Irena Srdanović

Odabir prikladnih primjera za učenički rječnik: značajke GDEX-a za japanski jezik i mogućnosti uporabe

Izvorni znanstveni rad
Original scientific paper
UDK 811.521'243:37.012>(030)
https://doi.org/10.32728/tab.16.2019.14

SAŽETAK

Primjeri su značajan sastavni dio rječnika i mogu u velikoj mjeri pomoći boljem razumijevanju značenja i uporabe riječi. Stoga je bitan korak u stvaranju rječnika odabrati prikladne primjere, pri čemu nam suvremene napredne tehnologije mogu znatno koristiti. Jedna takva tehnologija je alat GDEX (Good Dictionary EXamples) koji je osmišljen da iz korpusa automatski bira najprikladnije primjere i tako pomaže u leksikografskome postupku odabira primjera za rječnik. Alat GDEX za japanski jezik (engl. *GDEX for Japanese*) dizajniran je korištenjem i nadogradnjom metoda primijenjenih za druge jezike, za koje se je pokazalo da su korisni u leksikografskim projektima. Jedan korak dalje u razvoju tehnologije GDEX je i njegova dodatna prilagodba učenicima japanskoga jezika (engl. *Learner's GDEX for Japanese*) uzimajući u obzir razine leksema temeljene na JLPT popisu vokabulara japanskoga jezika. Ciljevi ovoga istraživanja su (1) opisati osnovne karakteristike alata GDEX za japanski jezik, (2) saznati mišljenja i stavove studenata japanskoga jezika o rezultatima automatski dobivenih primjera iz korpusa pomoću alata GDEX te (3) analizirati i vrednovati dobivene rezultate sa stajališta primjerenosti primjera i korisnosti alata u procesu odabira dobrih primjera za učenički rječnik japanskoga jezika, kao i prihvatljivosti angažiranja studenata u procesu odabira primjera za dvojezični japansko-hrvatski rječnik. Nakon educiranja o korištenju korpusa, alata za pretraživanje korpusa Sketch Engine (SkE) i alata GDEX za japanski jezik, studentima japanskoga jezika na srednjoj jezičnoj razini na preddiplomskome sveučilišnome studijskom programu Japanski jezik i kultura na Sveučilištu Jurja Dobrile u Puli, daje se zadatak vrednovanja primjera i popunjavanja upitnika o korištenju GDEX-a za japanski jezik. Rezultati analize pokazuju da studenti većinu primjera iz rječnika smatraju korisnim za uključivanje u učenički rječnik i uspijevaju

primijetiti i spomenuti neke poteškoće u odabiru primjerenoga primjera za rječnik. Može se zaključiti da se je GDEX za japanski jezik pokazao korisnim kao alat za poluautomatski odabir primjera za ovakvu vrstu rječnika, te da postoji veliki potencijal za uključivanje učenika u leksikografski rad na stvaranju trenutno nastajućega japansko-hrvatskoga rječnika za učenike, kao i za druge dvojezične parove. Prepoznati su i potencijalni problemi i koje je aspekte alata GDEX potrebno u budućnosti poboljšati.

Ključne riječi: rječnik učenika japanskoga jezika, GDEX (dobri primjeri za rječnik) za japanski jezik, Sketch Engine, korpusna leksikografija, automatski odabir primjera, uključivanje učenika u leksikografski rad

1. UVOD

Primjeri su značajan sastavni dio rječnika i mogu u velikoj mjeri pomoći boljem razumijevanju značenja i uporabe riječi. Prema Gouws i dr. (2013) primjeri imaju svoj značaj u bilo kojem tipu rječnika; dok kod monolingvalnih rječnika mogu poslužiti kao nadopuna opisu značenja odnosno definiciji, kod dvojezičnih se rječnika najčešće koriste za razjasniti nejednakosti u značenju i uporabi riječi dvaju jezika. Ponekada je teško razumjeti definiciju u rječniku bez čitanja primjera (Atkins & Rundell 2008), a često su primjeri informativniji i korisniji od same definicije (Frankenberg-Garcia 2012). Stoga je odabir prikladnih primjera jako bitan korak u obradi natuknica u rječniku te će kvalitetniji i prikladniji primjeri pridonijeti boljim podacima do kojih korisnici rječnika mogu doći, što je u skladu s korisnički usmjerenim stavom u suvremenoj leksikografskoj teoriji i praksi (Gouws i dr. 2013; Atkins & Rundell 2008; Prinsloo & Gouws 2000). Kvaliteta i prikladnost primjera ogleda se najčešće u četirima osnovnim kategorijama: prirodnost, tipičnost, informativnost i razumljivost (Atkins & Rundell 2008; Kosem 2019).[1] Prirodnost podrazumijeva da se primjer može očekivati u stvarnoj jezičnoj uporabi. Tipičnost primjera odnosi se na to da se u primjeru vidi tipična upotreba natuknice, npr. u tipičnome kontekstu, strukturi, s tipičnim kolokacijama i slično. Informativnost primjera postignuta je ako primjer u sebi nosi potrebne i dodatne informacije o značenju i uporabi riječi, njezinim kolokacijskim tendencijama u povezivanju s drugim riječima i pomaže razumijevanju definicije i uporabe. Razumljivost podrazumijeva da je primjer jasan i izvan šireg konteksta, da nema previše rijetke i teške riječi u sebi, da nije previše dugačak i samim time teško razumljiv (ibid.).

Pri odabiru primjera bitno je voditi računa o dobrome balansu među navedenim kategorijama, što je vrlo često zahtjevan i mukotrpan posao leksikografa. S razvojem suvremenih tehnologija s jedne su se strane povećale mogućnosti korištenja korpusa i različitih alata za pretraživanje autentičnih

1 Na engleskome kategorije glase: natural, typical, informative and intelligible (Atkins & Rundell 2008: 458-61).

primjera, s druge je strane sam postupak odabira primjera postao otežan zbog ogromne količine podataka, dugačkih i teških primjera, možebitnih nejezičnih elemenata i šuma (engl. *noise*) u podacima. Da bi se prešle te poteškoće i olakšao proces biranja primjera iz korpusa, počelo se je raditi na razvijanju naprednijih tehnologija za automatsko i poluautomatsko biranje prikladnih primjera za rječnike, memorijske kartice i druge namjene uz korištenje različitih kriterija (npr. vidi Nishina & Yoshihashi 2007; Hmeljak-Sangawa i dr. 2012; Tolmachev & Kurohashi 2017; Ljubešić & Peronja 2015). Jedan od takvih alata koji se koristi kao pomoć u odabiru primjera iz korpusa za potrebe rječnika je GDEX (Good Dictionary EXamples) (npr. Kilgarriff i dr. 2008; Kosem i dr. 2011, 2019; Srdanović i Kosem 2016) u okviru alata Sketch Engine (Kilgarriff 2004).

U ovome je radu cilj najprije predstaviti dizajn i karakteristike alata GDEX za japanski jezik (engl. *GDEX for Japanese*) i Učenički GDEX za japanski jezik (engl. *Learner's GDEX for Japanese*) koji su osmišljeni da automatski ponude dobre primjere za rječnik te tako pomognu leksikografima i učenicima japanskoga jezika u odabiru prikladnih primjera za rječnik (Srdanović i Kosem 2016). Alat GDEX za japanski jezik dizajniran je korištenjem i nadogradnjom metoda primijenjenih za druge jezike, za koje se je pokazalo da su korisni u leksikografskim projektima. Jedan korak dalje u razvoju tehnologije GDEX je i njegova dodatna prilagodba učenicima japanskoga jezika (Learner's GDEX for Japanese) uzimajući u obzir razine leksema temeljene na JLPT popisu vokabulara japanskoga jezika (Japanska zaklada i Udruženje za međunarodno obrazovanje u Japanu 2004). Zatim, cilj nam je predstaviti mišljenja i stavove učenika o rezultatima automatski dobivenih primjera japanskoga jezika iz korpusa pomoću alata GDEX. Nakon educiranja o korištenju korpusa, alata za pretraživanje korpusa Sketch Engine (SkE) i alata GDEX za japanski jezik, učenicima japanskoga jezika na srednjoj jezičnoj razini na preddiplomskome sveučilišnome studijskom programu Japanski jezik i kultura na Sveučilištu Jurja Dobrile u Puli daje se zadatak vrednovanja primjera i popunjavanja upitnika o korištenju GDEX-a za japanski jezik. Na kraju je cilj rada analizirati i vrednovati dobivene rezultate sa stajališta primjerenosti primjera i korisnosti alata u procesu odabira dobrih primjera za učenički rječnik japanskoga jezika, kao i prihvatljivosti angažiranja studenata u procesu odabira primjera za dvojezični japansko-hrvatski rječnik.

