razprave FF

Irena Srdanović

Kolokacije in kolokacije na daljavo v japonskem jeziku: korpusni pristop

Univerza v Ljubljani
FILOZOFSKA FAKULTETA

Kolokacije in kolokacije na daljavo v japonskem jeziku: korpusni pristop

Zbirka: Razprave FF (e-ISSN 2712-3820)

Avtorica: Irena Srdanović
Recenzenta: Andrej Bekeš, Bor Hodošček
Lektor slovenskega dela: Damjan Popič
Lektor japonskega dela: Iztok Ilc
Tehnično urejanje in prelom: Aleš Cimprič
Slika na naslovnici: Stock photo © akiyoko

Založila: Znanstvena založba Filozofske fakultete Univerze v Ljubljani
Izdal: Znanstvenoraziskovalni inštitut Filozofske fakultete
Za založbo: Roman Kuhar, dekan Filozofske fakultete

Ljubljana, 2021
Prva e-izdaja
Publikacija je brezplačna.

Publikacija je v digitalni obliki prosto dostopna na https://e-knjige.ff.uni-lj.si/
DOI: 10.4312/9789610605119

Kataložni zapis o publikaciji (CIP) pripravili v
Narodni in univerzitetni knjižnici v Ljubljani

COBISS.SI-ID 70974723
ISBN 978-961-06-0511-9 (PDF)

Kazalo vsebine

Uvodni del

V pričujoči monografiji je jezik obravnavan kot dinamičen fenomen, z empiričnim pristopom pa so analizirane obsežne podatkovne baze. Tovrstni pristop se v jezikovni teoriji v širšem pogledu sklada s funkcionalizmom, poudarja pa uporabne in operativne vidike jezika. V tem smislu se precej razlikuje od formalizma, ki v osnovi temelji na intuiciji raziskovalca. V pričujočo raziskavo je bila vključena empirična metodologija, in sicer objektivno opazovanje predmeta preučevanja – kolokacij, na podlagi obsežnega gradiva japonskega jezika. Rezultati analize so aplicirani na področje izobraževanja japonščine kot tujega jezika.

Za učence tujega jezika so kolokacije, istočasno pojavljanje dveh ali več besed ali pa besede in določenih stavčnih vzorcev, eden najpomembnejših delov učenja. Kolokacije so v vsakem jeziku edinstvene in pogosto težko predvidljive le na podlagi maternega jezika ali jezika, ki ga že obvladamo, kar predstavlja veliko breme pri učenju. Vendar pa kolokacije do sedaj še niso bile zadostno vključene v učne načrte pri poučevanju japonščine, primanjkuje tudi učnih materialov, kot je npr. slovar kolokacij, zasnovan na uravnoteženem korpusu japonskega jezika, ki bi bili lahko učencem v pomoč. Po drugi strani pa je izdelava obsežnega korpusa japonskega jezika BCCWJ v zadnjem desetletju (2006–2011) omogočila izvajanje podrobnih raziskav o kolokacijah ter korpusno zasnovano pripravo učnih gradiv, slovarjev in drugih pripomočkov, tako da se na tem področju kot tudi pri pomoči učencem japonščine kot tujega jezika pričakuje napredek.

Osnovni namen raziskave je analizirati in ovrednotiti kolokacijske povezave s pomočjo več različnih obsežnih korpusov japonskega jezika in posebej na primeru prislovov in modalnih oblik pokazati pomembnost uporabe korpusa pri izdelavi učnih načrtov in učnega gradiva. Kot je v nadaljevanju natančneje povedano, je cilj raziskave dvojen.

Prvi cilj je razviti in ovrednotiti japonsko različico besednih skic orodja Sketch Engine (SkE), ki omogoča pridobitev kolokacij japonskega jezika, in lahko služi kot osnova za izdelavo kolokacijskega slovarja za učence japonščine. Za to je bilo treba uporabiti obsežni spletni korpus JpWaC in izdelati slovnično datoteko v japonščini, ki zajema različne kolokacijske odnose japonskega jezika s pomočjo korpusne poizvedbene sintakse (ang. *corpus query syntax*), in sicer z uporabo regularnih izrazov (ang. *regular expressions*), angleških prevodov besednih vrst morfološkega analizatorja ChaSen, osmišljenih kolokacijskih vzorcev japonskega jezika in korpusnega orodja SkE. Vrednotenje rezultatov opravimo z metodo primerjanja z obstoječimi slovarski viri ter izvajanjem anket.

Drugi cilj je omogočiti, da se prislovi in modalne oblike na koncu stavka zajamejo kot kolokacije na daljavo ter pridobijo iz korpusov in uporabijo za učenje japonščine kot tujega jezika. Zato je treba v korpusih raziskati naravo kolokacijskih povezav in na podlagi rezultatov kombiniranja izrazov izdelati seznam pogostih modalnih oblik in njihovih variant. Na podlagi ugotovitev želimo podati predlog učnega načrta za besedišče in dodatne podatke za slovar, ki bi vključevali izluščene informacije o prislovih in modalnih oblikah. V tem delu raziskave empirično opazujemo pojav v trinajstih različnih korpusih in podkorpusih japonskega jezika ter primerjamo in kategoriziramo korpuse in kolokacije s statističnimi metodami. Vrednotenje rezultatov opravimo z izvajanjem ankete ter jih primerjamo z obstoječimi učnimi pripomočki in tako potrdimo pomembnost empiričnega pristopa.

V pričujoči monografiji uporabimo trinajst (pod)korpusov, ki so opisani v Tabeli 1. Naslednje podatke korpusa JpWaC uporabimo glede na namen raziskav: a) naključni primeri stavkov, b) vzorčni korpus z 20 milijoni besed, sestavljen naključno iz celotnega korpusa, c) celotni korpus JpWaC (400 milijonov besed).

Tabela 1: *Seznam uporabljenih korpusov.*

Vrsta korpusa		Kratica	Razlaga	Dostopnost
Govorni		NUJCC	Neformalni pogovori (3584 KB)	Univerza v Nagoyi http://tell.cla.purdue.edu/chakoshi/public.html
		Oikawa	Formalni intervjuji (817 KB)	Podiplomska univerza za napredne študije (Sokendai)
Spletni		KokkenOC	Yahoo! Chiebukuro, forum z vprašanji in odgovori (primeri iz korpusa BCCWJ iz leta 2007) (16,3 MB)	Nacionalni inštitut za japonski jezik in jezikoslovje
		JpWaC	Obsežni spletni korpus (7,3 GB)	Japonska različica orodja SkE http://www.sketchengine.com
Pisani	Učbeniki	KokugoK	Učbeniki za japonščino za osnovno šolo (3723 KB)	Tokijski inštitut za tehnologijo
		KokkenK	Učbeniki za srednjo šolo (BCCWJ) (437 MB)	Nacionalni inštitut za japonski jezik in jezikoslovje
		KKK	Učbeniki za japonščino, ki so zajeti v KokkenK (788 KB)	Nacionalni inštitut za japonski jezik in jezikoslovje

Vrsta korpusa		Kratica	Razlaga	Dostopnost
Pisani	Učbeniki s področja naravoslovja	16K	16 učbenikov s področja naravoslovja za študente (2,45 MB)	Tokijski inštitut za tehnologijo
	Članki s področja naravoslovja	NLP	Članki iz japonske revije NLP (719 KB)	Japonsko združenje za NLP (*Natural Language Processing* – obdelava naravnega jezika)
	Bela knjiga	KokkenOW	Vladna bela knjiga (primeri iz korpusa BCCWJ) (16,4 MB)	Nacionalni inštitut za japonski jezik in jezikoslovje
	Uravnoteženi korpus	KokkenBK	Knjige, revije in časopisi iz »publikacij« in »knjižnice« (primeri iz korpusa BCCWJ) (68,6 MB; 140 MB)	Nacionalni inštitut za japonski jezik in jezikoslovje
	Časopis	Mai2002	Časopis Mainichi shimbun, podatki za leto 2002 (100MB)	Časopisna hiša Mainichi shimbun (CD-ROM)
	Drugo	Kudō	Časopisi, moderna literatura	Kudō (2000)

Na koncu uvodnega dela se želim še posebej zahvaliti prof. Kikuko Nishina, častni profesorici na Tokijski tehnološki univerzi, in prof. Andreju Bekešu, rednemu profesorju na Filozofski fakulteti Univerze v Ljubljani, ki sta me med raziskavo skrbno in potrpežljivo usmerjala, mi pomagala z nasveti ter mi tudi sicer stala ob strani. Zahvala gre tudi prof. Chikako Shigemori Bučar, prof. Masanoriju Nakagawi, prof. Shin-ichiju Mayekawi, prof. Hiroyukiju Akami, prof. Masau Muroti za dragocene nasvete. Za podporo se zahvaljujem tudi številnim kolegom: dr. Boru Hodoščku, ki je izdelal program za označevanje modalnih oblik, prof. Asuki Terai s Tokijskega inštituta za tehnologijo za obilico nasvetov o uporabi statističnih metodologij, prof. Terryju Joyceu, mag. Kenjiju Yoshihashiju, prof. Takafumiju Utashiru, prof. Hong-Quan Cao, dr. Naomi Idi, članom raziskovalne skupine orodja SkE, vključno s žal prezgodaj pokojnim prof. Adamom Kilgarriffom, dr. Vojtěchom Kovářom in dr. Vidom Suchomelom. Hvaležna sem tudi skupini za izdelavo korpusov in skupini za didaktiko japonščine kot tujega jezika na Nacionalnem inštitutu za japonski jezik in jezikoslovje, ki so mi vseskozi omogočali uporabo potrebnega gradiva. Zahvaljujem se tudi študentkama Filozofske fakultete Mateji Žabjek in Poloni Vehar za prevajanje raziskovalnega dela, ki je vključen v pričujočo monografijo, ter velikemu številu ljudi, ki so v okviru raziskave sodelovali pri vrednotenju. In nazadnje, *od*

srca se zahvaljujem svojoj porodici, mami Zorici, tati Jovanu, suprugu Željku i ćerkama Alji i Iris za ogromnu podršku i sve dragocene reči pune podsticaja i ljubavi.

Dodatna opomba

Japonske besede so zapisane po sistemu Hepburn, kjer se *ch* približno izgovarja kot slovenski »č«, *j* kot slovenski »dž«, *ts* kot »c«, *sh* kot »š« in *y* kot »j«. Dolgi samoglasniki se zapišejo *aa*, *ii*, *uu*, *ee* ali *ei*, *oo* ali *ou*, razen v primeru že v slovenščini ustaljenih imen, kot je npr. Tokio, in že ustaljenih transkripcij osebnih imen, kot je npr. Kudō.

1 Empirični in aplikativni pristop v analizi kolokacij

V tem poglavju najprej na kratko predstavimo začetke razvoja korpusnega jezikoslovja v svetu in na Japonskem, opredelimo metodološki okvir tega področja ter razložimo vidike uporabe korpusa pri učenju tujega jezika. Zatem povzamemo izsledke dosedanjih raziskav o kolokacijah in kolokacijah na daljavo ter predstavimo raziskovalne cilje, metode in uporabljeno gradivo v analizi.

1.1 Razvoj korpusnega jezikoslovja

Za raziskovanje pojavov v jeziku obstajajo različne metode in ena izmed njih je analiza zbranih jezikovnih podatkov, imenovanih korpusi. Razvoj računalnikov in drugih naprednih tehnologij je omogočil digitalizacijo tovrstnih podatkov in digitalno pregledovanje, tako da se dandanes korpusi gradijo pretežno kot obsežne zbirke načrtno zbranega gradiva v elektronski obliki. Gradnja in uporaba korpusov sta napredovali z razvojem korpusnega jezikoslovja in prinesli velike spremembe na različnih področjih, kot sta na primer jezikovno izobraževanje in leksikografija. Korpusi namesto intuicije in introspekcije posebno poudarjajo jezik v dejanski rabi in tako poskušajo orisati dejansko stanje v jezikovni komunikaciji. Poleg tega so korpusi pripomogli k oživitvi deskriptivnega pristopa k jeziku, saj poudarjajo opis »slovnice, ki je taka, kot je« in se izogibajo tradicionalno preskriptivnemu pristopu, tj. »slovnici, kakršna bi morala biti« (Ishikawa, 2008).

Korpusno jezikoslovje se je začelo razvijati v 50. in 60. letih 20. stoletja. Leta 1964 je bil objavljen Brownov korpus, ki se je proglasil za prvi korpus na svetu. Izdelan je bil na univerzi Brown v ZDA, obsegal pa je milijon besed iz knjig in leposlovja. V istem času je Noam Chomsky izoblikoval teorijo o generativni slovnici, ki je bolj poudarjala zmožnost ustrezne tvorbe jezikovnih oblik (slovnična kompetentnost) kot pa dejansko tvorbo jezikovnih oblik (jezikovna dejavnost). Zaradi močnega vpliva teorije Chomskega raziskave, zasnovane na korpusih, niso bistveno napredovale. V Evropi se v 60. in 70. letih pojavljata korpus Lancaster-Oslo/Bergen (LOB) in korpus London-Lund (The Survey of English Usage), slednji le v obliki zapisov posnetih pogovorov. Močan vpliv sta imela angleški jezikoslovec Halliday (1966, 1978, 1991) in njegova funkcionalistična teorija, zato se je korpusno jezikoslovje začelo razvijati tudi na tem področju. V 90. letih sta bila v Angliji razvita dva obsežna korpusa, in sicer British National Corpus (BNC) in The Bank of English, oba pa sta bila korpusa britanske angleščine. BNC je upošteval načeli reprezentativnosti in uravnoteženosti, obsegal pa je 100 milijonov besed. The Bank of English je bil usmerjen v opazovanje lingvističnih sprememb v realnem času, sčasoma pa se je razširil do obsega 650 milijonov besed, ki so del zdaj že zelo obsežnega korpusa

Collins.[1] Izdelava obeh korpusov je vplivala na založnike slovarjev, ki so ugotovili, da sta gradnja korpusov in leksikografija tesno povezani (Ishikawa, 2008).

Zgodovina sestavljanja slovarjev angleščine je še posebej dolga in pestra, njihova ciljna publika pa so predvsem učenci, ki se učijo tega jezika kot tuji oz. drugi tuji jezik. Prvi slovar na svetu, ki je temeljil na korpusu v elektronski obliki, je bil slovar angleškega jezika Collins COBUILD English Language Dictionary, ki ga je Collins izdal leta 1987 v Angliji. Pozneje so bili izdelani učbeniki in leksikoni, ki temeljijo na korpusih in se jih že več desetletij uporablja za poučevanje angleškega jezika (Lüdeling in Kytö, 2008). Rečemo lahko, da raziskave, ki temeljijo na korpusih angleškega jezika, s številnimi ugotovitvami služijo kot referenca za korpusno jezikoslovje v japonskem in drugih jezikih.

Za japonščino so v 50. in 60. letih z uporabo računalnika opravili prve raziskave besedišča (npr. National Language Research Institute, 1970, 1962–64; Ito, 2003; Minami, 1974). Kljub izredno zgodnji uporabi računalnikov za procesiranje jezika se japonsko korpusno jezikoslovje ni razvijalo tako hitro kot angleško. V 60. letih so bili korpusi sestavljeni iz časopisov ali internih dokumentov inštitucij in javnosti niso bili dostopni. V 80. letih se je začel projekt Japonski elektronski slovar (Japan Electronic Dictionary, EDR), leta 1995 pa je korpus skupaj s slovarjem EDR postal splošno dostopen. Korpus EDR v glavnem vsebuje časopisne članke, približno 200.000 stavkov iz revij v japonskem jeziku in tudi okoli 100.000 stavkov v angleškem jeziku.

