conference on Computational Linguistics, Poster Volume*. Beijing: COLING. 276--284.

François, T., 2009. Combining a Statistical Language Model with Logistic Regression to Predict the Lexical and Syntactic Difficulty of Texts for FFL. V *Proceedings of the Student Research Workshop at EACL 2009*, Athens: ACL.

Goetz, E., Anderson, R. in Schallert, D.. 1981. The representation of sentences in memory, *Journal of Verbal Learning & Verbal Behavior.* 20:4. 369-385.

Graesser, A., McNamara, D., Louwerse, M. in Cai, Z., 2004. Coh-Metrix: Analysis of text on cohesion and language, *Behavior Research Methods, Instruments, & Computers* 36:193-202.

Hazelbeck, G. in Saito, H., 2009. A Corpus-based E-learning System for Japanese Vocabulary, *Information and Media Technologies* 4:4. 1104-1128.

Heilman, M., Collins-Thompson, K., Callan, J. in Eskenazi, M., 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. V *Proceedings of the Human Language Technology conference*, Rochester, NY: ACL. 460--467.

Kawamura, Y., 1999. Kanji no nan'ido hantei shisutemu 'Kanji checker' o mochiita tekisuto no bunseki - Analysis of Japanese Textbook Using the``Kanji Level Checker". *Tokyo kokusai daigaku ronsô.* 59:73-87.

Kendall M.G., 1970. *Rank correlation methods*. London: Griffin

Kitamura, T., Tomioka, Y. in Kawamura, Y., 2009. Development and evaluation of a word level rating system based on inverse document frequency. *JLEM* 16:1. 52-53.

Maekawa, K., Yamazaki, M., Maruyama, T., Yamaguchi, M., Ogura, H., Kashino, W., Ogiso, T., Koiso, H. in Den, Y., 2010. Design, compilation, and preliminary analyses of Balanced Corpus of Contemporary Written Japanese. *Proceedings of the 10th LREC*, Valletta: LREC. 1483-1486.

Mizuno, S., Ooyama, H., Kobayashi, T. idr. 2008. Nihongo dokkai shien no tame no gogigoto no yôrei chûshutsu shisutemu no kôchiku. *Kyôiku-gakushû o shien suru gengoshori.* Tokio: NLP. 63-66.

Mizushima, H., Uchida, S., Kitamura, T. in Kawamura, Y., 2011. Gakushûsha ni totte nankai na kôbun no jidôkenshutsu. *JLEM*. 18:1. 64-65.

Morikawa, Y., Nagasu, M., Haruna, H. in Kitamura, T., 2010. Nihongo dokkai gakushû shien saito "tutor.bunko" no kôsô to kaihatsu. *Kônan daigaku jôhô kyôiku kenkyû sentaa kiyô.* 9(3).

Morioka, K., 1952. Yomiyasusa no kisoteki kenkyû. V *Shôwa 26 nendo Kokuritsu kokugo kenkyûjô nenpô*. Tokio: Kokken. 91-108.

Nakamura, K., Kitamaura, T. in Kawamura, Y., 2012. Kensaku enjin o mochiita shukaku shôryakubun no jidô hantei. *JLEM*. 19:1.

Pichl, L. in Narita, J., 2007. Readability Factors of Japanese Text Classification. *Databases in Networked Information Systems*, Berlin: Springer. 132-138.

Pitler, E. in Nenkova, A., 2008. Revisiting readability: a unified framework for predicting text quality. V *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg: ACL. 186--195.

Sakamoto, I., 1962. Jidô yomimono no goihijû no hyôteihô. *Kyôiku shinrigaku nenpô.* 1:107.

Sano, M. and Maruyama, T., 2008. Lexical Density in Japanese Texts: classifying text samples in the Balanced Corpus of Contemporary Written Japanese (BCCWJ). *Proceedings of ISFC 35: Voices Around the World*, Sydney: ISFC. 359-364.

Sato, S., Matsuyoshi, S. in Kondoh, Y., 2008. Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. V *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech: LREC.

Sato, S., 2011. Measuring Text Readability Based on Balanced Corpus, *IPSJ Journal,* 52:4, 1777-1789.

Schwarm, S. in Ostendorf, M., 2005. Reading level assessment using support vector machines and statistical language models. V *Proceedings of the 43rd Annual Meeting*, Stroudsburg: ACL. 523--530.

Shibasaki, H., Tamaoka, K., Yamamoto K., Kanô, M., Hara, S. in Lee, J., 2008. Nihongo koopasu o ôyôshita bunshô no nan'ido sokutei no kenkyû. V *Tokutei ryôiki kenkyû 'Nihongo koopasu' Heisei 19 nendo kôkai waakushoppu.* 125-130.

Shibaraki, H. in Tamaoka, K., 2010. Constructing a formula to predict school grades 1-9 based on Japanese language school textbooks. *Nihongo kyôiku kôgakukai ronbunshi.* 33:4. 449-458.

Tateisi, Y., Ono, Y. in Yamada, H., 1988. A computer readability formula of Japanese texts for machine scoring. V *Proceedings of the 12th conference on Computational linguistics*, Morristown: ACL. 649--654.

Wessa, P., 2012. *Free Statistics Software*, Office for Research Development and Education, v. 1.1.23-r7. [http://www.wessa.net/]

Yamura-Takei, M., Aizawa, T. in Fujiwara, Mo., 2005. Diversity of zeros in Japanese discourse: A corpus analysis and a tool for language teachers. *Proceedings of PACLING 2005*, Tokyo: PACLING. 358-367.

Yoshihashi, K. in Nishina, K., 2007. Japanese composition support system displaying co-occurrences and example sentences. *Symposium on large-scale knowledge resources (LKR2007).* 119-122.