U usporedbi s drugim pristupima i alatima za automatski odabir primjera (npr. Nishina & Yoshihashi 2007; Hmeljak-Sangawa i dr. 2012; Tolmachev & Kurohashi 2017; Ljubešić & Peronja 2015), za GDEX je karakteristično da je usmjeren prije svega na leksikografsku podršku, a zatim i na učenje jezika. Njegova prednost jest što je ugrađen u već postojeću platformu uz druge potrebne funkcije za korpusno leksikografski rad te je uz pojedinačne dokonfiguracije dostupan za različite jezike. Jedan od pogodnosti korištenja

alata su i mogućnosti prilagodbe konfiguracije u skladu s potrebama različitih jezika i korisnika te mogućnosti nadogradnje s obzirom na napretke u tehnologiji i nove spoznaje o potrebama korisnika i prikladnosti primjera. Što se tiče istovrsnih tehnologija za japanski jezik, Hmeljak-Sangawa i dr. (2012) koriste sličnu metodologiju usmjerenu na učenike japanskoga jezika s ciljem izgradnje korpusa primjera za učenike tako da su primjeri označeni po razinama težine na osnovi podataka JLPT popisa. Rezultati su zatim direktno povezani s dvojezičnim japansko-slovenskim rječnikom, gdje se iz opisa natuknica može poveznicom doći do konkordance više različitih jezičnih razina. Razlika je, osim u već navedenoj širokoj namjeni i specifičnostima alata, i to da GDEX uz podatke JLPT rabi i druge brojne parametre u konfiguraciji i sortiranju prikladnih primjera.

2. GDEX I AUTOMATSKO SORTIRANJE DOBRIH PRIMJERA

2.1. Razvoj GDEX tehnologije

GDEX (Good Dictionary EXamples = dobri primjeri za rječnik) je tehnologija koja se je razvila u okviru računalne podrške za pretraživanje i analizu korpusa u leksikografske svrhe, mrežne platforme Sketch Engine (Kilgarriff i dr. 2004; Kilgarriff i dr. 2008). Osnovni je cilj ove tehnologije podrška leksikografima i drugim korisnicima u odabiru primjera za rječnik rangiranjem rečenica iz korpusa prema tome koliko su prikladne kao kandidati za primjer u rječniku. Takvo rangiranje rečenica vrši se automatski uz pomoć konfiguracije GDEX u kojoj su određene različite sintaktičke i leksičke značajke poput dužine riječi i rečenica, prisustvo ili izostanak određenih riječi npr. izostavljanje deiktičkih kategorija i drugih riječi koje upućuju na kontekst izvan rečenice ili izostavljanje rijetkih riječi, prisustvo ili izostanak određenih simbola, slova, rečeničnih struktura, pozicije ključne riječi u rečenici i slično. Ova je tehnologija najprije bila razvijena za engleski jezik (Kilgarriff i dr. 2008), a zatim se razvijala i za druge jezike, pri čemu je važno prilagoditi parametre konfiguracije u skladu sa specifičnostima određenoga jezika. Osim toga, konfiguraciju je moguće prilagoditi i prema namjerama uporabe, te se je npr. razvijala pojedinačno za određene vrste riječi u slučaju GDEX-a za slovenski jezik (npr. Kosem i dr. 2011; Kosem i dr. 2019) ili za različite jezične razine učenika stranoga jezika u slučaju GDEX-a za japanski jezik (Srdanović & Kosem 2016), što ćemo opisati detaljnije u sljedećemu poglavlju.

Ako tehnologija GDEX nije posebno razvijena i nadograđena za određeni jezik, moguće je upotrebljavati GDEX konfiguraciju s početnim vrijednostima (engl. *Default configuration*) koja se može rabiti za bilo koji jezik (Kilgarriff i dr. 2008). Ova konfiguracija sadrži obvezne parametre koji zahtijevaju da je primjer cijela rečenica, da se ne uključuju primjeri koji sadrže

znakove s crne liste[2] i sadrže riječi s minimalnom frekvencijom. Osim toga sankcionira[3] (engl. *penalizing*) primjere uzimajući u obzir dužinu rečenice, rijetke riječi, slova i simbole u rečenici.[4] Ovako pojednostavljena konfiguracija može biti dobro polazište za stvaranje novih konfiguracija za specifične jezike i namjene.

Pristup GDEX tehnologiji moguć je u sučelju Sketch Enginea u okviru konkordancera[5] (engl. *Concordance*) i skica riječi (engl. *Word Sketches*).[6]

2.2. GDEX konfiguracije za japanski jezik

Srdanović i Kosem (2006) osmislili su više različitih konfiguracija za japanski jezik te detaljno opisali metode dizajna i rezultate evaluacije: 1) GDEX za japanski jezik koji je za opću leksikografsku upotrebu, 2) Učenički GDEX za japanski jezik koji koristi učeniku usmjeren pristup i sadrži više podkonfiguracija ovisno o jezičnoj razini u skladu s JLPT popisom riječi prema razinama 1 – 5, te je tako prilagođen za upotrebu od strane učenika, potrebama učenja ili stvaranju učeničkih rječnika. Za izradu GDEX konfiguracija koristio se je japanski ogledni korpus od tristo milijuna pojavnica[7] JpTenTen11 [SUW, uzorak] koji je dio većega mrežnoga korpusa JpTenTen11 [SUW] od 10 milijardi pojavnica (Pomikálek & Suchomel 2012; Srdanović i dr. 2013). Navedeni korpusi rabe morfološki analizator MeCab s elektroničkim rječnikom UniDic koji se također koriste za segmentaciju i označavanje uravnoteženoga korpusa suvremenoga japanskoga jezika BCCWJ (Maekawa i dr. 2013). Svaki od ovih korpusa koristi se tzv. gramatikom skica (engl. *sketch grammar*) za japanski jezik (Srdanović i dr. 2008, 2013) koja omogućava korištenje gramatičkoga i kolokacijskoga opisa riječi tzv. skica riječi (engl. *Word sketch*) unutar platforme Sketch Engine.

2.2.1. GDEX za japanski jezik (GDEX for Japanese language)

Konfiguracija GDEX-a za japanski jezik u svojoj konačnoj verziji *Japanese-v1u* sastoji se od sljedećih parametara (Srdanović & Kosem 2016):

2 Crna lista je lista riječi koje ako se pojave u primjeru, on se sankcionira, odnosno ne odabire kao kandidat dobroga primjera.

3 Sankcioniranje primjera podrazumijeva niže rangiranje primjera.

4 GDEX konfiguracija je detaljno opisana u Kilgarriff i dr. (2008). Propratna dokumentacija dostupna je na: https://www.sketchengine.eu/documentation/manual-for-gdex/, posjećeno: 12. 11. 2019.).

5 Konkordancer (engl. Concordance) se definira kao „popis riječi sa svim oblicima u kojima se pojavljuju zajedno s kontekstom i oznakom izvora, koji se nalaze u nekome korpusu“ (http://ihjj.hr/mreznik/page/pojmovnik/6/, posjećeno: 12. 11. 2019.).

6 Novo sučelje Sketch Enginea iz 2019. godine zahtijeva posebnu dozvolu korisnika za upotrebu GDEX tehnologije u okviru skica riječi i tzv. funkcionalnosti Tickbox lexicography.