Pozneje so bili na Japonskem zgrajeni korpusi časopisnih člankov, npr. Nikkei shimbun in Mainichi shimbun, korpus Univerze v Kjotu, ki je vključeval podatke o besednih vrstah in strukturni odvisnosti,[2] ter prvi govorni korpus japonskega jezika (Corpus of Spontaneous Japanese, CSJ). Čeprav so začeli korpuse pogosto uporabljati kot enotne podatkovne baze za statistično obdelavo jezika (Matsumoto, 2003), je bila uporaba korpusov na področju jezikoslovja in japonskega jezika precej omejena, učni načrti za besedišče, učbeniki in učna gradiva pa so bili zasnovani na podlagi izkušenj in intuicije učiteljev in avtorjev.[3] V tem času se je v Evropi v veliki meri uporabljala metoda gradnje korpusov na podlagi vnaprej določene strukture, ki temelji na ideji Brownovega korpusa in upoštevanja ravnovesja v jeziku, izdelava uravnoteženega korpusa japonščine pa je nekoliko zamujala.

V zadnjem desetletju pa se je položaj za japonski jezik na področju korpusnega jezikoslovja močno spremenil. Za ta preobrat je v veliki meri zaslužen Nacionalni

1 http://www.mycobuild.com/about-collins-corpus.aspx (dostop 19. 9. 2014)

2 V japonščini *kakariuke jouhou*（係り受け情報）.

3 Razlog je lahko v tem, da so imeli lingvisti formalne struje, kot so Chomsky in drugi, ki so v začetnem obdobju kritizirali korpusno jezikoslovje, v Ameriki in na Japonskem močan vpliv.

inštitut za japonski jezik in jezikoslovje, ki se je usmeril v gradnjo obsežnih korpusov japonskega jezika, njihovo vrednotenje in splošno uporabo v jezikoslovnih in drugih raziskavah. Največji korak je bil napravljen s petletnim projektom[4] izdelave obsežnega Uravnoteženega korpusa sodobne pisane japonščine (Balanced Corpus of Contemporary Written Japanese, BCCWJ), od leta 2006 do 2011.

Korpus BCCWJ se uporablja in vrednoti ne samo v okviru jezikoslovja v ožjem pomenu, temveč tudi širše, na primer za jezikovno izobraževanje in poučevanje japonščine, leksikografijo in obdelavo naravnega jezika (Maekawa, 2006a,b; Maekawa et al., 2013). V zadnjem desetletju se tudi pri poučevanju japonščine v veliki meri širi ideja o uporabi korpusa in o potrebi, da se jezikovna gradiva za učence oblikujejo na osnovi objektivnih jezikovnih podatkov. Z uporabo korpusa BCCWJ je bilo opravljenih več raziskav o praktični uporabi korpusov pri poučevanju japonščine (npr. Hashimoto, 2008; Sunakawa, 2008; Nishina, 2008; Hodošček in Nishina, 2012; Srdanović et al. 2009a,b, 2014; Srdanović in Sakoda, 2013).

1.2 Uporaba korpusov pri učenju tujega jezika

Korpusi imajo močan vpliv na razvoj jezikoslovja, uporabljajo se pa tudi pri poučevanju tujega jezika. Tudi na tem področju so v ospredju raziskave, opravljene za angleščino. Od 19. stoletja sta imeli historična lingvistika in fonetika močan vpliv na poučevanje jezikoslovja, lahko pa rečemo, da je korpusno jezikoslovje svoje mesto dobilo šele po letu 1990 (Howatt, 1984; Ishikawa, 2008). Hunston (2002) meni, da je uporaba korpusov povzročila dve veliki spremembi na področju poučevanja jezikov. Prva je ta, da so s tem omogočeni novi načini opisovanja jezika in so izluščene jezikovne informacije bogatejše, npr. frekvenca gesel, kombinacija besed, kombinacija besed in slovničnih vzorcev ter besede, značilne za določeno zvrst. Druga sprememba je možnost priprave in vrednotenja učnega gradiva, ki ni odvisno od subjektivnih presoj in ocen, na primer pri pripravi slovarjev, učbenikov, učnih načrtov in pri neposrednem učenju jezika.

Pri poučevanju jezikov rabo korpusov delimo na dva tipa – direktno in indirektno. Pri direktni rabi poznamo podatkovno usmerjeno učenje (*data-driven learnig* ali DDL). Podatkovno usmerjeno učenje je poučevalna metoda, pri kateri učenci uporabljajo korpus, da odkrijejo rabo jezika (Johns, 1991). Na primer, učenec primerja učbenik slovnice z rezultati korpusne analize ter z vrednotenjem pride do novih znanj. V ta namen se uporabljajo različna orodja kot so sistemi za pomoč pri učenju, konkordance in bolj napredni korpusni iskalni sistemi

4 Projekt MEXT Grant-in-Aid for Scientific Research Priority Area Program: Japanese Corpus, v nadaljevanju projekt Korpus japonščine.

(Smrž, 2004; Smith et al, 2007, 2008). Hirotani (2007) podaja primer raziskave o izdelavi in uporabi konkordanc pri poučevanju japonščine, pri katerem učenci uporabljajo konkordance in s tem sami odkrijejo svoje napake v rabi jezika. Poleg tega je bilo razvitih več sistemov za podporo učenju, ki so namenjeni za učence japonščine. To sta na primer podporna sistema za bralno razumevanje ASUNARO (Nishina in Yoshihashi, 2007) in Reading Tutor (Kawamura, 1993) ter sistem za pomoč pri pisanju Natsume (Nishina, 2008; Hodošček in Nishina, 2012).

Indirektna raba korpusa pomeni razvijanje učnih pripomočkov, ki temeljijo na korpusih. Pri tem raziskovalci in avtorji učnega gradiva uporabijo prednosti korpusa, da dopolnijo slovarje in učbenike. Ta pristop teži k izboljšavi učnega načrta in oblikovanju učnega gradiva, ki je v skladu z dejanskim stanjem v jeziku. Korpus ravno tako zagotovi merila za učno gradivo in omogoča celostno analizo gradiva na mikro- in makroravni.

Pri izdelavi slovarjev in slovničnega gradiva za angleščino so korpusi prinesli velike spremembe (Sinclair, 1987). Korpusi so postali temeljno orodje za leksikografijo, velika založniška podjetja pa spodbujajo izdelavo slovarjev, osnovanih na korpusih (Hunston, 2002). V slednjih so za razliko od tradicionalnih možne dopolnitve z informacijami o frekvenci, s pomeni besed, primeri in opisi registra ter kolokacijami. Uporabniku je na primer mogoče podati objektivne informacije o frekvenci in spremeniti barvo besede glede na frekvenco ali dodati oznako frekvence. Longmanov slovar sodobne angleščine (Longman Dictionary of Contemporary English) ima denimo označene tri stopnje frekvence za govorni in pisni jezik. Angleški slovar za višjo stopnjo Collins Cobuilt Advanced Learners English Dictionary ima besede prikazane glede na vrstni red frekvenc. Z uporabo korpusa za učence lahko poleg tega učenci sami najdejo morebitne napake ter vnašajo in dopisujejo opombe (Gillard in Gadsby, 1998; Ishikawa, 2008).

Korpusi imajo pomembno vlogo tudi pri pripravi učnih načrtov, še posebno pri pripravi besedišča (Sinclair in Renouf, 1988; Willis, 1990). Besedišče je osnova za znanje tujega jezika in je poleg slovarja eden izmed osnovnih učnih pripomočkov. Pri pripravi pride do težav pri izbiri, katere besede so bolj relevantne. S tradicionalnega vidika jih po subjektivni presoji glede na frekvenco in druge faktorje izbere učitelj oz. sestavljalec besedišča (Palmer, 1930; West, 1953; JACET4000, 1993). Obstaja pa tudi stališče, ki daje prednost objektivnim podatkom. Sem spadata npr. seznam, ki temelji na Brownovem korpusu (Kučera in Franci, 1967), in seznam, ki temelji na korpusu BNC (Leech et al., 2001). Omenjena seznama vsebujeta podatke o stopnji frekvence besedne vrste, stopnji frekvence enot v govornem in pisnem jeziku ter posebnosti enot glede na register,

ki jih sicer drugače redko vidimo.[5] Obstajajo tudi primeri seznamov, oblikovanih z uporabo različnih virov. Seznam besedišča JACET8000 (2003), namenjen japonskim učencem angleškega jezika, je denimo sestavljen na podlagi razmerja frekvenc v korpusu BNC in za te namene izdelanega podkorpusa angleških besedil, s katerimi se učenci pogosto srečujejo.

Nekateri raziskovalci izražajo negativno stališče do uporabe korpusov pri poučevanju jezika. Widdowson (2000) kritizira uporabo korpusov, saj ti okrnijo kontekst, zato to ne more biti avtentičen jezik. Cook (1998) meni, da s prevelikim zaupanjem v frekvence tvegamo, da učenje usmerimo preozko in tako zanemarimo besede z nizkimi frekvencami, ki pa imajo lahko specifičen pomen. A kljub kritikam to še ne pomeni, da je obstoječa metoda subjektivne analize in razvoja učnega gradiva zadovoljiva. Leech (1997) priznava, da je načelo »najpogostejše je tudi najpomembnejše« preveč poenostavljena, vendar pa je težko reči, da frekvenčni podatki, pridobljeni iz korpusa, niso pomemben empirični prispevek k učnim gradivom (Ishikawa, 2008).

Uporaba korpusov je pritegnila nekaj pozornosti tudi na področju poučevanja japonščine, v precejšnji meri pa je bila uresničena pri delu na projektu Korpus japonščine (BCCWJ). Leksikografska sekcija, vključena v projekt, je opravila raziskavo v povezavi s pripravo japonskega slovarja, ki bi temeljil na korpusu (Ogino, 2008), sekcija za poučevanje japonščine pa je imela nalogo, da za učence japonščine pripravi raznovrstno učno gradivo (Sunakawa, 2008). Tudi za pripravo besedišča je obsežni korpus BCCWJ postal pomemben jezikovni vir (Yamauchi, 2006, 2008).

1.3 Začetki raziskovanja kolokacij in pomembnost učenja kolokacij za tuje učence

Tradicionalno je bila enobesedna enota ena od glavnih predmetov preučevanja v jezikoslovju, skupaj s slovničnimi pravili o tvorbi pravilnih stavkov. Rezultati, pridobljeni s korpusnojezikoslovnimi metodami, pa kažejo, da beseda ne deluje neodvisno, ampak se so-pojavlja z drugimi besedami in drugimi jezikovnimi elementi, da doseže svoj namen (Ishikawa, 2008; Srdanović 2014). Te ugotovitve so prispevale k vse aktivnejšemu raziskovanju kolokacij oz. kolokacijskih povezav. Začetki tovrstnih raziskav segajo že v predkorpusno obdobje, in sicer na področju poučevanja angleščine na Japonskem (npr. Saitou, 1905; Palmer, 1938; Hornby, 1954). Te raziskave so vplivale na J.R. Firtha (1951) iz londonske šole in so prispevale k razvoju

5 Na primer, pri frekvenci en milijon se beseda *exercise* »vaditi, vadba« kot samostalnik pojavi 79-krat, kot glagol pa 49-krat. Beseda *what* »kaj« se v govornem jeziku pojavi 7313-krat, v pisnem jeziku pa 1936-krat. Tovrstni podatki postanejo razvidni iz korpusne analize.

Hallidayeve funkcionalistične jezikovne teorije in nekoliko pozneje še Sinclairjeve korpusne raziskave o kolokacijah. Sinclair (1991) je izrazil pomembnost kolokacij in idiomov v obliki tako imenovanega načela idiomov (ang. *Principle of Idiom*).[6] Gre za to, da govorec v naravnem jeziku v veliki meri uporablja fraze in skladnjo, ki je že vnaprej pripravljena v vzorcih. Erman in Warren (2000) menita, da se približno polovica besedila naravnih govorcev ujema z načelom idiomov (Fukada, 2008). Ellis (2001) poudarja pomembnost kolokacij in trdi, da je poznavanje kolokacijah osnova za poznavanje jezika. Znanje jezika je odvisno od jezikovnih koščkov (angl. *language chunks*), ki so pomensko zaključeni in shranjeni v dolgoročnem spominu, ter od izkušenj o kolokacijskih povezavah teh posameznih jezikovnih koščkov.

Za to, da posamezne besedne fraze postajajo zaključene enote, obstajajo trditve na podlagi človeške intuicije kot tudi dokazi, ki jih ponujajo korpusi. Glede na raziskave o napačni uporabi kolokacij učencev tujega jezika so tovrstne enote včasih slovnično in besedno nepredvidljive, zato se učenci hitreje zmotijo (Nation, 2001). Na primer, kolokacija *faasuto fuudo* »hitra prehrana« (ファーストフード, angl. *fast food*) v japonščini ne obstaja kot **hayai tabemono* »dobesedno: hiter + hrana« (早い食べ物) ali **kyuukou tabemono* (急行食べ物), kot bi nekdo, ki pravilne kolokacije ne pozna, lahko sklepal na podlagi poznavanja besed za »hitro« in »hrano«. V japonščini se za »nehalo je deževati« ne uporablja **ame ga tomaru* »dobesedno: dež se ustavi« (*雨が止まる), ampak *ame ga yamu* »dobesedno: dež preneha« (雨がやむ) in *ame ga agaru* »dobesedno: dvigne se« (雨があがる). Tudi »pripraviti čaj« se ne kombinira kot *ocha* + *wo* + *tsukuru* »čaj + pripraviti, narediti« (お茶 +を +作る), ampak kot *ocha* +*wo* + *ireru* »čaj + vstaviti« （お茶 + を + いれる) (Fukada, 2008).

Številne raziskave so opozorile na pomembnost preučevanja kolokacij. Kjellmer (1991) tako denimo meni, da so kolokacije pomembna študijska snov, če želimo govoriti kakor naravni govorci. Nujno je, da se ne učimo le posameznih besed, ampak da se osredotočimo na sopojavljanje teh besed. Tudi James (1998) na podlagi rezultatov raziskave poudarja, da je učenje kolokacij pomembno za »naravnost jezika«. V njegovi raziskavi so bili primerjani korpusi naravnih govorcev z učenci, in pri kolokacijah, ki jih uporabljajo učenci, je moč opaziti, da so ene kolokacije rabljene prepogosto, druge pa preredko. Kolokacije so za učence težje obvladljive kot za naravne govorce. Poleg tega znanje maternega jezika vpliva na rabo kolokacij v tujih jezikih (Granger, 1998; De Cock et al., 1998; Komori, 2004). Nation (2001) meni, da je breme učenja (ang. *learning-burden*) kolokacij povezano z njihovo predvidljivostjo. Vpliv bremena je torej odvisen od tega, v kolikšni meri je možno

6 »The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices« (Načelo idioma pomeni, da imajo jezikovni uporabniki na voljo številne napol predpripravljene zveze, ki predstavljajo posamezne sporočanjske izbire; prev. I. S.) (Sinclair, 1991).

predvideti kolokacije v maternem ali že naučenem tujem jeziku. Pri uvajanju kolokacij je nujno razmisliti, katere kolokacije so zelo frekventne in katere zelo pogoste besede se lahko pojavljajo kot nepredvidljive kolokacije (Nation, 2001).

Vzporedno z razvojem korpusnega gradiva se je hitro razvijala tudi tehnologija za iskanje kolokacij. Razvoj konkordanc se je začel v 80. letih 20. stoletja. Z uporabo statističnih metod, npr. MI (ang. *mutual information*), so raziskovalci začeli računati statistično pomembnost kolokacij (Church in Hanks, 1989). Pozneje se pojavijo korpusni iskalni sistemi, ki lahko pridobijo sezname kolokacij po padajočem vrstnem redu glede na statistične vrednosti, kot so npr. WordSmith,[7] Sketch Engine (SkE)[8] in The IMS Open Corpus Workbench[9] (CWB). Orodje SkE, ki so ga izdelali Kilgarriff et al. (2004) najprej za potrebe angleščine, pozneje pa tudi za druge jezike, gre še korak naprej in ponuja dodatne funkcije z naprednim načinom iskanja in prikaza leksikalnih informacij. Zaradi tega se sistem pogosto uporablja za sestavljanje slovarjev, na primer slovarja Macmillan Dictionary for Advanced Learners of English (Rundell, 2002) ter za poučevanje tujih jezikov in jezikoslovne raziskave.