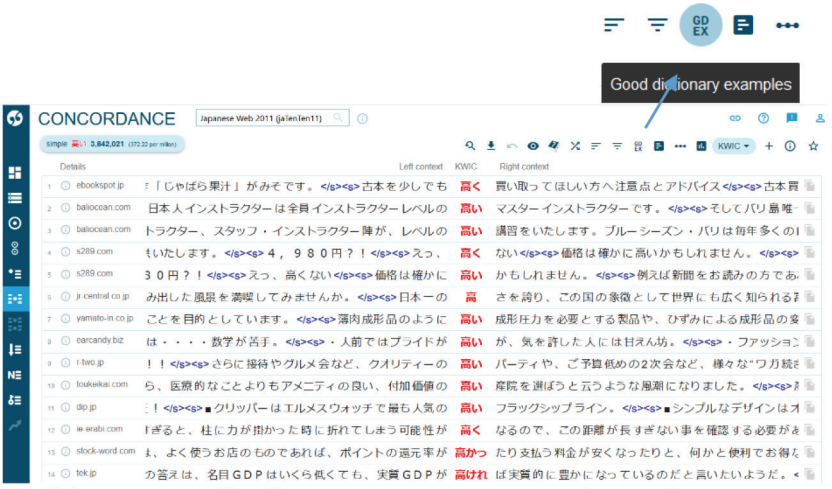
7 Pojavnica (engl. token) je „svaka pojava jezične jedinice u korpusu, na razini riječi svaki oblik uključen u leksem“(http://ihjj.hr/mreznik/page/pojmovnik/6/, posjećeno: 12. 11. 2019.).

- Obvezne značajke primjera: japanska točka ili upitnik koji se pojavljuju samo jednom u rečenici s ciljem da se izbjegnu primjeri koji nisu u obliku rečenice ili sadrže više rečenica.
- Dužina rečenice: poželjna dužina rečenice je od 10 do 25 pojavnica, dok je dozvoljena dužina nešto duža, između 8 i 30 pojavnica.
- Sankcioniranje simbola, znakova i alfabetskoga pisma uz korištenje crne liste i oznaka za simbole.
- Sankcioniranje rijetkih znakova i određenih tipova zagrada.
- Sankcioniranje samo otvorenih ili zatvorenih zagrada u rečenici.
- Sankcioniranje riječi dužih od 7 znakova.
- Sankcioniranje vlastitih imenica uz pomoć oznaka za imena, prezimena, toponime i slično.
- Sankcioniranje lema s manjom učestalošću od 10000 u uzorčnome korpusu (za druge korpuse je potrebno povećati vrijednost).
- Prednost rečenicama koje sadrže prvih 10 kolokata ključne riječi s ciljem dobivanja tipičnijih primjera.

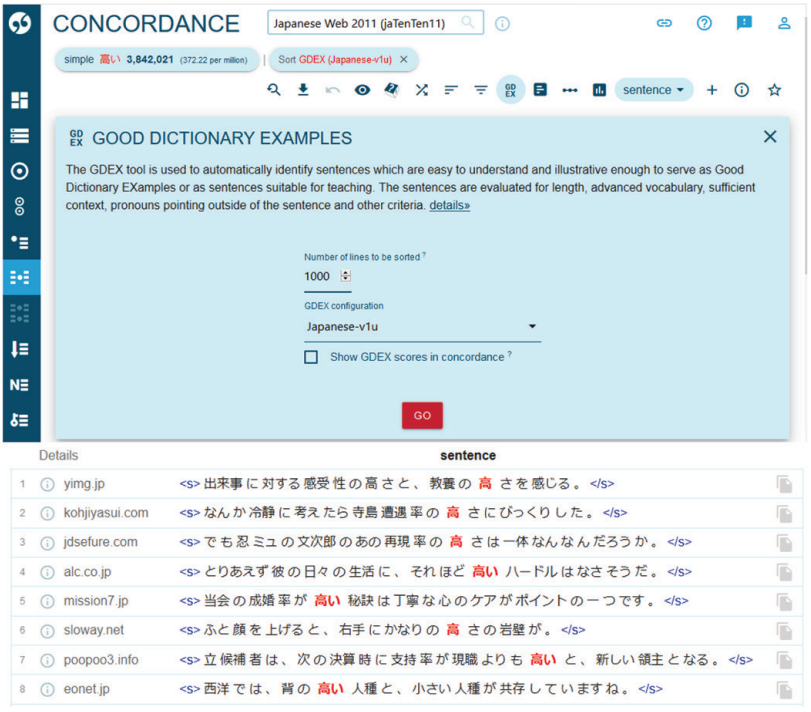
Slika 1 i Slika 2 ilustriraju način korištenja alata GDEX u okviru konkordancera (engl. *Concordance*) u Sketch Engineu. U konkordanceru su prikazani primjeri za lemu *takai* 高い 'visok, skup'. U gornjem desnom uglu konkordancera je ikonica GDEX (Slika 1) koja vodi do prozora gdje se odabire broj primjera[8] iz korpusa koje GDEX analizira za automatsko sortiranje najboljih kandidata primjera, i naziv konfiguracije koja se koristi za sortiranje. S odabirom 1000 primjera iz japanskoga mrežnoga korpusa JaTenTen11 i konfiguracijom *Japanese-v1u* (GDEX for Japanese language) GDEX automatski sortira kandidate primjera prikazanih u donjemu dijelu Slike 2. U rezultatima se vide primjeri za lemu *takai* 高い 'visok, skup' u različitim oblicima i rečeničnim funkcijama – u ulozi pridjeva uz imenice, priložnoj ulozi uz glagole i kao imenica *takasa* 高さ'visina' izvedena od pridjeva dodavanjem sufiksa *-sa*.

Funkciju GDEX moguće je koristiti i u Skicama riječi uz pomoć alata Tickbox lexicography (Kilgarriff i dr. 2010) koji omogućuje odabir kolokacija za koje želimo da GDEX automatski sortira primjere, kao i odabir primjera koje želimo upotrijebiti u rječniku, sačuvati ili direktno umetnuti u program za obradu rječnika. Slika 3 pokazuje skicu riječi za lemu *takai* 'visok, skup' i pored svakoga kolokatora kvadratić koji se može mišem označiti i odabrati za crpljenje primjera. Na Slici 4 vide se rezultati automatski ponuđenih primjera za priložnu upotrebu pridjeva *takai* uz glagole, u kolokaciji 高く売れる *takaku ureru* 'prodavati se skupo'

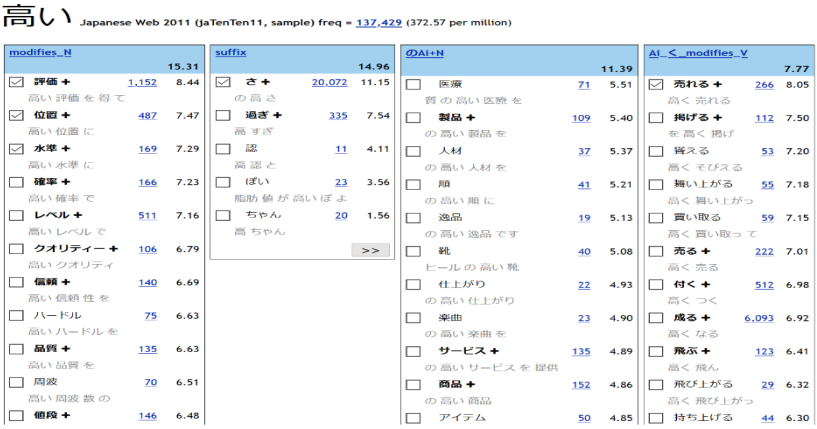
8 Početna vrijednost (engl. *default*) je 300 prvih primjera korpusa.



Slika 1 Funkcija GDEX u konkordanceru (u gornjem desnom uglu)



Slika 2 Funkcija GDEX u konkordanceru: Broj primjera, GDEX konfiguracija i primjeri u obliku rečenica



Slika 3 Funkcija GDEX u Skicama riječi: Tickbox leksikografija

Slika 4 Funkcija GDEX u Skicama riječi: sortirani primjeri
(lijevo početna GDEX konfiguracija; desno GDEX za japanski jezik)

Evaluacija GDEX-a za japanski jezik (GDEX for Japanese language: Japanese-v1u) u usporedbi s početnom konfiguracijom GDEX-a (engl. *Default configuration*) pokazala je da konfiguracija prilagođena japanskome jeziku daje bolje rezultate primjera koji su informativniji, jasniji i bolje oblikovani (Srdanović & Kosem 2016), no bilo bi ih poželjno još dodatno dorađivati prije svega što se tiče rijetkih i težih riječi, dužine rečenica, proširivanja crne liste i sankcioniranja izraza kojima je često potreban širi kontekst za razumijevanje, pojavljivanje brojeva. Da bi se došlo do primjera koji su još prilagođeniji potrebama učenika, napravili smo novu skupinu konfiguracija, tzv. Učenički GDEX za japanski jezik, opisan u sljedećemu poglavlju.

2.2.2. Učenički GDEX za japanski jezik (Learner's GDEX for Japanese language)

Konfiguracije Učeničkoga GDEX-a za japanski jezik u osnovi imaju opći GDEX za japanski jezik, ali se razlikuju po tome što je svaka konfiguracija prilagođena jednoj od razina JLPT-a (Japanska zaklada i Udruženje za međunarodno obrazovanje u Japanu 2004). Glavna razlika je u tome da se sankcioniraju riječi koje su teže od specificirane razine u konfiguraciji i davanju prednosti riječima koje sadrže određeni postotak riječi u ciljanoj razini (Srdanović & Kosem 2016). Za konfiguraciju se rabe pet postojećih razina znanja koje JLPT (Japanese Language Proficiency Test = Test poznavanja japanskoga jezika) definira i specificira u listama riječi po razinama, prema kojima je peta razina najlakša, a prva najteža. To je široko rasprostranjena lista riječi koja se koristi kako za testiranje razine znanja japanskoga jezika, tako i za izradu udžbenika za strance. Kako je u javnoj upotrebi stara lista od četiri razine, a ne pet, koristili smo se izvorom koji je pripremljen na osnovi informacija o novoj i staroj listi i analize frekvencije riječi.[9]

Konačni popis konfiguracija učeničkoga GDEX-a za japanski jezik (Learner's GDEX for Japanese language) je sljedeći:

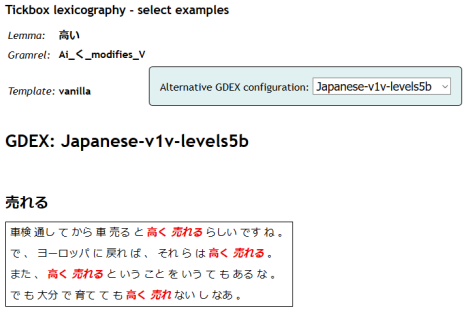
- Japanski-v1v-level1-5 (najteža razina, sankcioniranje riječi izvan JLPT popisa, prednost svim riječima s popisa JLPT razine od 1 do 5)
- Japanski-v1v-level2-5 (sankcioniranje riječi izvan razine 2, 3, 4 i 5, prednost riječima u listi JLPT razine od 2 do 5)
- Japanski-v1v-level3-5 (sankcioniranje riječi izvan razine 3, 4 i 5, prednost riječima u listi JLPT razine od 3 do 5)
- Japanski-v1v-level4-5 (sankcioniranje riječi izvan razine 4 i 5, prednost riječima u listi JLPT razine 4 i 5)
- Japanese-v1v-level5 (najlakša razina, sankcioniranje riječi izvan razine 5, prednost riječima u listi JLPT razine 5), koji je detaljnije bio testiran i dorađivan te je zadnja inačica: *Japanese-v1v-level5b*.

Konfiguracija upotrebljava listu lema da bi obuhvatila varijacije u japanskome pismu i promjene pridjeva i glagola, pri čemu je lista ujedno donekle prilagođena tako da odgovara segmentaciji i načinu označivanja korpusa. Slika 5 pokazuje rezultate učeničkoga GDEX-a za japanski jezik, zadnje inačice konfiguracije *Japanese-v1v-level5b* s rezultatima sortiranih primjera za kolokaciju 高く売れる *takaku ureru* 'prodavati se skupo'.

Vrednovanje učeničkoga GDEX-a za japanski jezik (Srdanović & Kosem 2016) pokazalo je da su rečenice dobivene ovim konfiguracijama manje teške od rečenica dobivenih općom konfiguracijom GDEX-a za japanski

9 Vidi: http://www.tanos.co.uk/jlpt/ i www.jlptstudy.net/ (posjećeno: 11. 9. 2019.)

jezik te time primjerenije za učenike. Ujedno je potvrđeno da primjeri postaju teži s konfiguracijama namijenjenih višim razinama, odnosno lakši za niže razine znanja. Ukazano je i na mogućnosti unapređenja alata s prilagodbom podataka o kanjijima i gramatičkim konstrukcijama podijeljenim prema razinama, kao i s dodatnim podešavanjem parametara i nadopunjavanjem konfiguracije novim spoznajama o pogodnosti primjera i tehničkim napretkom (Srdanović & Kosem 2016; Kosem i dr. 2019).



Slika 5 Učenički GDEX za japanski: rezultati automatski dobivenih primjera za kolokaciju 高く売れる takaku ureru 'prodavati se skupo'

3. Korištenje Učeničkoga GDEX-a za japanski jezik i analiza primjerenosti primjera

Ovo poglavlje nakon kratkoga opisa o potrebama za resursima japanskoga jezika na izvornome jeziku učenika i provedenoj edukaciji studenata o uporabi korpusa, GDEX-a i drugih alata, opisuje rezultate vrednovanja primjerenosti primjera dobivenih korištenjem Učeničkoga GDEX-a za japanski jezik od strane učenika jezika, te uz dodatnu analizu vrednovanja obrazlaže prihvatljivost angažiranja studenata japanskoga jezika u procesu odabira primjera za dvojezični japansko-hrvatski rječnik.

3.1. Potreba za resursima na izvornome jeziku: pilot-projekt dvojezičnoga rječnika

S razvojem studijskoga programa japanologije na Odsjeku za azijske studije Filozofskoga fakulteta Sveučilišta Jurja Dobrile u Puli i rastućim interesom studenata diljem regije i šire za studiranjem japanskoga jezika i kulture, javlja se i sve veća potreba za resursima na materinskome jeziku, a između ostaloga i za dvojezičnim japansko-hrvatskim i hrvatsko-japanskim rječnikom. Dok postoji priličan broj izvora na englesko-japanskom, japansko-engleskom, vrlo je ograničen broj resursa u paru s hrvatskim ili nekim drugim srodnim slavenskim jezikom, npr. srpskim ili slovenskim jezikom. Među relevantnim leksikografskim resursima ističe se *Japansko-hrvatski, hrvatsko-japanski rječnik* (Yamasaki-Vukelić 2006) koji je nastao na osnovi japansko-srpskoga

i srpsko-japanskogaa izdanja te postoji u tiskanome obliku s nešto više od 12000 natuknica jednostavnije strukture. Zatim, JaSlo, koji je japansko-slovenski mrežni rječnik za slovenske učenike japanskoga jezika s oko 10000 natuknica, detaljnom mikrostrukturom i poveznicama na podatke paralelnoga korpusa i podatke iz mrežnoga korpusa (Erjavec i dr. 2006; Hmeljak Sangawa & Erjavec 2010). Osim toga postoje hrvatski i slovenski prijevodi s manjim brojem riječi u okviru višejezičnoga mrežnoga alata za pomoć u čitanju i analizi japanskih tekstova, tzv. Reading Tutor.[10] Manji japansko-hrvatski popis riječi postoji i u prvome japanskom udžbeniku na hrvatskome jeziku *Ippo ippo* (Marković i dr. 2018).

Kako su studentima japanskoga jezika postojeći resursi na materinskome jeziku često nedostatni i nedovoljno detaljni, pomažu se pretraživanjem nepoznatih riječi, njihove upotrebe i prijevoda pomoću internetskih pretraživača i mobilnih aplikacija, uz pomoć mrežnoga prevoditelja i drugih resursa, uglavnom preko engleskoga jezika. Ovi im resursi pomažu do određene mjere, ali još uvijek nisu dostatni da odgovore na njihove potrebe za informacijama na materinskome jeziku.

S obzirom na jasnu potrebu za japansko-hrvatskim rječnikom koji bi sadržao detaljnije informacije, poput učestalih obrazaca, primjera upotrebe riječi, jezične razine i sl., ulažu se početni napori u razvoju resursa te se je započelo nekoliko pilot-projekata dvojezičnih rječnika uključujući i japansko-hrvatski opći mrežni rječnik i rječnik specijaliziran za područje turizma, koji ciljaju na studente ne samo kao korisnike već i kao sukreatore rječnika (Srdanović 2018). U okviru ovoga projekta od velikoga potencijalnoga značaja je i korištenje alata GDEX.