Slovarji, osnovani na korpusih, vključujejo podatke o kolokacijah in so še posebno pogosti v angleškem okolju (npr. Collins Cobuild Dictionary, Longman Dictionary of Contemporary English, Oxford Advanced Learner's Dictionary of Current English). Obstaja tudi angleško-japonski slovar, ki temelji na korpusu, in sicer Wisdom English-Japanese Dictionary (Inoue in Akano, 2003). Izdelan je bil tudi prvi slovar kolokacij za učence japonščine (Himeno, 2004, 2012), toda podatki o kolokacijah v teh slovarjih niso zadostni, tako da je nujno nadaljnje delo na tem področju (Siepmann, 2005, 2006; Ogino et al., 2006; Ogino, 2008). Ena od tovrstnih raziskav je bila vzpostavljena v okviru projekta Korpus japonščine, usmerjena pa je bila v izdelavo in vrednotenje japonskega slovarja kolokacij (Ogino, 2008). V zadnjem času je prišlo do zavedanja o pomembnosti podatkov o kolokacijah za učence japonščine (Noda, 2007), in zdaj so tovrstni podatki podani tudi v japonskih slovarjih (Muraki, 2007). Poleg tega potekajo raziskave o neustrezni uporabi kolokacij in o tem, kako jo preprečiti (Cao in Nishina, 2006; Oso in Takizawa, 2003).

1.4 Luščenje kolokacij japonskega jezika iz korpusov

Na področju obdelave naravnega jezika že obstajajo številne raziskave o pridobivanju kolokacij v japonščini. Prvi koristen vir o kolokacijskih podatkih je *Nihongo*

7 http://www.lexically.net/wordsmith/ (dostop 11.4.2015)
8 http://www.sketchengine.co.uk/ (dostop 11.4.2015)
9 http://cwb.sourceforge.net/ (dostop 11.4.2015)

kyouki jisho »Slovar japonskih kolokacij«（日本語共起辞書), ki je bil razvit v 80. letih kot del slovarja in korpusa EDR (EDR, 1994). Slovar kolokacij opisuje ujemalne strukturne odnose,[10] in sicer sklonske odnose glagolov, prisamostalniški (adnominalni) odvisni odnos, vezni atributivni odnos pregibne besedne vrste, števnike, ki modificirajo nepregibne besedne vrste in sklonski členek *no*（の）v odnosu z nepregibnimi besednimi vrstami.[11] Viri, ki so bili prvotno razviti za računalniško obdelavo jezikov, se v določeni meri uporabljajo tudi pri učenju in poučevanju japonščine (gl. naslednje poglavje).

Še posebno pogoste so raziskave kolokacij s sklonskimi členi. Kawahara in Kurohashi (2006) sta izdelala orodje za obsežno luščenje kolokacij, tako imenovani sklonski okvir (ang. *case frame*),[12] ki je pridobljen samodejno iz spletnih podatkov. Sklonski okvir sestavljajo pregibne besedne vrste ter njihove povezave s samostalnikom in so urejene po pravilih pregibnih besednih vrst. Na spletu je tako mogoče poiskati povezave samostalnik + sklonski členek + glagol ter primere iz izvirnega besedila.[13]

Raziskovalci poskušajo kolokacijske podatke uporabiti tudi pri obdelavi naravnega jezika, primeri tovrstnih raziskav so denimo izdelava semantične mreže z uporabo kolokacijskih podatkov (Akama et al., 2008), izdelava semantičnega koncepta, ki temelji na kolokacijski verjetnostni odvisnosti med besedami (Kameya in Sato, 2005), in analitična raziskava sestavljenih samostalnikov z uporabo kolokacijskih podatkov (Kobayashi et al., 1996).

Razvitih je bilo tudi nekaj japonskih korpusnih iskalnih sistemov, ki lahko iščejo kolokacijske podatke. S sistemom za luščenje kolokacijskih podatkov *Chakoshi* (茶漉)[14] (Fukada, 2007; Fukada in Oso, 2007) lahko prikažemo kolokacijske sezname iskanih besed, pridobljenih iz spletnega korpusa Aozora bunko (青空文庫コーパス) in govornega korpusa Univerze v Nagoyi（名古屋大学会話コーパス). Korpusno orodje *Chaki*（茶器）poleg funkcije iskanja strukturno razčlenjenih japonskih besedil omogoča tudi statistično obdelavo kolokacij (Matsumoto et al., 2009). V okviru projekta Korpus japonščine sta bili za pregledovanje korpusa BCCWJ razviti še dve orodji, Shonagon[15] in Chunagon[16].

10 Jp. *kakariuke* (係り受け)

11 http://www2.nict.go.jp/out-promotion/techtransfer/EDR/JPN/TG/Doc/EDR_J07a.pdf_(dostop 14.9.2014). Opis vsebine slovarja v izvirniku je naslednji: 日本語共起辞書に記述している共起関係は、格関係、連体修飾関係、連用修飾関係、助数詞による　体言の修飾関係の４種類に類別される。格助詞「の」を介した体言間の連体修飾関係や、助詞の「に」「で」などを介した格関係以外の連体修飾関係も格関係に準じて記述されている。

12 Jp. *kaku fureemu*（格フレーム）

13 http://nlp.kuee.kyoto-u.ac.jp/nl-resource/caseframe.html (dostop 14. 9. 2014).

14 http://tell.fll.purdue.edu/chakoshi/public.html (dostop 14. 9. 2014).

15 http://www.kotonoha.gr.jp/shonagon/ (dostop 14. 9. 2014).

16 http://chunagon.ninjal.ac.jp/ (dostop 14. 9. 2014).

Shonagon omoča iskanje niza znakov (ang. *string search*, jp. *mojiretsu kensaku*) po korpusu, ki ni označen z besednovrstnimi oznakami in drugimi podatki, Chunagon pa omogoča več različnih načinov iskanja po podrobno označenem korpusu. V obeh orodjih ni mogoče neposredno iskati kolokacij, mogoče pa je pridobiti rezultate s pomočjo nadaljne analize podatkov v tabelaričnem formatu.

Zelo napreden sistem za raziskovanje kolokacij je že omenjeno orodje Sketch Engine, ki v japonski inačici s korpusom JpWaC obstaja od leta 2008 (Srdanović et al., 2008), leta 2013 pa je bil nadgrajen z uporabo obsežnega spletnega korpusa JpTenTen (Srdanović et al., 2013). Izdelavo japonske verzije sistema podrobno predstavimo v 2. poglavju. Na podoben način je zasnovano orodje Natsume,[17] katerega osnovni namen je pomoč učencem japonskega jezika pri pisanju spisov, orodje pa pridobiva in primerja številne kolokacijske odnose iz različnih zvrsti oz. različnih tipov korpusov in podkorpusov (Nishina, 2008; Hodošček in Nishina, 2012). Podoben način iskanja in prikaza kolokacij ponuja tudi sistem NINJAL-LWP z uporabo korpusov BCCWJ in TWC (Tsukuba Web Corpus) (Pardeshi in Akasegawa, 2010).

Polega tega lahko uporabimo še korpusne iskalne sisteme, kot so Himawari,[18] TextFinder,[19] AntConc[20] ali sisteme za iskanje po spletu, npr. WebCorp,[21] ki lahko do neke mere prikažejo kolokacijske podatke. V zadnjih desetih do petnajstih letih so bile opravljene različne jezikoslovne raziskave kolokacij z uporabo korpusov in zgoraj omenjenih korpusnih orodij (npr. Backhouse, 2004; Takizawa, 2006, 2007).

Na voljo so tudi drugi viri, ki prvotno sicer niso bili namenjeni učencem japonščine, vendar so do določene mere v uporabi tudi na tem področju. Tovrstne vire nekoliko podrobneje predstavimo v naslednjem poglavju.

1.5 Razvoj pripomočkov za učenje kolokacij

Sočasno z razvojem raziskav na področju kolokacijskih povezav so raziskovalci začeli izpostavljati pomembnost učenja kolokacij, a kljub velikemu številu učencev in poudarjanju pomembnosti učenja kolokacij pri poučevanju japonščine za to še vedno ni dovolj učnega gradiva, denimo slovarjev (Kjellmer, 1991; James, 1998; Lewis, 2000). Ogino et al. (2006) glede opisov kolokacij v slovarjih in poskusnih slovarjih pravijo naslednje:

17 https://hinoki-project.org/natsume/ (dostop 14. 9. 2014).
18 http://www.kokken.go.jp/lrc/index.php (dostop 14. 9. 2014).
19 www.biwa.ne.jp/~lagoinst/textfinder.htm (dostop 14. 9. 2014).
20 http://www.laurenceanthony.net/software/antconc/ (dostop 14. 9. 2014).
21 http://www.webcorp.org.uk/live/ (dostop 14. 9. 2014).

Slovarski opisi japonščine so v zadnjem času dosegli izjemen napredek, vendar pa pravih slovarjev, ki bi uporabljali obsežne korpuse, še vedno nimamo. Če pogledamo vsebino slovarjev, je semantični opis besede podroben, vendar pa je opis povezav in odnosov besede z drugimi besedami izredno pomanjkljiv. Tovrstni podatki so pri učencih japonščine močno zaželeni. V prihodnjih slovarjih bodo še posebno pomembni.

Angleški Longmanov slovar sodobne angleščine (Longman Dictionary of Contemporary English) in nekatera raziskovalna podjetja na področju leksikografije zelo dobro vključujejo kolokacije v slovarje. Za japonščino je tu Himeno Masako (2004), toda želeli bi si jih več.

Slovar japonskih kolokacij (Himeno, 2004) je prvi slovar za učence japonščine, ki vsebuje kolokacijske izraze. Omenjeni slovar vsebuje veliko stavčnih primerov in kolokacijskih podatkov in je posledično koristen vir, vendar ima dve pomanjkljivosti. Prva je neuravnoteženost korpusa. Slovar je bil izdelan v času, ko uravnoteženega korpusa japonskega jezika še ni bilo na voljo, tako da je nastal na podlagi različnih jezikovnih virov, kot so drugi slovarji, literarna dela, časopisni članki. Pri posameznih primerih ni omenjeno, iz katerih virov so pridobljeni, zato težko sklepamo, ali so pridobljeni z upoštevanjem visokofrekvenčnih vzorcev iz uravnoteženega gradiva. Druga pomanjkljivost je ta, da je vrsta kolokacijskih odnosov, vključenih v slovar, omejena in zajema le glagole ter pridevnike. Terashima in Moriguchi (2005) sta naredila primerjavo tega slovarja z angleškim slovarjem kolokacij Oxford Collocations Dictionary for Students of English in pokazala, katere segmente bi bilo treba v prihodnjih slovarjih izboljšati.

Druga omenjena pomanjkljivost slovarja je bila odpravljena z novo, obsežnejšo izdajo (Himeno, 2012), ki vključuje več različnih kolokacijskih odnosov, vendar pomanjkljivost v zvezi z uporabljenim gradivom ostaja. To je vidno med rezultati, ki v določenih primerih ponujajo kolokacije iz moderne in sodobne japonščine ali kolokacije, vezane na specifično zvrst, npr. leposlovje, to pa v slovarskem opisu ni izpostavljeno.

Kolokacijski slovar EDR, ki je omenjen v poglavjih 1.1 in 1.4, je bil uporabljen pri izdelavi dela *Nihongo doushi no ketsugouka* »Vezljivost japonskih glagolov«（日本語動詞の結合価）(Ogino et al., 2003). EDR-jev glagolski koncept je bil izdelan iz pridobljenih odvisnikov in primerov vezljivosti na približno 9400 glagolih. To obsežno bazo podatkov je mogoče vsestransko uporabiti pri obdelovanju podatkov ter pri poučevanju in raziskovanju japonščine (Ogino et al., 2007).

V sklopu projekta Korpus japonščine so bile opravljene raziskave v zvezi z urejanjem japonskega slovarja kolokacij, predlagana je bila tudi metoda za sestavo kolokacijskega slovarja z uporabo spletnega korpusa (Ogino, 2009).

Za poučevanje in učenje kolokacijskih odnosov v japonščini lahko uporabljamo tudi korpusne iskalne sisteme. Nekateri so bili razviti posebej za poučevanje in raziskovanje japonščine, denimo orodje KH Coder, ki lahko izbira glagolske vzorce (Sano in Lee, 2007), ali že omenjeno orodje Natsume (Nishina, 2008; Hodošček in Nishina, 2012). Poleg tega obstajajo tudi raziskave na področju razvoja osebnih učnih pripomočkov, ki podpirajo učenje kolokacij. To so na primer spletne konkordance za učenje kombinacij s pridevniki in prislovi (Sato et al., 2007).

Obenem so raziskovalci na področju poučevanja japonščine v zadnjem desetletju začeli iskati načine, kako uporabiti korpuse pri razvijanju učnih načrtov za usvajanje besedišča. Hashimoto (2008) tako predlaga, da se k učnemu načrtu za besedišče pripiše oznako »tema«, zato da poveže učenčeve vsakodnevne »jezikovne aktivnosti in naloge« z besediščem. Z dodano oznako težavnostne stopnje se potem znotraj iste teme razporedi besede po vrstnem redu od lažjih do težkih. Rezultati raziskave so povzeti v Yamauchi (2008),[22] obstaja pa tudi novejša študija o primerjanju več različnih virov učnih načrtov (Yamauchi, 2015). V prihodnje bo tako treba razmisliti o učnem načrtu za poučevanje japonščine, ki bo vključeval ne le besedišče, temveč bo sistematično zajemal tudi kolokacijske podatke (gl. npr. Srdanović, 2013, 2014).

1.6 Kolokacije na daljavo in ujemanje med prislovi in modalnimi oblikami na koncu stavka

Pojav kolokacij na daljavo v korpusnem jezikoslovju ni kaj dosti uveljavljen, načeloma pa velja, da razdalja med sopojavnimi besedami šteje do pet besed (Sinclair, 1991; Church in Hanks, 1990; Partington, 1998; Manning in Schütze, 1999). V raziskavah o kolokacijah najdemo tudi izraze »oddaljena kolokacija« (ang. *distant collocations*), »prekinjena kolokacija« (ang. *interrupted collocations*) (Ikehara et al., 1996) in »pretrgana kolokacija« (ang. *discontinous collocations*) (Milkov, 2001). Pri kolokacijah na daljavo morata biti med besedama, ki se sopojavljata, vsaj ena do dve besedi, in kar je manj od tega, ne moremo uvrstiti med kolokacije na daljavo. V japonščini je to tradicionalno prepoznano kot odvisno razmerje, npr. med prislovom in obliko modalnosti na koncu stavka.[23] Odvisno razmerje se obravnava kot

22 Osnutek standarda za besedišče za učenje japonskega jezika, jp. *Nihongo kyouiku sutandaado shian: goi* 『日本語教育スタンダード試案　語彙』.

23 Modalnost opredeljuje pomen govornega dejanja in kot takšna v osnovi prikazuje način izražanja govorca v času govorjenja (Moriyama, 2000). Modalne oblike na koncu stavka so oblike kot npr. *darou* (だろう), *kamo shirenai* (かもしれない), *rashii* (らしい), ki označujejo predvidevanje.

razmerje v stavku, pri katerem so posamezne besede navezane na druge. Odvisnost oz. ujemanje modalnih prislovov in modalnih oblik na koncu stavka v japonščini so že bili predmet raziskovanja (Minami, 1974; Kudō, 2000; Bekeš, 2006, 2008a,b). Z vidika korpusnega jezikoslovja se takšno razmerje, kjer je opazna verjetnost sočasnega pojavljanja, obravnava kot močna kolokacijska povezava, ki jo opazujemo kot pogosto v obsežnem jezikovnem gradivu in je statistično gledano utemeljena.[24]

V nekaterih primerih se kolokacije pojavljajo na razdalji, kar imenujemo kolokacije na daljavo (gl. primer 1-1). Pomembno je opomniti, da se kolokacije na daljavo ne pojavljajo le na razdalji do pet besed, ampak je ta razdalja lahko tudi večja.