3.2. Upoznavanje studenata s leksikografskim radom i tehnologijama

Studenti treće godine programa japanskoga jezika i kulture mogli su tijekom akademskih godina 2017./2018. i 2018./2019. sudjelovati na kolegiju Uvod u leksikologiju i leksikografiju, koji je bio organiziran 90 minuta tjedno 15 puta u semestru i nosio je 4 ECTS-a.[11] Cilj je kolegija definirati i objasniti osnovna načela opće i japanske leksikologije i leksikografije te primijeniti znanje u kreiranju i analizi rječnika koristeći se novim tehnologijama.

Nakon osnovnoga upoznavanja s korpusnom leksikografijom i različitim korpusima japanskoga jezika, studentima se demonstrira način i mogućnosti korištenja alata Sketch Engine te ostali alati za pretraživanje

10 Mrežni rječnik Reading Tutor Web Dictionary postoji za različite jezike uključujući i mini hrvatski rječnik dostupan unutar alata za analizu teksta i pomoć u čitanju: http://chuta.jp/ (posjećeno 8. 10. 2019.).

11 Od akademske 2020./2021. godine, kolegij će se izvodi na prvoj godini diplomskoga studija.

korpusa. Studenti dobivaju niz zadataka s ciljem da se upoznaju s korpusima i alatima, poput tehnologije za gradnju korpusa, skicama riječi (word sketches), usporedbom skica riječi, tezaurusom, GDEX-om i ostalim funkcionalnostima u okviru Sketch Enginea. Studenti su potaknuti da se njima koriste za razne aktivnosti i upućeni su u samostalno obavljanje zadataka. Jedan od zadataka bio je i unošenje novih natuknica u japansko-hrvatski rječnik te skupno definiranje pravila i načina zapisivanja natuknica u rječniku. Studentima je demonstrirana upotreba alata Lexonomy za potrebe uređivanja rječnika, koji se je zatim koristio za izgradnju unutarnje strukture rječnika i njezina prikaza s ciljem timskoga i dosljednoga korištenja i uređivanja rječnika (Srdanović 2018).

3.3. Učeničko vrednovanje primjera dobivenih Učeničkim GDEX-om

Nakon nekoliko tjedana edukacije o leksikografiji, korpusnoj leksikografiji, jezičnim tehnologijama za japanski jezik, studenti japanskoga jezika na srednjoj jezičnoj razini dobili su detaljne upute i zadatak vrednovati primjere dobivene uz pomoć Učeničke GDEX funkcije u Sketch Engineu. Za te su se potrebe koristili zadnjom inačicom konfiguracije *Japanese-v1v-level5b*, s obzirom na to da je namjenjena učenicima jezika, da je u prethodnome istraživanju (Srdanović & Kosem 2016) dala najbolje rezultate u odabiru prikladnih primjera te da su prosječni rezultati težine primjera pokazivali srednju jezičnu razinu.

Zadatak se je sastojao od sljedećih koraka.[12]

Studenti su najprije odabrali po svojoj želji tri različite riječi iz liste novih riječi udžbenika *Tobira* (Oka i dr. 2009), kojim se koriste u učenju japanskoga jezika na srednjoj razini (kraj druge i treća godina studija). Napomenuto im je da se po mogućnosti koriste različitim vrstama riječi, npr. jednu imenicu, jedan glagol i jedan pridjev.

Zatim su potražili svaku od tih riječi koristeći funkciju Word Sketch na platformi Sketch Engine u japanskome uzorčnom mrežnom korpusu JaTenTen11 (301 miliona pojavnica) s označenom naprednom opcijom *Use tickbox lexicography* (opcija koja omogućava izravni odabir primjera automatski sortiranih alatom GDEX) .

U Word sketch rezultatima tražene riječi odabrali su neku od kolokacija po želji, pri čemu im je sugerirano da se koriste čestim odnosno značajnijim kolokacijama te za njih izlistaju primjere.

U prozoru s primjerima odabrali su alternativnu GDEX konfiguraciju

12 Upute uručene studentima su bile znatno detaljnije i postupnije, ali su ovdje zbog svrhe i dužine rada sažete u nekoliko koraka.

Japanese-v1v-levels5b te odabrali za svaku kolokaciju po četiri primjera koja su kopirali u upitnik i za svaki od primjera ispunili odgovore (jednomu primjeru posvećena je jedna stranica u upitniku).

Osnovna pitanja u upitniku ugrubo pokrivaju sljedeća četiri područja: a) Primjer je dobar za uključivanje u japansko-hrvatski dvojezični rječnik (Ako biste prilagodili primjer (skratili, promijenili), napišite kako bi taj primjer tada bio i što odnosno zašto biste promijenili), b) Primjer je jednostavan i razumljiv, c) Primjer ukazuje na tipičnu uporabu (kolokacije, gram. i sl.), d) Druga opažanja (Ako je teška gramatika ili kanji, ima vama teških ili nejasnih riječi, navedite ih).

3.4. Rezultati vrednovanja

Analizirano je ukupno 60 dobivenih primjera 60 kolokacija od 15 različitih riječi koje je pet studenata[13] samo biralo i vrednovalo te su rezultati prikazani u Slici 6. Rezultati pokazuju da se 64 % primjera ocjenjuje kao prikladno za uvrštavanje u rječnik, 28 % kao potencijalno prikladno, uglavnom zbog teže razumljivoga primjera (teže riječi ili gramatike), osobnih imena u primjerima ili nejasnosti primjera bez šireg konteksta, te ih je u nekim slučajevima moguće prilagoditi upotrebi za rječnik, dok je samo 5 % vrednovano kao neprimjereno za uključivanje u japansko-hrvatski dvojezični učenički rječnik. Gotovo polovica primjera (48 %) procjenjuje se jednostavnim i razumljivim, dok je trećina (35 %) učenicima razumljiva uz upotrebu rječnika. Za većinu primjera procijenjeno je da ukazuju na tipičnu uporabu riječi (90 %).

U komentarima studenti navode većinom probleme u vezi s težim riječima (11 puta), uljudnim frazama (1), imenima (1), sa strukturom i gramatičkim konstrukcijama (5 puta), dužinom rečenica (2) te o tome da nisu sigurni kako nešto prevesti ili jesu li ispravno preveli rečenice (6 puta).

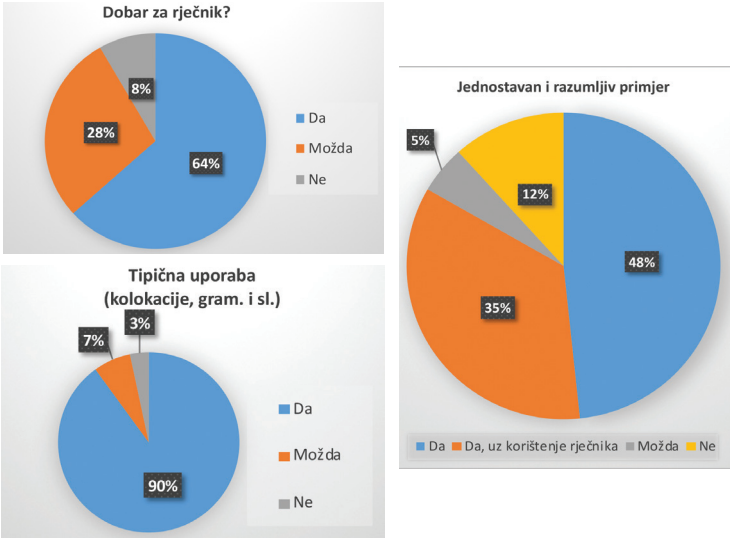
U vezi s gramatikom spominje se za različite primjere: a) nesigurnost oko određivanja subjekta rečenice (1 puta), b) dvostruka negacija koja može biti zbunjujuća učenicima (1 puta), c) kompliciranа, naprednija ili zbunjujuća gramatička struktura (4 puta), d) raspored surečenica (1 puta).

Nesigurnost u prijevodu se vidi u nekoliko tipova komentara: a) nesigurnost kako (točno) prevesti ili kako je prevedeno (3 puta), b) navođenje da je potreban bolji prijevod (1 puta), c) teško je prevesti zbog nedostatka konteksta (učenik je odlično detektirao problem nedostatka konteksta) (1 puta), d) prijevod nema smisla (učenik je odlično detektirao problem da primjer nije prikladan) (1 puta).

Među riječima koje su navedene u komentarima kao teške, spominju se npr.