Primer 1-1:

Tabun ashita hareru deshou。

»Verjetno se bo jutri razvedrilo«

Verjetno/jutri/razvedriti se/<modalna oblika:predvidevanje>

prislov (Adv) + modalna oblika na koncu stavka (Mod)

močno ujemanje/kolokacijski odnos
(Minami, 1974; Kudō, 2000; Bekeš, 2006)

Kolokacija prislova in modalne oblike lahko pogosto oklepa druge besede. Ko je Minami (1974) odkril večplastnost povedi v japonščini, je predlagal, da se to večplastno strukturo označi s štirimi stopnjami ABCD. To kaže naslednji primer: plast A) povedek *iru* »biti« (いる), plast B) jedro stavka *kono machi ni mo gonin gurai wa iru* »tudi v tem mestu je vsaj približno pet ljudi« (この町にも五人ぐらいはいる), plast C) prislov *douyara* »videti je, da« (どうやら) in modalni izraz *rashii* »bojda« (～らしい), kar ustreza kolokaciji na daljavo.

24 Kolokacija z obliko modalnosti na koncu stavka se lahko obravnava kot sopojavljanje besed in slovnične kategorije, zato je možna tudi uporaba izraza koligacije (ang. *colligation*, jp. *korigeeshon* コリゲーション). V nasprotju s tem pa kolokacijo besede in besede poimenujemo kolokacije (ang. *collocation*, jp. *korokeeshon* コロケーション) (Ishikawa 2008).

Primer 1-2 (primer je iz korpusa NUJCC v Bekeš, 2006):

{どうやら [この町にも五人ぐらいは(い)$_A$る]$_B$らしい}$_C$

{*Douyara* [*kono machi nimo gonin gurai wa* (*i*)$_A$ *ru*]$_B$ *rashii*}$_C$

{Videti je, da [to / mesto / v / tudi / pet ljudi /vsaj približno / <členek_topik> (<osnova gl. biti>)$_A$ <slovarska gl. oblika>]$_B$ <modalni izraz_domnevanje>}$_C$

»Izgleda, da je tudi v tem mestu vsaj približno pet ljudi.«

Raziskave o modalnosti in prislovih v japonščini sta opravila Nitta (2002) in Masuoka (1991), in čeprav sta raziskavi ločeni, je pri obeh poudarek na kolokacijskih odnosih med prislovi in modalnimi oblikami. Tudi Kudō (2000) podobno kot Minami uporablja empirični pristop. Med raziskavo povezave med modalnim prislovom in povedno modalnostjo je razjasnil stopnjo kolokacijskega odnosa med posameznim prislovom in specifično modalno obliko ter to močno povezavo obravnaval kot ujemanje oz. kolokacijsko povezavo. Prislovi ugibanja vsebujejo določeno stopnjo prepričanja, ki jo izkazuje govorec v odnosu do uresničitve situacije (Kudō, 2000). Glede na to stopnjo prepričanja govorca so prislovi razdeljeni v štiri skupine, ki so po intenziteti razporejene na: »nujnost«, npr. *ni chigai nai* in *hazu da* (NEC, necessity), »pričakovanje«, npr. *darou* in *to omou* (EXP, expectation), »domnevanje«, npr. *rashii* in *mitai* (CON, conjecture), in »možnost« , npr. *kamo shirenai* ter *kana* (POSS, possibility) (Kudō, 2000). Na prislov *tabun* »verjetno« se na primer močno vežejo modalne oblike predvidevanja *darou*, *no darou ka*, *to omou* in *no de wa nai ka*. Podrobnosti o Kudōjevih podatkih sledijo v poglavju 4.3.

V raziskavi analize pogovorov v japonščini (Bekeš, 2006) je bila kolokacija med prislovom in modalno obliko na koncu stavka za desetino višja kot pa samopojavljanje modalne oblike na koncu stavka. Kolokacije, ki se pojavljajo v takšnem razmerju, so s stališča jezikovne teorije zaznamovane (Halliday, 1991) in rečemo lahko, da ima tolikšen delež kolokacij pri modalnih izrazih pomembno mesto. Po drugi strani, gledano z vidika prislovov ugibanj, je primerov kolokacij s prislovi ugibanj in modalno obliko na koncu stavka veliko, zato je ta pojav mogoče sistematično raziskovati. Pri poučevanju japonščine kolokacijski odnosi med prislovi in modalno obliko na koncu stavka niso bili prepoznani, zato se jih tudi ni ciljno poučevalo. V drugih jezikih imamo različne odnose odvisnosti, toda v jezikih SOV,[25] kot sta npr. korejščina in tamilščina, lahko opazimo podobne pojave odvisnosti

25 SOV: ang. *subject – object – verb*, slo. *osebek – predmet – glagol*

kot v japonščini, ki je tudi jezik SOV. Po drugi strani pa v SVO-jezikih, na primer kitajščini, angleščini in tudi v slovanskih jezikih, ki imajo prosto stavo, ne najdemo odvisnih odnosov oz. kolokacijskih razmerij med prislovi in modalnostjo na koncu stavka kot pri japonščini. Predvidevamo lahko, da ima zaradi maternega jezika veliko učencev japonščine težave pri predvidevanju kolokacijskih odnosov. V tem smislu je nujnost po učenju kolokacijskih odnosov še toliko večja.

Bekeš (2008a) je analiziral tendence ujemanja prislova ugibanja in modalne oblike na koncu stavka v konverzacijskem gradivu japonskega jezika, ter podal koristne predloge za poučevanje japonščine. Prvič, modalna oblika, ki se sopojavlja po določenem sistemu, pospešuje razumevanje vsebine konverzacije, zato je priporočljivo, da se učencem pokaže ta sistem kolokacij in se jih na ta sistem privadi. Drugič, priporočljivo je učencem jasno razložiti, da sopojavnost prislova in modalne oblike na koncu stavka pomensko funkcionira kot ena enota.

Na področju poučevanja japonščine obstajajo tudi raziskave, ki se ukvarjajo posebno s prislovi in modalnimi oblikami na koncu stavka (npr. Hirata, 2001; Sunakawa, 2007), vendar ni dovolj sistematičnih raziskav, temelječih na korpusih, in opisa tega pojava v učbenikih za japonski jezik.

1.7 Povezave s pričujočo raziskavo

V zadnjem desetletju je prišlo do razvoja raziskav za podporo učenju kolokacij v japonščini ter pripravi učnih gradiv in slovarjev kolokacij za učence japonskega jezika, vendar pa je še precej prostora za tovrstne raziskave in izdelavo referenčnih materialov o kolokacijah za učence. Pojavilo se je tudi nekaj korpusnih orodij, ki omogočajo luščenje kolokacijskih podatkov, vendar se je treba zavedati različnih omejitev teh orodij in delati na nadgrajevanju in dodajanju funkcij, tako da se čim učinkoviteje in izčrpneje pridobijo statistično obdelani seznami kolokacij za iskane besede. V pričujoči monografiji opišemo potek oblikovanja japonske različice besednih skic korpusnega iskalnega sistema, ki omogoča izčrpno luščenje kolokacij in kolokacijskih odnosov iz obsežnih podatkovnih baz in prikaže rezultate v obliki urejenega seznama kolokacij. Takšen sistem je lahko uporaben pri sestavi učnega gradiva za japonščino in lahko prispeva k izboljšavi slovarjev kolokacij japonskega jezika.

O odvisnih in ujemalnih odnosih, ki imajo dolgo raziskovalno tradicijo v japonski slovnici, že obstajajo številne teoretične raziskave. Med njimi so tudi raziskave o odnosu povednih prislovov ugibanja in modalnih oblik na koncu stavka, a o tem pojavu primanjkuje korpusno temelječih raziskav, prav tako pa ga japonski

učbeniki obravnavajo nezadostno. Zato se v pričujoči monografiji, ki temelji na rezultatih teoretičnih raziskav o prislovih in modalnih oblikah ter njihovih odvisnostih, osredotočimo na raziskovanje verjetnosti pojavljanja teh kolokacijskih povezav v obsežnem korpusnem gradivu japonskega jezika ter na preučevanje uvajanja omenjenega pojava na področju izobraževanja. Z novega gledišča raziščemo prislove ugibanja in modalno obliko na koncu stavka kot kolokacijski odnos (na daljavo) in pojasnimo kolokacijske trende v korpusih. Poleg tega razvijemo metodo označevanja modalnih oblik in tako omogočimo njihovo analizo s korpusnim iskalnim sistemom in pridobimo kolokacije modalnih oblik. Na podlagi pridobljenih rezultatov prikažemo predloge za izboljšavo opisov kolokacij, ki do zdaj še niso bili sistematično uporabljeni v učbenikih in slovarjih.

Med raziskovanjem možnosti uporabe obsežnih korpusov japonščine v poučevanju se dotaknemo tudi metod za vrednotenje korpusov s pomočjo njihove analize in kategorizacije glede na razpršenost modalnih prislovov in oblik.

2 Razvoj japonske različice orodja za luščenje kolokacij[26]

2.1 Uvod

V času, ko so se korpusi v jezikoslovju začeli širiti, tj. v 80. letih, je težišče takratnih razprav bilo, ali bi se korpusi morali uporabljati ali ne. Po skoraj desetih letih so se razprave obrnile k problematiki obširnosti in reprezentativnosti, obsežni seznami podatkov pa so postali še posebno aktualna tema s širjenjem rabe interneta. Intenzivne razprave so se dotikale tudi vprašanja, kako lahko na najhitrejši možni način pridobimo primerne podatke iz obsežnih korpusov. V 80. letih so bile razvite prve konkordance za luščenje podatkov iz korpusov, tovrstna orodja pa zdaj veljajo že za tradicionalna (Kilgarriff in Rundell, 2002).

Iskanje po obsežnem korpusu je težavno tudi v obliki konkordanc, če imamo za iskano besedo 500, 1000 ali 20.000 in več prikazanih primerov. Zaradi tega so okoli leta 2000 začeli razvijati korpusne iskalne sisteme, ki poleg konkordanc vsebuje tudi druge funkcije (Heid et al., 2000; Kilgarriff in Tugwell, 2001). Eden od sistemov je Sketch Engine (SkE) (Kilgarriff et al., 2004), ki vključuje napredne funkcije za hitro luščenje kolokacijsko-slovničnih odnosov ter ponuja različne tipe jezikovnih podatkov. Orodje SkE se že več let v številnih jezikovnih različicah uporablja na področju korpusnega jezikoslovja, korpusne leksikografije ter za namene poučevanja drugega tujega jezika.

V tem poglavju predstavimo orodje SkE in opišemo metodologijo ter rezultate razvoja japonske različice. Za razvoj japonske različice sta bila potrebna dva glavna koraka: 1) gradnja obsežnega spletnega korpusa (JpWaC) s 400 milijoni besed in implementacija v orodju SkE in 2) priprava slovnične datoteke[27] za japonščino ter uvoz te datoteke v SkE. V tretjem poglavju sledi vrednotenje japonske različice in predstavitev njene veljavnosti pri uporabi.

2.2 Povzetek delovanja sistema

Orodje SkE v osnovi uporablja korpusni iskalni sistem Manatee, poleg tega vsebuje še posebne načine iskanja in prikaze pridobljenih podatkov. Najprej je bil izdelan za angleški jezik, pozneje pa je bil večkrat prirejen za uporabo v različnih jezikih.

26 Določeni rezultati raziskave predstavljene v tem poglavju so objavljeni v reviji *Shizen gengo shori* (*Journal of Natural Language Processing*) (Srdanović et al., 2008: o izdelavi japonskega korpusa JpWaC in japonskih besednih skicah) v *Nihongo kagaku* (*Japanese Linguistics*) (Srdanović & Nishina, 2008: o izdelavi besednih skic za japonski jezik in njihovi uporabnosti).

27 Jp. *bunpou kankei fairu*（文法関係ファイル）, ang. *gramrel file*.

Če so na voljo potrebni viri, orodja in strokovnjaki za določen jezik, je različica v tem jeziku sestavljena iz specifičnega jezikovnega korpusa in slovnične datoteke, ki je izdelana tako, da vsebuje specifično slovnico in kolokacijske povezave tega jezika. Korpusni vhodni formati temeljijo na standardu »ena beseda na vrstico«, ki ga je razvil Christ (1994).[28] Vsaka beseda je v novi vrstici, za vsako besedo pa so v tabulatorju uporabljeni podatki za pojavno obliko oz. pojavnico (besede in morfemi v obliki, kot se pojavljajo v korpusu), lemo (izbrana reprezentativna beseda ali morfem, ki vključuje vse pregibne oblike te besede oz. morfema), besedno vrsto ipd. Slovnična datoteka je pripravljena s pomočjo korpusne iskalne skladnje,[29] ki jo je predlagal Gahl (1998) in temelji na regularnih izrazih, besednih vrstah in besedah.[30]

Statistika,[31] ki se jo lahko uporabi v sistemu SkE, so vzajemne informacije[32] (MI--score, Church in Hanks, 1989), T-score, MI3-score, log-likelihood, minimum sensitivity, MI.log-f, Dice in relativna frekvenca. Pri besednih skicah je uporabljena statistika DiceLog (jp. *daisu rogu* ダイスログ)[33] in na podlagi tega je izračunana kolokacijska izpostavljenost (jp. *seriansu*セリアンス).[34] DiceLog temelji na koeficientu podobnosti (jp. *daisu keisuu* ダイス係数), z njim pa merimo frekvenco kolokacijskih besed v posebnem kolokacijskem razmerju, ki sestoji iz treh delov (beseda 1, kolokacijska povezava, beseda 2). Na primer *atsui oyu* »vroča/topla voda« (熱いお湯), *atsui* »vroč« (熱い) + Modifier_Ai + *oyu* »vroča/topla voda« (お湯) je prepoznana kot ena izmed kolokacijskih povezav za pridevnik + samostalnik.

V nadaljevanju so predstavljeni postopki za izdelavo japonske različice, predstavljeni pa so tudi njene glavne funkcije in primeri rezultatov iskanja v japonščini.

2.3 Priprava in lastnosti spletnega korpusa

Kot prvi korak pri izdelavi japonske različice je opisan način priprave spletnega korpusa in vnos v sistem SkE. Nadalje so opisane specifične lastnosti spletnega korpusa in s tem povezane razprave ter rezultati najnovejših raziskav.

28 Stuttgart Corpus Tools (Christ, 1994).

29 Ang. *corpus query syntax*; jp. *koopasu kensaku shintakusu*コーパス検索シンタクス

30 Za podrobnosti glej http://www.sketchengine.co.uk/documentation/wiki/SkE/CorpusQuerying (dostop 14. 4. 2015).

31 Za podrobnosti glej http://www.sketchengine.co.uk/documentation/attachment/wiki/SkE/DocsIndex/skestat.pdf (dostop 14. 4. 2015).

32 *Sougojouhouryou*（相互情報量）

33 Vrednost je bila predlagana v doktorski disertaciji (Curran, 2004; *From Distributional to Semantic Similarity*), Univerza v Edinburgu. Za več informacij gl. Rychlý (2008).

34 Izpostavljenost (ang. *salience*) predstavlja statistično pomembnost korelacije v korpusu.

2.3.1 Struktura in značilnosti spletnega korpusa

JpWaC je širokoobsežni japonski korpus s 400 milijoni besed oz. morfoloških enot. Korpus je bil zgrajen z metodami, ki jih predlagajo Sharoff (2006a,b), Ueyama in Baroni (2005), Baroni in Kilgarriff (2006) s približno 50.000 zbranih spletnih strani.[35] Način izdelave korpusa je v skladu z raziskavami projekta Wacky (Baroni in Bernardini, 2006), uporabljena pa so bila orodja, kot sta npr. BootCat (Baroni in Berdardini, 2006, Baroni et al., 2006) in WAC5 (Web as Corpus Toolkit). Cilj je bil pripraviti uravnoteženi korpus, zato so bile spletne strani izluščene z natančno metodo, ki jo predlaga Sharoff (2006a). V ta namen je bilo uporabljenih 500 visokofrekvenčnih besed iz korpusa BNC, ki so bile prevedene v japonščino. Za odstranjevanje neželenih oznak, kot so npr. HTML-oznake, je bilo uporabljeno orodje za t. i. *boilerplate removal*. Tako so bile odstranjene odvečne oznake, kot so navigacijski okvir, šifre in povezave. Z orodjem ChaSen je nato bila opravljena morfološka analiza, ki je določila pojavnice, leme in dodala besednovrstne oznake (izvorne japonske besedne vrste so bile še dodatno prevedene v angleščino). Tabela 2 kot primer prikazuje rezultate analize japonskega stavka とても幸せな気分でした。(*Totemo shiawasena kibun deshita* »Bil sem zelo veselo razpoložen«) z orodjem ChaSen. Besedne vrste in dodatni podatki o obliki se v orodju ChaSen prikažejo v japonščini, v tabeli so prevedeni iz japonščine v slovenščino. Angleški prevodi besednih vrst niso prikazani v orodju ChaSen, ampak so dodani v japonski različici orodja SkE (glej desni stolpec Tabele 2).

35 Za izdelavo spletnega korpus JpWaC je Adam Kilgarriff priskrbel seznam zbranih spletnih strani, ki jih je pripravil Serge Sharoff. Tomaž Erjavec je potem opravil vse potrebno za zbiranje in označevanje korpusa, avtorica monografije pa je pripravila ustrezne angleške oznake na podlagi japonskih oznak morfosintaktičnega analizatorja ChaSen, ki so bile zatem uporabljene za označevanje korpusa.

Tabela 2: *Analiza stavka* Totemo shiawasena kibun deshita *»Bil sem zelo veselo razpoložen«* (とても幸せな気分でした。) *z orodjem ChaSen.*

Pojavnica	Kana	Lema	Besedna vrsta	Dodatno o obliki	Besedna vrsta (ANG)
とても	ﾄﾃﾓ	とても	prislov		Adv.P
幸せ	ｼｱﾜｾ	幸せ	samostalnik – osnova pridevnika na *-na*		N.A na
な	ﾅ	だ	pomožni glagol	nepregibna besedna vrsta	Aux
気分	ｷﾌﾞﾝ	気分	samostalnik		N.g
でし	ﾃﾞｼ	です	pomožni glagol	vezna oblika	Aux
た	ﾀ	た	pomožni glagol	osnovna oblika	Aux
。	。	。	ločilo		Sym.p

2.3.2 Statistika spletnega korpusa

V tem podpoglavju je prikazana statistika korpusa JpWaC, in sicer velikost, značilnosti podatkov in razlike v besednih vrstah.

Korpus zajema 7,3 GB podatkov. Tabela 3 prikazuje število zbranih spletnih strani (npr. http://www.arsvi.com/0e/ps01.htm), število gostiteljev spletnih strani (npr. www.arsvi.com) ter povprečno število strani glede na gostitelja in število strani dveh domen (.jp in .com).

Tabela 3: *Statistika spletnih podatkov v korpusu.*

49.544	spletne strani
16.072	gostitelji (host)
3,1	št. strani gostitelja
34.911	št. strani domene .jp
14.633	št. strani domene .com

Tabela 4 kaže ključne besede, ki se pojavijo več kot 1000-krat v URL-ju spletnih strani. Iz ključnih besed lahko razberemo tipe strani, ki se največkrat pojavijo v korpusu. Na primer blog.livedoor.jp (1646 strani), www.amazon.co.jp (1048 strani), d.hatena.ne.jp (759 strani), blog.goo.ne.jp (690 strani).

Tabela 4: *Najpogostejše ključne besede v spletnih virih (število spletnih strani).*

6.486	Blog
3.471	Nifty
2.362	archives
2.075	livedoor
1.545	Diary
1.428	News
1.380	Cocolog
1.296	Exblog
1.051	Amazon
1.013	Archive
1.006	Geocities

Korpus, ki je bil analiziran z orodjem ChaSen, je sestavljen iz 12.759.191 povedi oz. 409.384.405 pojavnic. Tabela 5 kaže skupno število posameznih besed oz. razločnic (jp. *kotonarigo*, ang. *types*). Veliko je splošnih samostalnikov, neodvisnih glagolov in samostalnikov na *-sa*, poleg tega pa je tudi zelo veliko število neznanih besed, to so besede, ki se jih z orodjem ChaSen ne more analizirati. V korpusu je približno ena četrtina simbolov in posebnih znakov. Vsebuje pa tudi kitajske in klasične pismenke, izraze v katakani (npr. *burogu* »blog« ブログ, *merumaga* »elektronska revija« (メルマガ), *burauza* »brskalnik« ブラウザ, *keitai* »mobilni telefon« (ケイタイ), *kuraianto* »stranka« (クライアント) in angleške okrajšave (npr. IT, DVD, PC).

Tabela 5: *Skupno št. besed in št. razločnic v korpusu glede na besedno vrsto.*

Besedna vrsta (ChaSen)	Besedna vrsta (ANG)	Razlaga v SLO	Skupno št. besed	Št. razločnic
名詞-一般	N.g	Samostalnik (splošni)	50.121.542	49.268
動詞-自立	V.free	Glagol (neodvisni)	32.746.478	17.135
助詞-格助詞-一般	P.c.g	Členek (sklonski, splošni)	31.582.601	15
助動詞	Aux	Pomožni glagol	28.798.730	122
記号-アルファベット	Sym.a	Znak (abeceda)	27.583.164	152
名詞-サ変接続	N.Vs	Samostalnik (veza na *suru* glagol)	21.704.856	10.017
名詞-数	N.Num	Samostalnik (številnik)	20.160.119	50
記号-読点	Sym.c	Znak (vejica)	17.328.451	3
助詞-連体化	P.prenom	Členek (nominal)	14.786.763	1
助詞-接続助詞	P.Conj	Členek (vezni)	13.395.830	30
記号-一般	Sym.g	Znak (splošni)	13.312.393	93
記号-句点	Sym.p	Znak (točka)	13.260.780	3
助詞-係助詞	P.bind	Členek (vezni)	12.854.037	9
名詞-非自立-一般	N.bnd.g	Samostalnik (odvisni, splošni)	7.582.296	66
名詞-接尾-一般	N.Suff.g	Samostalnik (pripona, splošni)	6.945.415	639
動詞-非自立	V.bnd	Glagol (odvisni)	6.884.053	470
未知語	Unknown	Neznana beseda	6.857.818	182.823
記号-括弧閉	Sym.bc	Znak (zaprti oklepaj)	6.171.426	15
記号-括弧開	Sym.bo	Znak (odprti oklepaj)	6.061.836	15
名詞-副詞可能	N.Adv	Samostalnik (prislovni)	4.831.851	734
名詞-接尾-助数詞	N.Suff.msr	Samostalnik (pripona v številniku)	4.543.767	547
名詞-形容動詞語幹	N.Ana	Samostalnik (osnova pridevnika na *-na*)	4.436.970	2.892

Besedna vrsta (ChaSen)	Besedna vrsta (ANG)	Razlaga v SLO	Skupno št. besed	Št. razločnic
名詞-代名詞-一般	N.Pron.g	Samostalnik (zaimek, splošni)	4.081.314	108
形容詞-自立	Ai.free	Pridevnik na *-i* (neodvisni)	3.871.216	1.799
副詞-一般	Adv.g	Prislov (splošni)	3.603.890	2.303
助詞-格助詞-引用	P.c.r	Členek (sklonski, navajanje)	3.521.270	2
助詞-格助詞-連語	P.c.Phr	Členek (sklonski, sestavljenka)	3.514.071	88
連体詞	Adn	Prilastek/Adnominal	3.178.853	125
動詞-接尾	V.Suff	Glagol (pripona)	2.714.110	36
副詞-助詞類接続	Adv.P	Prislov (vezni)	2.537.640	517
接続詞	Conj	Veznik	2.450.712	169
名詞-非自立-副詞可能	N.bnd.Adv	Samostalnik (neodvisni, prislovni)	2.356.968	58
助詞-副助詞	P.Adv	Členek (prislovni)	2.302.728	27
助詞-並立助詞	P.coord	Členek (priredni vezni)	2.063.217	8
助詞-副助詞／並立助詞／終助詞	P.adv-coordfin	Členek (prislovni/ priredni vezni/ povedni)	2.005.823	1
接頭詞-名詞接続	Pref.N	Predpona (veza na samostalnik)	1.929.246	158

2.3.3 Vnašanje spletnega korpusa v sistem in funkcija konkordanc

Korpus JpWaC je integriran v sistem SkE, kar omogoča dostop do podatkov iz korpusa prek spletne strani SkE ter uporabo standardnih konkordanc in drugih načinov pregledovanja. Slika 1 kot primer prikazuje rezultate iskanja besede 意見 *iken* »mnenje« s pomočjo standardnih konkordanc. Nujno je iskanje z morfemi oz. besedami, kakor so določeni in analizirani s sistemom ChaSen.

Slika 1: *Rezultati iskanja s konkordancami v sistemu SkE.*

S pomočjo funkcije CQL (Corpus Query Language, jp. *koopasu kensaku gengo* コーパス検索言語) je z uporabo regularnih izrazov omogočeno iskanje kompleksnejših vzorcev. V nadaljevanju je podanih nekaj konkretnih primerov.

- Iskanje, ki vključuje več načinov zapisovanja ene besede: [word="きれい"| word="綺麗"]
- Iskanje enot, ki so analizirane v dveh delih ali več, kot so npr. besedne zveze, posamostaljeni glagoli, slovnični vzorci: [word="気"] [word="に"] [lemma="する"]
- Iskanje z besednimi vrstami: [tag="N.*"]&[word ="つる"]

2.3.4 Primerjava spletnega in korpusa časopisov

V tem poglavju predstavimo metodo in rezultate primerjanja korpusov JpWaC s korpusom časopisov. V analizo so bili vključeni podatki (30 milijonov besed) iz časopisa Mainichi shimbun (毎日新聞) iz leta 2002. Primerjalno metodologijo analize neskladja med korpusi sta predlagala Rayson in Garside (2000). S to metodo lahko med dvema korpusoma odkrijemo razlike pri besedah in besednih vrstah, in sicer tako, da najprej sestavimo seznam besed in njihovih frekvenc in nato izračunamo statistiko verjetnostnega algoritma (log-likelihood, v nadaljevanju LL) za vsako enoto posebej. Pri izračunavanju LL se upoštevata velikost posameznega korpusa ter frekvenca enot v vsakem korpusu. Večja kot je vrednost LL, v večji meri je izpostavljenost enote v enem korpusu večja kot v drugem. Začenši z enotami z najvišjo vrednostjo LL je mogoče odkriti največje razlike v dveh korpusih. Tabela 6 prikazuje razliko med enotami besednih oblik znotraj korpusa JpWaC in časopisnega korpusa glede na izračunano vrednost LL.

Tabela 6: *Največje razlike med korpusimi JpWaC (zgoraj) in časopisnimi podatki (spodaj) (po izračunih verjetnostnega algoritma; prikazanih je prvih dvajsetih besed).*

Besedna oblika	Oznaka	Razlaga (SLO)	LL	JpWaC	Časopis
ます	Aux	pomožna oblika *masu*	206122	5,76	074
です	Aux	pom. oblika *desu*	148844	4,55	0,70
まし	Aux	pom. oblika *mashi*	88599	2,69	0,41
月	N.Suff.msr	samostalniška pripona *gatsu/getsu* »mesec«	65734	1,12	0,00
て	P.Conj	vezni členek *te*	64966	21,04	14,54
で	Aux	pom. oblika *de*	59324	7,53	3,96
か	P.advcoordfin	členek *ka*	59069	4,90	2,10
の	N.bnd.g	odvisni sam. (sploš.) *no*	58270	6,31	3,10
な	Aux	pomožna oblika *na*	51056	6,63	3,52
ん	N.bnd.g	odvisni sam. (sploš.) *n*	42093	1,44	0,27
こと	N.bnd.g	odvisni sam. (sploš.) *koto*	36013	5,86	3,38
ん	Aux	pomožna oblika *n*	34973	1,22	0,24
ませ	Aux	pom. oblika *mase*	34454	1,07	0,17
私	N.Pron.g	sam. (zaimek, splošni) *watashi*	33163	1,72	0,53
ござい	Aux	pomožna oblika *gozai*	29663	0,55	0,01
その	Adn	prilastek *sono*	28998	2,25	0,93
もの	N.bnd.g	odvisni sam. (sploš.) *mono*	28935	1,80	0,64
それ	N.Pron.g	prilastek *sore*	28902	1,54	0,48
よう	N.bnd.Aux	odv. sam. -pomožna oblika *you*	28562	2,76	1,29
という	P.c.Phr	členek (fraza) *to iu*	28068	2,35	1,02
日	N.Suff.msr	sam. pripona *nichi* »dan«	110177	1,38	4,34
た	Aux	pom. obl. *ta*	104951	16,95	25,5

Besedna oblika	Oznaka	Razlaga (SLO)	LL	JpWaC	Časopis
同	Pref.N	samostalniška predpona (sam.) *dou*	93209	0,12	1,26
約	Pref.Num	predpona (številnik) *yaku* »okrog«	68451	0,17	1,20
市	N.Suff.p	sam. (pripona, regija) *shi*	64868	0,19	1,20
など	P.Adv	členek (prislovni) *nado*	64244	1,22	3,25
容疑	N.g	samostalnik (splošni) *yougi* »osumljenec«	63021	0,02	0,58
首相	N.g	samostalnik (splošni) *shushou* »predsednik vlade«	54740	0,05	0,67
円	N.Suff.msr	sam. pripona *en* »jen«	48315	0,48	1,67
人	N.Suff.msr	sam. pripona *hito* »človek«	46548	0,80	2,22
万	N.Num	sam. (številnik) *man* »10000«	41995	0,38	1,37
勝	N.Suff.msr	sam. pripona *shou* »zmaga«	40016	0,01	0,35
区	N.Suff.p	sam. pripona (regija) *ku*	39628	0,08	0,64
を	P.c.g	členek *wo*	38881	21,58	2,.2
東京	N.Prop.p.g	sam. (lastno ime, kraj) *Tokio*	38117	0,24	1,03
県	N.Suff.p	sam. pripona (provinca) *ken*	37005	0,17	0,85
で	P.c.g	členek *de*	36297	8,79	12,4
氏	N.Suff.n	sam. pripona na osebno ime *kun*	33179	0,28	1,04
後	N.g	samostalnik (splošni) *ato/go* »po (v časovnem smislu)«	33120	0,03	0,40
北朝鮮	N.Prop.p.c	sam. (lastno ime, regija, država) »Severna Koreja«	32973	0,05	0,48

Kot lahko vidimo pri enotah, podanih v nadaljevanju, je izpostavljenost v korpusu JpWaC večja kot v časopisih.