13 Za analizu je uvršteno vrednovanje pet studenata metodom slučajnoga odabira.

激しい *hageshii* 'intenzivno', 葬地 *souchi* 'groblje', 旦那 *dan'na* 'suprug', 当初 *tousho* 'na početku' i dr., riječi koje su izvan liste JLPT-a ili na težim razinama JLPT1 ili 2.



Slika 6 Rezultati vrednovanja primjera dobivenih Učeničkim GDEX-om za japanski jezik

3.5. Analiza vrednovanja i primjerenosti automatski dobivenih primjera

U ovome poglavlju analiziramo automatski dobivene primjere i rezultate učeničkoga vrednovanja primjerenosti primjera s ciljem rasvjetljavanja korisnosti uporabe alata GDEX od strane učenika japanskoga jezika na srednjoj jezičnoj razini u procesu odabira dobrih primjera za učenički rječnik japanskoga jezika. Kao kriteriji koriste se u uvodnome dijelu opisane četiri osnovne kategorije: prirodnost, tipičnost, informativnost i razumljivost (Atkins & Rundell 2008; Kosem 2019) za odabir kvalitetnih i prikladnih primjera te se s obzirom na te kategorije vrednuju primjeri i procjene studenata navedene u njihovome vrednovanju.

3.5.1. Prirodnost

Kako prirodnost podrazumijeva da se primjer može očekivati u stvarnoj jezičnoj uporabi (Atkins & Rundell 2008), za većinu odabranih primjera to možemo reći s obzirom na to da se radi o autentičnim primjerima iz jezičnoga korpusa. Iz toga možemo izdvojiti nekoliko primjera koji sadrže u sebi neprirodnu i nestandardnu uporabu uslijed nedovoljno pročišćenoga mrežnoga teksta ili segmentacije rečenica, što su studenti prilikom vrednovanja primjera uspjeli dobro zamijetiti i takve primjere vrednovati kao neprikladne ili potencijalno prikladne za uključivanje u rječnik.

3.5.2. Tipičnost

Tipičnost primjera odnosi se na to da se u primjeru vidi tipična upotreba natuknice, npr. u tipičnome kontekstu, strukturi, s tipičnim kolokacijama i slično (Atkins & Rundell 2008). Iz Slike 6 možemo vidjeti da su studenti vrednovali da 90 % primjera predstavlja tipičnu uporabu. Dobivanju tipičnih primjera korištenjem GDEX-a svakako pomaže funkcionalnost Skice riječi koja bira najfrekventnije i najznačajnije kolokacije, kao i nekoliko parametra u konfiguraciji GDEX-a koji daju prednost frekventnijim kolokacijama i penaliziraju rijetke riječi. Pogled u frekventnost i značaj kolokacija odabranih za primjere pokazuje da su, osim u nekoliko izuzetaka, birane najfrekventnije i najznačajnije kolokacije, te da je njihova prosječna frekvencija pojavljivanja u korpusu 580, dok je prosječna statistička značajnost kolokacije 6,16 što je razmjerno visoko. Međutim, detaljnija analiza samih primjera otkriva da se povremeno u primjerima javljaju i neke rjeđe riječi te bi bilo potrebno još podešavati konfiguraciju za postizanje što veće tipičnosti u primjerima.

3.5.3. Informativnost

Primjer je informativan ako u sebi nosi potrebne i dodatne informacije o značenju i uporabi riječi, njezinim kolokacijskim tendencijama u povezivanju s drugim riječima i pomaže razumijevanju definicije i uporabe (Atkins & Rundell 2008). Svakako prvi korak u informativnosti primjera je da imaju oblik pune rečenice normalne informativne strukture, što je u većini primjera dobivenim konfiguracijom GDEX-a za japanski jezik ostvareno. Konfiguracija je uspjela znatno umanjiti broj neinformativnih primjera u obliku fragmenata ili dvije ili više rečenica, tako da je u tome pogledu ispravne rečenične strukture dostignuto 97 % informativnosti. Nadalje, svaki ispravno formirani primjer na neki je način i informativan, dok jednim primjerom svakako ne možemo pokriti sve potrebne i dodatne informacije o značenju i uporabi jedne riječi. U tome smislu, informativnost jednoga primjera teško je vrednovati, a ujedno je i teško doći do „idealno informativnih primjera". Za veću informativnost svakako je potrebno ponuditi više primjera za istu riječ odnosno kolokaciju, što GDEX trenutno i omogućava (četiri primjera je početna vrijednost). Ono na čemu je poželjno dodatno poraditi je analizirati raznolikost primjera s obzirom na postojeću GDEX konfiguraciju i unaprijediti raznolikost primjera u budućoj nadgradnji s uvrštavanjem novih parametara i podešavanjem postojećih. U studentskome vrednovanju primjera informativnost kao kategorija djelomično je pokrivena procjenom o prikladnosti primjera i dodatnim komentarima. S obzirom na to da je većina primjera označena kao prikladna ili potencijalno prikladna za rječnik (64 %+28 %), može se reći da su studenti zadovoljni s informativnošću vrednovanih primjera, a iz analize komentara i procjene studenata, sljedeće točke se

mogu ispostaviti u svezi s informativnošću primjera: a) bilo bi bolje da je primjer napisan u obliku razgovora (1 put), b) zbunjujuće je pojavljivanje imena u primjeru (3) i nedostatak objašnjenja imena kulturološki relevantne osobe (1), c) neprimjerena tema za rječnik (prostitucija), a nisu ni toliko česte kolokacije (1), d) primjer je specifičan za određenu grupu ljudi (1).

3.5.4. Razumljivost

Primjer je razumljiv kada je jasan i izvan šireg konteksta, kada nema previše rijetke i teške riječi u sebi, i kada nije teško razumljiv zbog svoje dužine (Atkins & Rundell 2008). Iz rezultata vrednovanja u prethodnome se poglavlju vidi da je studentima razumljivo oko polovice dobivenih primjera, odnosno ukupno 83 % primjera uz mogućnost korištenja rječnika. Ispostavljeno je s čime su i koliko studenti imali probleme u razumijevanju primjera (npr. nepoznate/teške riječi, uljudne fraze, nerazumijevanje gramatike, imena, dužine rečenica). Što se tiče nerazumljivosti zbog nedostatka konteksta, četiri primjera je imalo ovaj nedostatak, od čega su studenti uspješno uočili kod dva primjera.

Kako bi utvrdili razinu težine primjera koje su studenti dobili iz korpusa, sve smo primjere analizirali uz pomoć slobodno dostupnoga alata Reading Tutor.[14] Alat vrednuje težinu teksta s obzirom na staru listu JLPT koja sadrži četiri različite jezične razine, počevši od JLPT4 kao najlakše do JLPT1 kao najteže. U Tablici 1 se može vidjeti da je najviše, oko 61 % pojavnica, najlakše JLPT4 razine, dok se pojavljuje oko 11 % pojavnica (odnosno 22 % različnica) koje ne postoje u listi JLPT, odnosno većinom se radi o težim riječima izvan liste. Znatan broj je i riječi na razini JLPT2 (13 % različnica) i JLPT3 (11,5 % različnica). Sveukupno analiza procjenjuje da su primjeri „少し難しい“ (pomalo teški) što se može smatrati prikladnim za pobliže objašnjavanje riječi koje pripadaju srednjoj razini (iz udžbenika srednje razine), posebice ako su uz primjer u rječniku dostupni korisnicima i prijevodi primjera.

Tražene riječi su analizirane u Reading Tutoru kao „Teške“ s više od polovice riječi na razini JLPT2 (srednje teškoj razini), čemu slijede teške riječi izvan JLPT liste i riječi na razini JLPT3, što ukazuje na to da su dobiveni primjeri lakše razine od traženih riječi i sugerira njihovu jednostavnost i razumljivost. Međutim, kako se u ovoj analizi upotrebljava konfiguracija za najnižu jezičnu razinu i daje rezultate srednje razine, svakako bi bilo poželjno da se u budućim konfiguracijama pokuša još sniziti razina težine primjera. Na to upućuju i neki od komentara studenata i procjena tijekom vrednovanja.

14 Alat je nastao u sklopu Chuta rječničkoga projekta i slobodno je dostupan na: http://chuta.jp/. Omogućava pomoć analizom teksta i prijevodom riječi prilikom čitanja japanskoga jezika, kao i vrednovanje težine teksta.