• Pomožni glagoli	*masu* ます, *desu* です, *mashi* まし, *mase* ませ Oblika *masu/desu* ます/ですse večkrat pojavi na spletu kot v časopisih.
• Modalne oblike	*ka* か, *deshou* でしょう, *you* よう, *omou* 思う, *node* ので, *wake* わけ
• Vljudnostne oblike	*o* お, *gozai* ござい
• Neformalne oblike	*tte* って, *yo* よ, *ne* ね, *n* ん
• Osebni zaimki in kazalnice	Osebni zaimek v prvi osebi *watashi* 私, kazalnice: *kono* この, *sono* その, *sore* それ
• Formalni samostalniki	*koto* こと, *mono* もの

Kot lahko vidimo pri enotah, podanih v nadaljevanju, je izpostavljenost spodnjih enot večja v časopisih kot v korpusu JpWaC.

• Pretekla oblika na *-ta*	Časopisni članki so povečini napisani v pretekli obliki.
• Pripone in zapone za: čas, kraj, števila, enote za količino, ljudi	*nichi* »dan« 日, *shi* »mesto« 市, *ken* »prefektura« 県, *yaku* »okrog, približno« 約, *en* »jen« 円, *meetoru* »meter« メートル, *hito* »človek« 人, *san* »gospod/gospa« さん
• Samostalniški prislovi	*sakunen* »prejšnje leto« 昨年, *go* »po« 後, *moto* »prvotno« 元
• Lastna imena	»Tokio« 東京, »Osaka« 大阪, »Suzuki« 鈴木
• Splošni samostalniki	*yougi* »osumljenec« 容疑, *shushou* »predsednik vlade« 首相, *Beikoku* »ZDA« 米国, *tero* »terorizem« テロ, *Jimintou* »Liberalna demokratska stranka« 自民党

Izpostavljenost predstavljenih enot je višja zaradi narave vsebin na spletu in v časopisih. Zanimivo pri tem je, da se splošni samostalniki z lastnostmi spletnih podatkov na spletu niso pojavili med stotimi najpogostejšimi enotami, kar lahko kaže na večjo uravnoteženost spletnega korpusa. Rezultat primerjave spletnih in časopisnih podatkov kaže, da imajo časopisni podatki z vidika oblike (rabe preteklega časa, oblike na *masu* in *desu* se ne uporabljajo) kot tudi z vidika vsebine (visoka frekvenca

splošnih samostalnikov) posebne lastnosti. Po drugi strani je korpus JpWaC manj formalen, ima izčrpnejše podatke in je vsebinsko bolj raznolik ter obsežen.

Že Biber (1988, 1995) in Heylighen (2002) sta raziskovala razlike med formalnim in specifičnim jezikom na eni strani ter neformalnim in splošnejšim jezikom na drugi. Razlike sta razumela kot pomemben del jezikovnih zvrsti. Omenjeni rezultati raziskave so podobni rezultatom primerjave med korpusom časopisov in spletnim korpusom v angleškem in nemškem jeziku, ki jih je opisal Sharoff (2006a). Poleg tega je bil v Sharoffovi primerjavi (2006a) uporabljen uravnoteženi korpus BNC, kjer se je potrdilo, da so podatki iz korpusa BNC bolj podobni spletnemu korpusu kot pa časopisnim člankom. Podobno ugotovitev potrdimo tudi v 4. poglavju v tej monografiji na primeru prislovov.

2.3.5 Spletni podatki kot vrsta korpusa

V zadnjem desetletju so se v mnogih jezikih, začenši z angleščino, spletni podatki začeli uporabljati kot obsežen jezikovni vir s pomočjo različnih metod in tehnik. Tudi raziskave so pokazale, da je splet lahko dober vir uporabnih jezikoslovnih podatkov.

Po drugi strani je veliko kritik glede spletnih podatkov kot predmeta jezikoslovne analize. Kritike se v največji meri nanašajo na razsežnost nepotrebnih podatkov oz. šuma v korpusu (ang. *noise*). Vendar pa raziskave o spletnih podatkih kažejo, da z naraščanjem količine podatkov nepotrebni podatki postanejo neznatni in ne vplivajo na kakovost rezultatov. Poleg tega je bilo ugotovljeno, da se rezultati, pridobljeni iz tovrstnih korpusov, skladajo s človeško subjektivno presojo (Keller in Lapata, 2003). Naslednja pogosta kritika je ta, da so spletni podatki specifični. V obstoječem spletnem korpusu, ki zajema tudi korpus JpWaC, vsebina podatkov ni kategorizirana glede na zvrst. Z raziskavami o spletnih korpusih, ki so postale popularne v zadnjem času in s katerimi se preučuje tudi metodologija klasificiranja spletnih podatkov (Sharoff, 2006; Ueyama in Baroni, 2005), so postale dostopne analize s kategoriziranjem in urejanjem podatkov glede na cilj in namen. Prav tako se s primerjanjem jezikov v spletnih korpusih, korpusih časopisov in uravnoteženih korpusih s spletnim korpusom bolj kot s korpusom časopisov približamo rezultatom uravnoteženega korpusa (Sharoff, 2006a,b; Ueyama in Baroni, 2005). Spletni korpusi odražajo aspekt običajnega jezika, poleg tega so številni in širokega obsega, zato dajejo boljše rezultate.

V zvezi s primerjavo angleškega uravnoteženega korpusa in spletnega korpusa so zanimive tudi naslednje ugotovitve. V uravnoteženem korpusu je veliko zaimkov v

tretji osebi, preteklika in pripovedovalnega stila (ang. *narrative style*), medtem ko je v spletnem korpusu veliko zaimkov v prvi in drugi osebi, sedanjika oz. prihodnjika in interaktivnega stila (ang. *interactive style*). Na splošno naj bi bil uravnoteženi korpus zanesljivejši, vendar pa ni konsenza glede tega, katera kategorija in v kolikšni meri naj bi bila vključena za primerno uravnoteženost. Če za določen jezik obstaja uravnoteženi korpus, potem se s primerjanjem spletnega in uravnoteženega korpusa lahko pokažejo značilnosti posameznega korpusa. Spletni korpus je lahko v pomoč pri izdelavi uravnoteženega korpusa in – gledano s tega vidika – je spletni korpus koristno vodilo. Po drugi strani pa lahko – v primeru, da zaradi različnih razlogov določen jezik nima uravnoteženega korpusa – spletni korpus zelo učinkovito nadomesti uravnoteženi korpus (Ghani et al., 2001).

V določenih primerih, odvisno od cilja raziskave, ne zadostujejo samo frekvenčni podatki, tako da je nujen pregled vsakega posameznega primera, za kar je potrebna tudi človeška presoja. Z drugimi korpusi je podobno, čeprav so v spletnih korpusih jezikovni standardi bolj ohlapni, več je variant kot v drugih korpusih, kar kaže na jezikovno pestrost. Z vidika standardnega jezika je možna tudi kritika, da uporaba spletnih korpusov ni ustrezna, a v praksi med spletnimi podatki pogosto vidimo podatke, ki so v skladu z jezikovnim standardom. Ocenjuje se, da so iz visokofrekventnih podatkov pridobljeni zelo kakovostni rezultati, tudi z vidika standardnega jezika.

Poleg zelo pogostih jezikovnih elementov pridobivamo tudi srednje- in nizkofrekventne podatke, s pomočjo človeške presoje, če je to potrebno. Poleg tega spletni podatki prikazujejo praktično rabo jezika in njegovo produkcijo, kar je ravno tako zelo koristna informacija. Tudi jezikoslovci poudarjajo, da se vloge spleta kot novega medija v jezikoslovju ne sme zanikati (Crystal, 2006).

Glede na povedano lahko glavne vloge spletnih podatkov strnemo v naslednji dve točki:

- spletni podatki so jezikoslovni vir za raziskovanje naravnega jezika;
- spletni podatki so vir za raziskovanje novega medija in njegovih lastnosti.

Priučujoči razdelek je podal raziskave spletnih podatkov za različne jezike, začenši z angleščino. Tudi v japonščini je nujna večstranska primerjava korpusov (npr. z naslednjimi korpusi: Aozora Bunko, korpus uradne bele knjige, korpus govornega jezika, korpus učbenikov, celotni uravnoteženi korpus BCCWJ ipd.), kar v 4. poglavju obravnavamo na primeru prislovov.

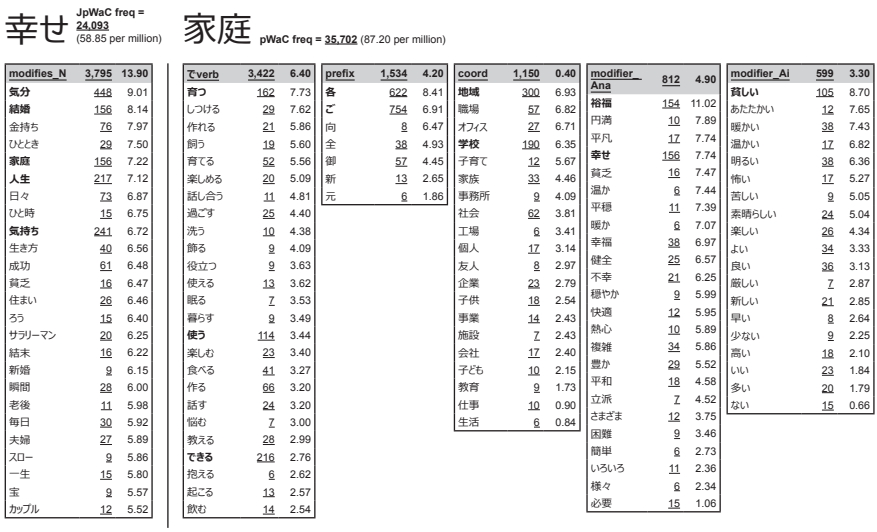
2.4 Izdelava slovnične datoteke v japonščini

V podpoglavju 2.3 smo predstavili pripravo spletnega korpusa in vnos v sistem SkE kot prvi korak pri izdelavi japonske različice besednih skic. Naslednji korak je določitev kolokacijskih in slovničnih odnosov v slovnični datoteki ter vnos te datoteke v orodje SkE. S tem poleg običajnih konkordanc omogočimo tudi prikaz kolokacijskih in slovničnih povezav za ključne besede, prikaz podobnosti in razlik med sopomenkami, med podobnimi ali različnimi besedami, tezaver in druge jezikovne podatke.

2.4.1 Vzorci v slovnični datoteki

Japonski slovnični vzorci (Srdanović et al., 2008; Srdanović in Nishina, 2008) imajo v slovnični datoteki 22 registriranih pravil in z njihovo pomočjo lahko iščemo potencialne zveze med ključno besedo in drugimi besedami v jeziku, obenem pa je s tem ponujen prikaz korpusnih podatkov na naprednejši način v okviru besednih skic, tezavra in primerjalnih skic. Teh 22 pravil zajema glagole, samostalnike, pridevnike na *-na*, pridevnike na *-i*, prislove in kolokacije z njimi. Posamezna besedna vrsta vključuje številne kolokacijske vzorce. Na primer, pri vnosu glagola kot ključne besede se v rezultatih iskanja pokažejo povezave tega glagola s sklonskimi členki *ga* (が), *wo* (を), *to* (と), *ni* (に), *de* (で), *made* (まで), *kara* (から), *he* (へ) in s tematskim členkom *wa* (は) in z njimi povezani samostalniki. Poleg tega se pri iskanju glagolov pokažejo tudi prislovi, odvisni glagoli, glagolske pripone in njihove kolokacije, drugi neodvisni glagoli in njihove paralelne povezave, ki se pojavljajo v kombinaciji z iskanim glagolom v japonskem jeziku. Slovnične povezave so sestavljene s pomočjo regularnih izrazov, besednih vrst orodja ChaSen in po predlogu korpusne poizvedbene sintakse, kot priporoča Gahl (1998).

Primer na levi strani Slike 2, tj. pridevnik na *-na shiawasena* »srečen, vesel« (幸せな) + samostalnik, prikazuje delni rezultat iskanja besednih skic. Vidimo lahko, kako iskalna beseda *shiawasena* modificira samostalnik (npr. *shiawasena kibun* »veselo/srečno razpoloženje« (幸せな気分), *shiawasena kekkon* »srečen zakon« (幸せな結婚), *shiawasena katei* »srečen dom, srečna družina« (幸せな家庭) ipd. Primer na desni strani Slike 2 prikazuje rezultate iskanja samostalniške besede *katei* »dom, družina« (家庭) in nekaj različnih vzorcev te besede, npr. kateri pridevniki jo modificirajo (*yuufukuna katei* »bogata/premožna družina« (裕福な家庭), *enmanna katei* »miren dom, složna družina« (円満な家庭) ipd.), s katerim glagolom in členkom *de* se pojavi (*katei de sodatsu* »odraščati doma/ v družini« (家庭で育つ) ipd.).



Slika 2: *Besedne skice: rezultati iskanja za pridevnik na* -na shiawasena »*srečen, vesel*« （幸せな） + *samostalnik (levo) in rezultati iskanja za* katei »*dom*«（家庭） *(desno).*

Leva in desna stran Slike 2 prikazujeta posamična stolpca s številkami; prvi stolpec prikazuje kolokacijsko frekvenco v korpusu, drugi pa izpostavljenost oz. pomembnost statistične kolokacije. S klikom na število v prvem stolpcu tabele se v konkordancah prikažejo stavčni primeri, ki vsebujejo posamezne kolokacije besed in ključne besede v korpusu. S klikom na povezavo slovničnega vzorca (npr. modifies_N) odpremo slovnično datoteko in tako ugotovimo, na kakšen način je bil sestavljen določen vzorec – pravilo slovnične povezave z uporabo regularnih izrazov, besednih vrst in drugih podatkov.

Slovnična povezava, prikazana na Sliki 2, je nastavljena kot dvostranska povezava (DUAL). Z iskanjem pridevnika na *-na* se prikaže samostalnik, ki ga pridevnik modificira, in če iščemo samostalnik, se prikaže pridevnik na *-na*, ki mu odgovarja. Tovrstni primer slovnične povezave je opisan v nadaljevanju.

Primer 2-1:

**DUAL*

=modifier_Ana/modifies_N

2:"N.Ana" "Aux" "Pref.*"? **1:**[tag="N.*" & tag!="N.Suff.*" & tag!="N.bnd.*"]

V zgornji formuli modifier_Ana označuje pridevnik na *-na*, modifies_N pa modificiran samostalnik. V formuli 2 ("N.Ana" "Aux" "Pref.*"?) pride pomožni glagol (Aux) za pridevnikom, obstaja pa verjetnost, da se za tem pojavi še predpona (Pref.*). Formula 1 ([tag="N.*" & tag! ="N.Suff.*" & tag! ="N.bnd.*"]) kaže metodo pridobivanja samostalnikov. Samostalniških oznak (N.*) je preveč, zato med njimi izbrišemo zapone (N.bnd.*) in odvisne samostalnike (N.Suff.*). S pomočjo tega pravila lahko pri iskanju kolokacij za pridevnik 幸せ-な pridobimo samostalnika *gokibun* (ご気分) in *kibun* (気分), ki pomenita enako, le da se uporabljata v različnih registrih (*gokibun* vsebuje predpono za spoštljivost *go* in je bolj spoštljiv kot *kibun*). Samostalnika se v korpusu pojavljata v vzorcu *shiawase-na-go-kibun* (幸せ-な-ご-気分), *shiawase-na-kibun* (幸せ-な-気分).[36]

2.4.2 Luščenje kolokacijskih podatkov iz korpusa s pomočjo sistema

V slovnični datoteki, ki jo je avtorica izdelala za japonsko različico sistema SkE, je torej 22 pravil in ta pravila zajemajo več kot 50 različnih kolokacijskih povezav (Srdanović et al., 2008; Srdanović in Nishina, 2008). Od besednih vrst lahko iščemo samostalnike, glagole, pridevnike na *-na*, pridevnike na *-i*, prislove in njihove kolokacijske povezave. Ker je morfološka analiza orodja ChaSen zelo natančna, slovnična datoteka vključuje pripone, predpone, odvisne glagole, glagolske pripone ipd. in njihove kolokacijske podatke. Tabela 7 kaže številne kolokacijske povezave, ki jih pokriva slovnična datoteka.

Samostalniške kolokacijske povezave so najpogostejše, zajetih je 16 vrst. Glagolskih kolokacijskih povezav je 14, kolokacijskih povezav pridevnikov na *-i* je 7 in pridevnikov na *-na* je 11. Zajeta je le ena vrsta povezave s prislovi, a v 5. in 6. poglavju predlagamo izboljšave. Še posebno se osredotočimo na modalno obliko in njeno kolokacijsko povezavo, ki je ni mogoče analizirati z orodjem ChaSen (gl. poglavji 5 in 6).

36 Nadomestna znaka, ki se uporabljata v regularnih izrazih nosijo naslednji pomen:

- * Število izraza je 0 ali več (uporablja se npr. za oznake besednih vrst, ki vsebujejo več podoznak. Npr. N.* vsebuje oznake N.g, N.Prop ipd.).
- ? Število izrazov je enako 0 ali 1.
- ! Izpustitev določenega izraza.
- & Povezovanje izrazov pred in po.

Tabela 7: *Kolokacijske povezave, ki jih omogoča japonska slovnična datoteka.*

Besedna vrsta		Slovnični odnos (gramrel)	Tip odnosa	Primer
Samostalniki	16	modifier_Ai	pridevnik na *-i* opisuje samostalnik	*atarashii chousen* »nov izziv« 新しい挑戦
		modifier_Ana	pridevnik na *-na* opisuje samostalnik	*kakanna chousen* »drzen izziv« 果敢な挑戦
		をverb	samostalnik + *wo* + glagol	*chousen wo ukeru* »sprejeti izziv« 挑戦を受ける
		でverb	samostalnik + *de* + glagol	*oyu de toku* »stopiti v vroči vodi« お湯で溶く
		がverb	samostalnik + *ga* + glagol	*chousen ga hajimaru* »izziv se prične/začne« 挑戦が始まる
		にverb	samostalnik + *ni* + glagol	*chousen ni tachimukau* »soočiti se z izzivom« 挑戦に立ち向かう
		はverb	samostalnik + *wa* + glagol	*chousen wa tsuduku* »izziv se nadaljuje« 挑戦は続く
		からverb	samostalnik + *kara* +glagol	*oyu kara agaru* »iti iz vroče vode« お湯から上がる
		pronomの	samostalnik 2 + *no* + samostalnik 1	*saigo no chousen* »zadnji izziv« 最後の挑戦
		のpronom	samostalnik 1 + *no* + samostalnik 2	*chousen no iyoku* »volja da odgovori na izziv« 挑戦の意欲
		がAdj	samostalnik + *ga* +Adj	*oyu ga ii* »vroča/topla voda je dobra« お湯がいい
		はAdj	samostalnik + *wa* +Adj	*oyu wa nurui* »voda je mlačna« お湯はぬるい
		Coord	paralelni odnosi	*chousen・kakushin* »izzivi in inovacije« 挑戦・革新
		particle	samostalnik + členek	*chousen to iu* »t. i. izziv« 挑戦という

Besedna vrsta		Slovnični odnos (gramrel)	Tip odnosa	Primer
Samostalniki	16	Suffix	samostalnik + pripona	*chousenjou* »pisni izziv« 挑戦状[37]
		Prefix	predpona + samostalnik	*hatsu chousen* »prvi izziv« 初挑戦
Glagoli	14	modifier_Adv	prislov opisuje glagol	*nikoniko warau* »nasmehniti se« にこにこ笑う
		nounは	samostalnik + *wa* +glagol	*kare wa warau* »on se smeji«彼は笑う
		nounが	samostalnik + *ga* + glagol	*oni ga warau* »hudič se smeji/nerealistično« 鬼が笑う
		bound_V	odvisni glagoli, vezani za neodvisne glagole	*waratchau* »(na)smejati se« わらっちゃう
		V_bound	neodvisni glagoli, vezani za odvisne glagole	*tsurete iku* »pripeljati koga s seboj« 連れて行く
		nounで	samostalnik + *de* + glagol	*hana de warau* »posmehovati se nekomu« 鼻で笑う
		nounに	samostalnik + *ni* + glagol	*takaraka ni warau* »naglas se smejati« 高らかに笑う
		nounから	samostalnik + *kara* + glagol	*(kokoro no) soko kara warau* »smejati se od srca« （心の）底から笑う
		nounまで	samostalnik + *made* + glagol	*saigo made warau* »smejati se do konca« 最後まで笑う
		nounを	samostalnik + *wo* + glagol	*hara wo (kakaete) warau* »počiti od smeha« 腹を（抱え て）笑う
		nounへ	samostalnik + *he* + glagol	*(kouen he iku)* »iti v park«（公園へ行く）
		Coord	paralelni odnos	*warau・naku* »smejati se in jokati« 笑う・泣く

37 »Takeshi no Chousenjou« »Pisni izziv Takeshija« (たけしの挑戦状) je znana japonska igrica.

Besedna vrsta		Slovnični odnos (gramrel)	Tip odnosa	Primer
Glagoli	14	Suffix	glagol + pripona	*waraippanashi* »še naprej se smejati« 笑いっぱなし
		Prefix	prefix + glagol	*chouwarau* »ful se smejati« 超笑う
Pridevniki na *-i*	7	modifies_N	pridevnik na *-i* opisuje samostalnik	*nagai rekishi* »dolga zgodovina« 長い歴史
		Nは	samostalnik + *wa* + pridevnik na *-i*	*michinori wa nagai* »pot je dolga« 道のりは長い
		Nが	samostalnik + *ga* + pridevnik na *-i*	*maeoki ga nagai* »uvod je dolg« 前置きが長い
		bound_N	odvisni/neodvisni samostalniki vezani na pridevnike na *-i*	*nagai wake* »razlog zakaj je dolg/ni dolg …« 長いわけ
		Coord	paralelni odnos	*nagai・mijikai* »dolg in kratek« 長い・短い
		Suffix	pridevnik na *-i* +pripona	*chounagai* »ful dolg« 超長い
		Prefix	prefix + pridevnik na *-i*	*nagasa* »dolžina« 長さ
Pridevniki na *-na*	11	bound_N	bound/free samostalnik nouns connecting to N.Ana samostalnik	*juuyouna ten* »pomembna točka« 重要な点
		Nは	samostalnik + *wa* + N.Ana samostalnik	*yakuwari wa juuyou* »vloga je pomembna« 役割は重要
		Nが	samostalnik + *ga* + N.Ana samostalnik	*koto ga juuyou* »pomembno je, da« ことが重要
		pronomの	samostalnik + *no* + N.Ana samostalnik	*nettowaaku no juuyousei* »pomembnost mreže« ネットワークの重要(性)
		のpronom	samostalnik N.Ana + *no* + samostalnik	*juuyou no kadai* »pomembna tema« 重要の課題
		modifier_Ai	pridevnik na *-i* opisuje N.Ana samostalnik	*monosugoi juuyou* »izjemno pomemben« ものすごい重要

Besedna vrsta		Slovnični odnos (gramrel)	Tip odnosa	Primer
Pridevniki na *-na*	11	modifier_Ana	N.Ana samostalnik opisuje N.Ana samostalnik	*fukaketsuni juuyou (na)* »neizogibna pomembnost« 不可欠に重要（な）
		Suffix	N.Ana samostalnik + pripona	*juuyousei* »pomembnost« 重要性
		Prefix	Predpona + N.Ana samostalnik	*saijuuyou* »najpomembnejše« 最重要
		Coord	paralelni odnosi	*juuyou・takai* »pomembno in visoko« 重要・高い
		particle	N.Ana samostalnik + členek	*juuyou to (naru・iu)* »postati pomemben / praviti, da je nakaj pomembno« 重要と（なる・いう）
Prislovi	1	modifies_V	prislov opisuje glagol	*yatto ochitsuku* »končno se pomiriti« やっと落ち着く

Ta komplet kolokacijskih povezav je bil pridobljen iz korpusa JpWaC in temelji na besednih vrstah orodja ChaSen. V Tabeli 8 je podana količina pridobljenih podatkov, tj. samostalnikov, glagolov, pridevnikov na *-i*, prislovov (za statistiko o podatkih drugih besednih vrst v korpusu gl. podpoglavje 2.3.2.).

Tabela 8: *Število samostalnikov, glagolov, pridevnikov in prislovov v korpusu JpWaC.*

PoS	ChaSen PoS	PoS (ANG)	PoS (SLO)	Pojavnice	Različnice
Samostalniki	名詞-一般	N.g	Samostalnik (splošni)	50.121.542	**49.268**
	名詞-固有名詞-地域-一般	N.Prop.p.g	Samostalnik (zemljepisna lastna imena)	1.878.093	**19.280**
	名詞-固有名詞-人名-名	N.Prop.n.f	Samostalnik (osebna lastna imena – ime)	1.570.530	**12.854**
	名詞-サ変接続	N.Vs	Samostalnik (gl. *suru*)	21.704.856	**10.017**
	名詞-固有名詞-人名-姓	N.Prop.n.s	Samostalnik (lastna imena – priimek)	1.499.072	**9.116**
	名詞-固有名詞-一般	N.Prop.g	Samostalnik – lastna imena	622.453	**7.517**
	名詞-固有名詞-組織	N.Prop.o	Samostalnik (lastna imena – organizacije)	940.397	**7.277**
	名詞-形容動詞語幹	N.Ana	Osnova pridevnika na *-na*	4.436.970	**2.892**
	名詞-固有名詞-人名-一般	N.Prop.n.g	Samostalnik (osebna lastna imena – splošna)	98.476	**872**
	名詞-副詞可能	N.Adv	Samostalnik – prislov	4.831.851	**734**
	その他	other	Drugo	52.160.162	**1.905**
	合計	**total**	**Skupaj**	139.864.402	**121.732**
Glagoli	動詞-自立	V.free	Glagol (neodvisni)	32.746.478	**17.135**
	動詞-非自立	V.bnd	Glagol (odvisni)	6.884.053	**470**
	動詞-接尾	V.Suff	Glagol (pripona)	2.714.110	**36**
	合計	**total**	**Skupaj**	42.344.641	**17.641**
Pridevniki	形容詞-自立	Ai.free	Pridevnik na *-i*	3.871.216	**1.799**
	形容詞-非自立	Ai.bnd	Pridevnik (odvisni)	248.190	**52**
	形容詞-接尾	Ai.Suff	Pridevnik (pripona)	21.871	**12**
	合計	**total**	**Skupaj**	4.141.277	**1.863**
Prislovi	副詞-一般	Adv.g	Prislov (splošni)	3.603.890	**2.303**
	副詞-助詞類接続	Adv.P	Prislov (vezni)	2.537.640	**517**
	合計	**total**	**Skupaj**	6.141.530	**2.820**

2.4.3 Slovnična datoteka in razmerje do orodja ChaSen

Sistem SkE uporablja rezultate morfološke analize orodja ChaSen (elektronski slovar, uporabljen za ChaSen, je IPADIC), zato je nujno iskanje glede na sistem besednih vrst v slovarju IPADIC.[38] Na primer – ker je pridevnik na *-na shiawasena* razdeljen na *shiawase* + *na*, ga je treba pri iskanju ene enote iskati kot *shiawase* (幸せ).

Treba je omeniti dobre in slabe plati slovarja IPADIC v sistemu ChaSen. Enote analize so zelo natančne, zato so koristne za definiranje slovničnih povezav. Za definiranje vzorca je bolj priporočljiva uporaba natančnih oznak, saj s tem znižamo potrebo po omejitvah. Po drugi strani je v primeru analize dveh enot ali več z orodjem ChaSen možno iskanje z zapletenimi vzorci, kar pa z besednimi skicami postane težavno, saj se kot ključna beseda lahko uporabi samo ena enota. Na primer, beseda *onna no ko* »dekle« (女の子) v sistemu ChaSen šteje za eno enoto, zato je tudi iskanje besednih skic enostavno. Po drugi strani pa je beseda *onna no hito* »ženska« (女の人) razdeljena na tri morfeme, zato je iskanje v Besednih skicah onemogočeno in je namesto tega potrebno jo iskati kot vzorec v konkordancah.

Natančnost morfološke analize je težko 100-odstotna, zato so tudi pri besednih skicah občasno vidne napake. Tovrstni primer je npr. stavek *Onsen he itta* »toplice + v/k (cilj gibanja) + iti (v pretekliku) = Šel sem v toplice« (温泉へ行った（いった）, ki je napačno analiziran kot **Onsen he okonatta* »toplice + v/k (cilj gibanja) + *izvesti (v pretekliku) = *Bil je izveden v toplice« (温泉へ行った（*おこなった）. Poleg tega so narečne in pogovorne besede v tovrstnih jezikovnih podatkih analizirane kot neznane besede. Izboljšava sistema za analizo ostaja izziv za prihodnost.

2.5 Glavne funkcije sistema

To poglavje predstavi glavne funkcije sistema SkE, to so besedne skice, tezaver in primerjalne skice.

2.5.1 Besedne skice: funkcija pridobivanja kolokacijskih povezav

Funkcija Besedne skice se uporablja za iskanje kolokacijskih povezav in kolokacijskih gesel za določeno ključno besedo v korpusu. Iskanje poteka tako, da določeno ključno besedo poiščemo s funkcijo Besedne skice, nato pa pridobimo izpis

38 Ogled sistema besednih vrst slovarja IPADIC na podlagi analizatorja ChaSen je možen na spletni strani orodja Chakoshi: http://tell.fll.purdue.edu/chakoshipub/index2.html (dostop 14. 9. 2014).

različnih kolokacijskih odnosov in številne kolokacije, ki se pojavljajo v korpusu, prikazane in razporejene glede na statistično izpostavljenost in frekvenco.[39] Vsak primer kolokacije lahko poiščemo s konkordancami. Na slikah 3, 4 in 5 v tem poglavju pokažemo rezultate iskanja samostalnikov, glagolov, pridevnikov na *-i*, pridevnikov na *-na* ter kolokacijske povezave posameznih delov besednih vrst.

Slika 3 prikazuje rezultate iskanja za samostalnik *oyu* »vroča/topla voda« (お湯). Skrajno desni stolpec (modifier_Ana) in četrti stolpec spodaj (modifier_Ai) prikazujeta pridevnike, ki modificirajo samostalnik *oyu*. Podobno tudi prvi (をverb), tretji (にverb) in četrti stolpci (でverb) prikazujejo glagole, ki se sopojavljajo s samostalnikom *oyu*. Poleg tega se v drugem stolpcu (pronomの) vidijo vezave s členkom *no* (の) in samostalniki. V enem tipu kolokacijske povezave lahko opazimo različne pomene ključne besede glede na različne kolokacije, pomenske razlike pa se pojavljajo tudi med različnimi slovničnimi odnosi. V kolokacijskih povezavah so razvidne kolokacije z najvišjo izpostavljenostjo (številke v drugem stolpcu vsakega okvirja) oz. z visoko frekvenco (številke v prveg stolpcu vsakega okvirja). Iz števila v prvem stolpcu je s pomočjo konkordanc mogoče poiskati primere stavkov posameznih kolokacij.