Zanimljivo je da su studenti za tražene riječi iz udžbenika u korpusu najviše birali kolokate koje su im više poznati, te je težina kolokata procijenjena na razini „少し難しい“ (pomalo teški) s najviše kolokata na najlakšoj razini (JLPT4 33 % i JLPT3 21 %), no studenti su povremeno birali i riječi teže razine (21 % riječi izvan JLPT liste, 18 % JLPT2, 6 % JLPT1).

Tablica 1. Jezična razina primjera dobivenih GDEX-om prema analizi Reading Tutora

Jezična razina: ★★★★ Pomalo teško.							
	Broj riječi	Izvan JLPT	JLPT1	JLPT2	JLPT3	JLPT4	Drugo
Pojavnice	790	89	20	105	91	485	79
	100%	11,30%	2,50%	13,30%	11,50%	61,40%	10%
Različnice[15]	340	75	16	69	52	128	7
	100%	22,10%	4,70%	20,30%	15,30%	37,60%	2,10%

Analizom rezultata vrednovanja ustanovljeno je da je 75 % primjera odlično procijenjeno od strane studenata pri čemu su studenti uspijevali uspješno detektirati nekoliko problema u vezi s težinom primjera, nejasnošću i sl. U više su navrata dali i pojednostavljeni primjer za rječnik u zamjenu za teži/složeniji. Studenti su 23 % odabranih kolokacija i vrednovanih primjera procijenili vrlo dobro, pri čemu je u većini slučajeva nedostatak bio u preciznosti prijevoda. Samo dvije procjene bile su slabije kvalitete zbog odabira rijetke kolokacije i zbog lošega prijevoda.

Detaljnijom analizom propusta u prijevodu, zaključili smo da je potrebno znatno popraviti 10 % primjera, te nadopuniti odnosno malo doraditi prijevod za 28 % primjera. U potpunosti je točan prijevod kod 62 % primjera, međutim iz analize rezultata možemo zaključiti da je za upotrebu podataka prikupljenih uz pomoć učenika svakako potrebno proći kroz provjeru kvalitete prijevoda za potrebe budućega rječnika, što je i jedan od planiranih koraka u procesu izgradnje dvojezičnoga japansko-hrvatskog rječnika.

Analiza rezultata i primjera rezultirala je i s nekoliko prijedloga za daljnje podizanje kvalitete konfiguracija i posljedično bolje rezultate primjera.

15 Različnica (engl. *type*) je pojedinačna riječ koja se razlikuje od druge riječi (npr. u korpusu riječ koja se bilježi samo pri prvome pojavljivanju jer se sa svakim sljedećim pojavljivanjem smatra pojavnicom), jedinstveni oblik pojavnice iz korpusa. (objašnjenje termina preuzeto sa stranice Pojmovnik: http://ihjj.hr/mreznik/page/pojmovnik/6/, posjećeno: 12. 11. 2019.)

4. ZAKLJUČAK

Ovaj rad predstavlja osnovna svojstva i mogućnosti uporabe alata za poluautomatski odabir prikladnih primjera za rječnik japanskoga jezika: GDEX za japanski jezik (GDEX for Japanese) i Učenički GDEX za japanski (Learner's GDEX for Japanese) dizajniranih kako bi se olakšala identifikacija i ekstrakcija kandidata za primjere japanskoga jezika u leksikografske svrhe i potrebe učenja jezika. Dok je GDEX za japanski jezik temeljen na najnovijoj metodologiji GDEX konfiguracije za druge jezike i prilagođen specifičnostima japanskoga jezika, Učenički GDEX japanskoga jezika primjenjuje pristup orijentiran na učenike stranog jezika, uzimajući u obzir različite razine leksema temeljene na JLPT popisu vokabulara japanskoga jezika tako da sadrži konfiguracije za svaku od pet različitih JLPT razina.

Nakon jednosemestralne edukacije o temeljima leksikografije i korpusne leksikografije, upoznavanja korpusa i alata za pretragu korpusa, studenti japanologije dobili su zadatak koristiti se korpusom japanskoga jezika JaTenTen11 (301 milijun pojavnica) unutar alata Sketch Engine i procijeniti mogućnost uključivanja prikupljenih primjera u rječnik uz korištenje GDEX funkcionalnosti. Analizirano je ukupno 60 dobivenih primjera 60 kolokacija za 15 različitih riječi koje su studenti birali iz vokabulara udžbenika *Tobira*, udžbenika japanskoga jezika na srednjoj razini. Rezultati pokazuju da se 64 % primjera ocjenjuje kao prikladno za uvrštavanje u rječnik, 28 % kao potencijalno prikladno, uglavnom zbog teško razumljivoga primjera, osobnih imena u primjerima ili nejasnoće primjera bez šireg konteksta, dok je samo 5 % vrednovano kao neprimjereno za uključivanje u japansko-hrvatski dvojezični učenički rječnik. Skoro polovica primjera procjenjuje se jednostavnim i razumljivim, dok je uz mogućnost korištenja rječnika ukupno 83 % primjera razumljivo studentima. Za većinu primjera (90 %) procijenjeno je da ukazuju na tipičnu uporabu riječi.

Vrednovanje studenata je zatim analizirano s ciljem rasvjetljavanja korisnosti uporabe alata od strane učenika japanskoga jezika na srednjoj jezičnoj razini u procesu odabira dobrih primjera za učenički rječnik japanskoga jezika. U obzir su uzete četiri osnovne kategorije za procjenu dobrih primjera za rječnik: prirodnost, tipičnost, informativnost i razumljivost. Ustanovljeno je da je 75 % primjera odlično procijenjeno od strane studenata, pri čemu su studenti uspijevali uspješno detektirati većinu problema u vezi s težinom primjera, nejasnoćama i sl. Rezultati analize jezične razine primjera pokazali su da su primjeri „pomalo teški" s najvećim brojem pojavnica na najlakšoj razini, ali i s određenim brojem težih pojavnica te je još potrebno poraditi na smanjivanju težine primjera dobivenih GDEX-om. Osim toga, studenti su povremeno imali problema s preciznošću prijevoda (znatno je potrebno popraviti 10 % primjera te

nadopuniti odnosno malo doraditi 28 % primjera, točno je prevedeno 62 % primjera), zbog čega je potrebno proći kroz provjeru kvalitete prijevoda za potrebe budućega rječnika, što je i jedan od planiranih koraka. Istraživanje je pokazalo da je funkcionalnost GDEX-a korisna za odabir primjera za dvojezični japansko-hrvatski rječnik za učenike jezika u čijoj izradi sudjeluju i studenti i profesori japanskoga jezika. Analiza rezultata i primjera rezultirala je i s nekoliko prijedloga za daljnje podizanje kvalitete konfiguracija i posljedično bolje rezultate primjera.

Osim toga, slična metodologija u izradi GDEX-a za učenike može biti primijenjena i na mnoge druge jezike te se njezini rezultati mogu koristiti ili za rad s učenicima na izradi rječnika u skladu s korisnički usmjerenim stavom u suvremenoj leksikografskoj teoriji i praksi (Gouws i dr. 2013; Atkins & Rundell 2008; Prinsloo & Gouws 2000) ili za rad s učenicima jezika u okviru pristupa koji zagovara učenje zasnovano na podacima (engl. *Data-driven learning*) pristupa (Johns 1991).

LITERATURA

ATKINS – RUNDELL 2008

B.T.S Atkins – Michael Rundell, „*The Oxford Guide to Practical Lexicography*“, Oxford University Press, Oxford 2008.

ERJAVEC i dr. 2006

Tomaž Erjavec – Kristina Hmeljak Sangawa – Irena Srdanović, „jaSlo, A Japanese-Slovene Learners' Dictionary: Methods for Dictionary Enhancement“, *Proceedings of the 12th EURALEX International Congress,* Turin 2006.

FRANKENBERG-GARCIA 2012

Ana Frankenberg-Garcia, „Learners' Use of Corpus Examples*“, International Journal of Lexicography,* 25, 3, 2012, 273–296.

GOUWS i dr. 2013

Rufus Gouws – Ulrich Heid – Wolfgang Schweickard – Herbert Ernst Wiegand (eds.), *„Dictionaries. An International Encyclopedia of Lexicography“*, Walter de Gruyter 2013.

HMELJAK SANGAWA i dr. 2009

Kristina Hmeljak Sangawa – Tomaž Erjavec – Yoshiko Kawamura, „Automated collection of Japanese word usage examples from a parallel and a monolingual corpus“, *eLexicography in the 21st century: new challenges, new applications: Proceedings of eLex*, Presses Universitaires de Louvain, Louvain 2009, 137-147.

JAPANSKA ZAKLADA 2004
Japanska zaklada i Udruženje za međunarodno obrazovanje u Japanu (Japan Foundation and Association of International Education), „*Nihongo Nouryoku Shiken Shutsudai Kijun (Kaitei-ban) [Japanese Language Proficiency Test Test Content Specifications (drugo izdanje)]* ", Bonjinsha, Tokyo 2004.

JOHNS 1991
Tim Johns, *„Chapter 2: Should you be persuaded: Two examples of data-driven learning"* , *Classroom Concordancing*, Birmingham 1991.

KILGARIFF i dr. 2004
Adam Kilgarriff – Pavel Rychly – Pavel Smrž – David Tugwell, „The Sketch Engine", Proc. Euralex, Lorient 2004, 105-116.

KILGARRIFF i dr. 2008
Adam Kilgarriff – Miloš Husák – Katy McAdam – Michael Rundell – Pavel Rychlý, „GDEX: Automatically finding good dictionary examples in a corpus", *Proceedings of the 13th EURALEX International Congress*, Barcelona 2008, 425–432.

KILGARRIFF i dr. 2010
Adam Kilgarriff – Vojtěch Kovář – Pavel Rychlý, „Tickbox Lexicography", *eLexicography in the 21st century: New challenges, new applications*, Brussels 2010, 411–418.

KOSEM i dr. 2011
Iztok Kosem – Miloš Husak – Diana McCarthy, „GDEX for Slovene", *Proceedings of eLex 2011*, Ljubljana 2011, 151-158.

KOSEM 2015
Iztok Kosem, „Interrogating a Corpus", *The Oxford Handbook of Lexicography,* 76-93.

KOSEM i dr. 2019
Iztok Kosem – Kristina Koppel – Tanara Zingano Kuhn – Jan Michelfeit – Carole Tiberius, „Identification and automatic extraction of good dictionary examples: the case(s) of GDEX", *International Journal of Lexicography, 32, Issue Special Thematic Section: Natural Language Processing and Automatic Knowledge Extraction for Lexicography*, 2019, 119–137.

LJUBEŠIĆ – PERONJA 2015
Nikola Ljubešić – Mario Peronja, „Predicting Corpus Example Quality via Supervised Machine Learning", Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age, Proceedings of the eLex 2015 Conference, Ljubljana/Brighton 2015, 477–485.

MAEKAWA i dr. 2013
Kikuo Maekawa – Makoto Yamazaki – Toshinobu Ogiso – Takehiko Maruyama – Hideki Ogura – Wakako Kashino – Hanae Koiso – Masaya Yamaguchi – Makiro Tanaka – Yasuharu Den, „Balanced corpus of contemporary written Japanese", *Language Resources and Evaluation*, Springer, Netherlands 2013.

MARKOVIĆ i dr. 2018
Ljiljana Marković – Divna Tričković – Irena Srdanović, „Udžbenik japanskoga jezika Ippo Ippo: Glavna knjiga", Sveučilište Jurja Dobrile u Puli, Pula 2018.

NISHINA – YOSHIHASHI 2007
Kikuko Nishina – Kenji Yoshihashi, „Japanese composition support system displaying co-occurrences and example sentences", *Proceedings of the Symposium on large-scale knowledge resources (LKR2007)*, Tokyo Institute of Technology, Tokyo 2007, 119-122.

OKA i dr. 2009
Mayumi Oka – Michio Tsutsui – Junko Kondo – Shoko Emori – Yoshiro Hanai – Satoru Ishikawa, „*Tobira Gateway to Advanced Japanese - Learning Through Content and Multimedia",* Kurosio Publishers, Tokyo 2009.

POMIKÁLEK – SUCHOMEL 2012
Jan Pomikálek – Vid Suchomel, „Efficient web crawling for large text corpora", *Proceedings of the Seventh Web as Corpus Workshop (WAC7)*, Lyon 2012.

SRDANOVIĆ i dr. 2008
Irena Srdanović – Tomaž Erjavec – Adam Kilgarriff, „A web corpus and word sketches for Japanese", Shizen gengo shori (Journal of Natural Language Processing), 15, 2, Tokyo 2008, 137-159. (reprinted in Information and Media Technologies 3, 3, 529–551.)

SRDANOVIĆ i dr. 2013
Irena Srdanović – Vid Suchomel – Toshinobu Ogiso – Adam Kilgarriff, „Japanese Language Lexical and Grammatical Profiling Using the Web Corpus JpTenTen", *Proceeding of the 3rd Japanese corpus linguistics workshop*, NINJAL, Tokyo 2013, 229-238.

SRDANOVIĆ – KOSEM 2016
Irena Srdanović – Iztok Kosem, „GDEX for Japanese: Automatic extraction of good dictionary examples", GLOBALEX 2016 Lexicographic Resources for Human Language Technology, Portorož 2016, 57-64.

SRDANOVIĆ 2018
Irena Srdanović, „Engaging Students in Creating Bilingual Online Learner's Dictionaries: General and Specializing in Tourism", *International Symposium Japanese Language Learning for New Generations: Book of Abstracts*, Pula 2018a.

TOLMACHEV – KUROHASHI 2017
Arseny Tolmachev – Sadao Kurohashi, „Automatic Extraction of High-Quality Example Sentences for Word Learning Using a Determinantal Point Process“, *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications,* Copenhagen 2017, 133–142.

UEYAMA – SRDANOVIĆ
Motoko Ueyama – Irena Srdanović (eds.), „*Digital Resources for Learning Japanese*“, Bognona University Press, Bologna 2018.

YAMASAKI-VUKELIĆ 2006
Hiroshi Yamasaki-Vukelić, „*Japansko-hrvatski, hrvatsko-japanski rječnik*“, Dominović, Zagreb 2006.

ABSTRACT

Choosing suitable examples for the learner's dictionary: GDEX for Japanese and its usage options

Examples make up a significant component of a dictionary and can greatly assist in better understanding the meaning and usage of words. Therefore, an important step in creating a dictionary involves choosing suitable examples, where modern advanced technologies can be useful. The aims of this study are (1) to present the design and characteristics of the GDEX (Good Dictionary EXamples) tool for Japanese, devised to automatically extract good dictionary example candidates and thus assist lexicographers and language learners in choosing good dictionary examples, (2) to assess students' opinions and attitudes towards the results of Japanese examples automatically extracted from corpora using the GDEX tool, and (3) to investigate how useful the tool can be in the process of choosing good dictionary examples for a Japanese learner's dictionary by Japanese language students at an intermediate language level who were trained for lexicographic work using corpora and advanced technology. The GDEX tool for Japanese is designed by using and upgrading methods employed for other languages, which have proven to be useful in lexicographic projects. One step further in the development of the GDEX technology is the Learner's GDEX for Japanese, which applies a language-learner oriented approach by taking into account the different difficulty levels of lexemes based on the Japanese Language Proficiency Test vocabulary list. A questionnaire on the GDEX for Japanese is distributed to intermediate level students at the Juraj Dobrila University of Pula bachelor's program in the Japanese language and culture who were previously trained to use corpora, the corpus query tool Sketch Engine (SkE), and GDEX for Japanese. The results indicate that students most value the dictionary examples, whose inclusion into the learner's dictionary they deem to be useful,

and they manage to notice and mention some difficulties in choosing an appropriate dictionary example. A further assessment of the results reveals that trained students are in most cases able to make a good decision, but in some cases fail to translate examples properly or to mention and/or notice a need to simplify the example by excluding some context related expressions or difficult words. Finally, it can be concluded that the GDEX for Japanese proves to be useful as a tool for the semi-automatic extraction of dictionary examples, and that there is great potential in engaging students while monitoring their work in the ongoing lexicographic project of creating a Japanese-Croatian learner's dictionary.

Keywords: Japanese language learner's dictionary, GDEX (Good Dictionary Examples), Sketch Engine, lexicography, automatic extraction of examples, corpus lexicography, engaging students in lexicographic work