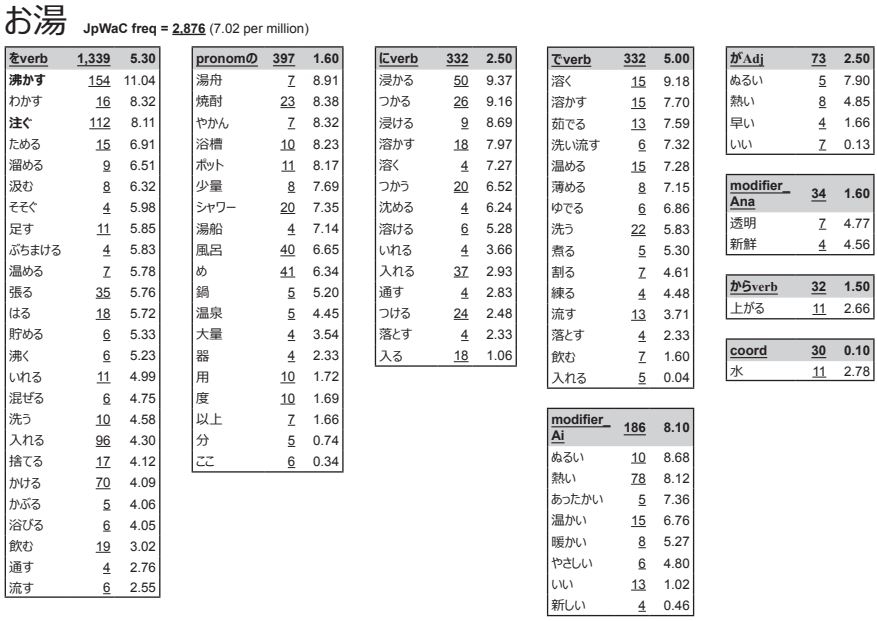
S seznama kolokacij za samostalnik *oyu* so razvidni naslednji posamezni pomeni besede *oyu*.

1. Vroča/topla voda (pogreta voda): *oyu wo wakasu* (お湯を沸かす) »zavreti vodo«, *atsui oyu* (熱いお湯) »vroča/topla voda«, *potto no oyu* (ポットのお湯) »vroča/topla voda iz termovke/v termovki«, *nabe no oyu* (鍋のお湯) »vroča/topla voda iz lonca/v loncu, *oyu de yuderu* (お湯で茹でる) »skuhati v vroči vodi«, *oyu de atatameru* (お湯で温める) »pogreti z vročo vodo«,
2. Kopanje (voda za kopanje): *oyu ni hairu* (お湯に入る) »stopiti v vročo vodo«, *oyu kara agaru* (お湯から上がる) »stopiti iz vroče vode«, *yubune no oyu* (湯舟のお湯) »vroča/topla voda v kopalni kadi«, *yokusou no oyu* (浴槽のお湯) »vroča/topla voda v kopalni kadi«, *furo no oyu* (風呂のお湯) »vroča/topla voda v (tradicionalni) kadi«, *yasashii oyu* (やさしいお湯) »prijetna vroča/topla voda«, *ii oyu* (いいお湯) »dobra, prijetna vroča/topla voda«.[40]

Razmerje *coord* kaže, da se beseda *oyu* (お湯) v paralelnem razmerju sopojavlja z (mrzlo/hladno) vodo *mizu* (水).

39 Kot že omenjeno v prejšnjem razdelku, japonska različica pokriva različne besedne vrste (glagole, samostalnike, pridevnike na *-i* in *-na* ter prislove) in vsebuje več kot 50 različnih tipov kolokacijskih povezav (na osnovu 22 vzorčnih pravil v slovnični datoteki).

40 Pri iskanju besede *yu* (brez morfema *-o-*) lahko najdemo še dodatni pomen »toplice«.



Slika 3: *Besedne skice: rezultati kolokacij za samostalnik* oyu *»vroča/topla voda«*

Slika 4 prikazuje rezultate iskanja kolokacijskih povezav glagola *warau* »smejati se« (笑う). Tretji stolpec (modifier_Adv) kaže prislove, ki modificirajo glagol. Na primer, *nikkori warau* »nasmehniti se«, *omowazu warau* »nehote se zasmejati«, *kusukusu warau* »hihitati se«, *nikoniko warau* »nasmejati se« ipd. V stolpcu »noun は« so prikazani osebki oz. teme. Stolpec »bound_V« prikazuje neodvisne in odvisne glagole ter glagolske pripone, stolpec na skrajnji levi »coord« pa prikazuje paralelne povezave. Od četrtega do enajstega stolpca (»nounを«, »nounに«, »nounで«, »nounが«, »nounから«, »nounまで«, »nounへ« so prikazani samostalniki, ki se najlažje povežejo z glagolom *warau*, in sicer v kombinaciji z določenim členkom. Na koncu so prikazane še pripone in predpone ter njihove kolokacije.

Slika 5 prikazuje rezultate iskanja kolokacijskih povezav pridevnika *surudoi* »oster« (鋭い). Stolpec na skrajni levi prikazuje katere samostalnike modificira ta pridevnik. Na primer, *dousatsu* »vpogled« (洞察), *gankou* »pogled, uvid, vid« (眼光), *shiteki* »izpostavljanje, pokazanje« (指摘). Stolpca »Nは« in »Nが« kažeta, s katerimi temami oz. subjekti se pridevnik *surudoi* pogosto sopojavlja. Stolpec »bound_N« kaže povezave s samostalniki ter samostalnikškimi priponami, vidijo pa se tudi paralelne kolokacijske povezave (coord), pripone in zapone.

笑う JpWaC freq = 20,910 (51.07 per million)

coord	9,250	3.10
泣く	270	7.98
ごまかす	58	7.29
怒る	157	7.28
笑える	48	6.81
答える	127	6.29
済ませる	35	6.05
すませる	19	5.81
誤魔化す	16	5.64
話せる	22	5.63
はいる	42	5.63
済ます	18	5.60
興じる	15	5.58
泣ける	17	5.58
こみ上げる	15	5.56
暮らせる	16	5.54
うなずく	16	5.49
合える	14	5.36
貼る	37	5.34
許せる	16	5.30
さざめく	11	5.27
許す	66	5.26
話す	105	5.23
絶える	14	5.19
楽しめる	26	5.08
振る	30	5.04

bound_V	7,329	4.30
ちゃう	486	7.30
しまう	1,491	6.37
みせる	33	5.98
ちまう	23	5.98
だす	67	5.90
合う	89	5.50
出す	307	5.35
てる	769	5.34
あう	33	5.24
すぎる	83	4.99
くれる	310	4.92
いける	93	4.41
もらえる	22	4.10
とく	12	4.03
くださる	106	3.98
下さる	43	3.88
いる	2,579	3.79
なさる	21	3.73
続ける	67	3.72
つづける	8	3.59
始める	37	3.03
はじめる	8	3.00
過ぎる	15	2.91
く	22	2.84
あげる	19	2.77

modifier_Adv	3,140	12.10
にっこり	216	10.95
思わず	252	10.74
くすくす	105	10.05
ゲラゲラ	103	10.02
にやりと	100	9.96
ニヤリ	97	9.93
ニコニコ	103	9.79
ニッコリ	80	9.62
ニヤニヤ	79	9.48
ヘラヘラ	59	9.22
にこにこ	55	9.11
ニコッ	52	9.05
げらげら	51	9.03
にやにや	52	9.01
へらへら	42	8.73
心から	48	8.26
大いに	53	8.03
どっと	30	7.99
つい	29	7.67
思いっきり	24	7.55
ひとしきり	18	7.46
やっと	34	7.22
すっと	16	7.22
いつも	60	7.20
にんまり	14	7.16

nounを	2,868	1.30
腹	79	7.63
能収	12	7.09
歯	39	6.80
鼻くそ	8	6.47
冗談	17	6.42
鼻糞	6	6.09
お腹	17	5.93
顔	141	5.88
声	157	5.82
セックス	9	5.46
舌	9	5.41
ジョーク	6	5.38
大声	8	5.27
肩	16	5.11
頬	6	4.97
不幸	9	4.73
口	34	4.63
涙	12	4.52
失敗	9	4.23
怒り	7	4.16
眼	7	3.99
表情	8	3.88
こちら	18	3.70
文	6	3.32
様子	10	3.23

nounは	2,746	5.10
ディック	7	6.07
隣人	9	6.03
彼女	98	5.64
彼	144	5.35
ふたり	9	5.30
オレ	12	5.22
わ	8	4.98
夫人	6	4.88
筆者	9	4.87
少女	11	4.71
少年	15	4.66
赤ちゃん	7	4.63
観客	8	4.60
あたし	6	4.52
老人	7	4.44
今度	13	4.42
わたし	24	4.37
僕	46	4.24
男	29	4.16
女	19	4.11
先輩	7	4.07
ぼく	10	4.07
父	11	4.06
俺	14	3.75
母	9	3.53

nounに	2,117	1.80
傍ら	81	9.59
高らか	8	6.80
水族館	8	6.51
たび	22	6.27
いっしょ	19	6.23
げ	32	6.01
一緒	86	5.84
久々	9	5.30
やりとり	10	5.30
片手	6	5.29
ホント	6	5.12
冗談	6	5.09
鬼	6	5.07
そう	239	5.06
絶対	17	4.82
皆	14	4.81
みんな	25	4.69
気味	8	4.66
お互い	9	4.52
久しぶり	10	4.52
最後	42	4.11
誰か	8	4.08
ネタ	10	4.01
瞬間	6	3.95
逆	9	3.73

nounで	1,803	3.10
鼻	162	9.08
大声	91	9.02
口元	7	6.41
模試	6	6.36
ジョーク	8	6.16
皆	27	5.79
みんな	40	5.38
陰	7	5.29
隣	15	4.96
ひとり	16	4.94
笑顔	13	4.84
顔	64	4.76
声	62	4.49
向こう	9	4.28
横	11	4.27
端	6	4.20
本気	6	4.17
表情	9	4.14
そば	6	3.94
場面	6	3.38
シーン	6	3.37
会話	6	3.31
ネタ	6	3.30
感じ	13	3.05
ここ	37	2.95

nounが	1,476	1.60
鬼	35	7.78
口元	7	6.59
膝	11	5.82
みんな	43	5.50
皆	17	5.15
ニューヨーク	9	5.15
赤ちゃん	7	4.84
観客	7	4.59
全員	10	4.55
表情	7	3.81
顔	32	3.77
母	10	3.75
犬	7	3.48
妻	7	3.48
女の子	6	3.45
生徒	8	3.02
彼女	15	2.96
女	8	2.91
男	10	2.66
友人	6	2.54
君	15	2.36
彼	16	2.20
僕	11	2.20
ちゃん	9	2.18
相手	8	2.16

suffix	1,193	0.50
っぱなし	26	6.92
どころ	6	4.59
れる	998	3.51
そう	29	2.02
方	46	1.59
せる	43	1.04

nounから	268	1.40
底	41	6.78
朝	7	3.04
みんな	6	2.71

nounまで	96	1.40
ぶん	6	6.84
いつ	11	3.18
最後	8	1.75
今	8	0.28

nounへ	58	1.10
へ	18	8.14
トップ	15	4.56

Slika 4: *Besedne skice: rezultati kolokacij za glagol* warau »*smejati se*«.

鋭い JpWaC freq = 4,859 (11.86 per million)

modifies_N	2,795	52.80
洞察	49	8.44
眼光	28	8.28
牙	27	7.88
刃	29	7.67
観察	49	7.46
突っ込み	19	7.43
視線	49	7.26
対立	47	7.22
爪	25	7.07
指摘	165	7.02
感性	31	6.83
メス	17	6.79
分析	77	6.64
打球	10	6.57
批判	75	6.57
批評	20	6.56
眼差し	12	6.54
切れ味	9	6.50
刃物	11	6.48
直感	15	6.47
嗅覚	9	6.45
追及	11	6.42
感受性	11	6.37
切り口	13	6.37
痛み	36	6.35

Nは	501	20.50
舌鋒	6	8.50
眼光	6	8.24
勘	9	7.25
洞察	6	6.39
指摘	32	4.74
眼	8	4.44
感性	4	4.31
歯	4	3.84
感覚	12	3.77
分析	4	2.51
批判	4	2.47
視点	5	2.38
彼	12	1.79
音	4	1.18
言葉	7	0.83
目	9	0.22
これ	16	0.20
氏	5	0.17

Nが	489	13.80
勘	22	8.55
眼光	7	8.49
嗅覚	8	7.95
感受性	8	6.99
眠気	4	6.38
カン	6	6.32
刃	5	6.11
つき	18	6.05
振り	7	5.81
直感	5	5.66
感覚	23	4.71
視線	5	4.32
感性	4	4.31
眼	6	4.02
犬	4	2.74
音	8	2.18
目	16	1.05
力	14	1.00
感	5	0.58
声	4	0.56
先	4	0.43
意見	5	0.28
氏	5	0.17

suffix	321	1.00
さ	303	4.67

coord	262	0.60
鈍い	5	6.49
冷たい	6	4.72
深い	23	4.45
優しい	8	4.27
暖かい	4	4.25
激しい	8	3.99
黒い	4	3.89
速い	5	3.85
細かい	4	3.32
厳しい	8	3.08
長い	11	2.65
強い	12	2.25
楽しい	4	1.66
高い	12	1.52
多い	11	0.93
よい	6	0.83
新しい	5	0.78
良い	6	0.55
ない	10	0.08

bound_N	149	7.80
ん	9	1.15
もの	38	1.10
ところ	13	0.74
方	17	0.16

prefix	7	0.10
最	5	3.80

Slika 5: *Besedne skice: rezultati kolokacij za pridevnik* surudoi *»oster«.*

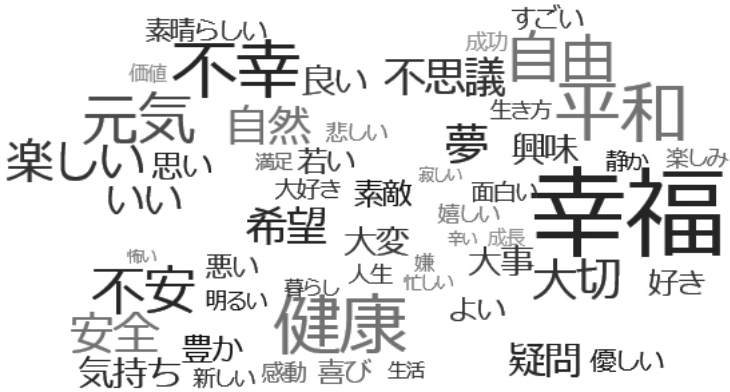
2.5.2 Tezaver: funkcija slovarja sopomenk

Tezaver temelji na »skupnih trojicah« (ang. *shared triples*), uporablja pa samodejno tehniko za luščenje besed, ki se podobno obnašajo glede na kolokacijske vzorce in gesla. Na primer, besedi *zasshi* »revija«（雑誌）in *hon* »knjiga«（本）se vsaka posamezno sopojavljata s povedkom *X wo yomu* »brati X«（？を読む）in tako oblikujeta trojno strukturo, ki kaže na podobno obnašanje elementov. S tovrstno tehniko je mogoče prikazati sopomenke in protipomenke. Osnovni koncept temelji na izdelavi samodejnega tezavra, ki so ga predlagali Sparck (1986), Grefenstette (1994) in Lin (1998).

Na primer, z iskanjem besede *shiawase* »sreča«（幸せ）v tezavru se pokažejo pomensko podobne besede, kot so *koufuku* »sreča«（幸福）, *tanoshii* »zabaven«（楽しい）, *yorokobu* »veseliti se«（喜ぶ）, in pomensko nasprotne besede, kot so *kiken* »nevarnost«（危険）, *fuan* »tesnoba«（不安）, *kodoku* »samota«（孤独）(gl. Sliko 6). Poleg tega tezaver slikovno prikaže podobne besede iskani besedi.

幸せ JpWaC freq = 24,093 (58.85 per million)

Lemma	Score	Freq	Cluster
幸福	0.370	8,023	**不幸** [0.247, 7,821]
健康	0.259	39,764	**平和** [0.258, 26,450] **自由** [0.218, 66,424] **安全** [0.183, 51,659] **自然** [0.182, 61,199]
元気	0.245	21,465	
不安	0.224	28,221	**夢** [0.186, 39,559] **希望** [0.178, 31,622] **疑問** [0.164, 24,016] **気持ち** [0.16, 54,620] **興味** [0.151, 32,593] **思い** [0.147, 30,111]
楽しい	0.223	50,233	**大切** [0.214, 41,607] **不思議** [0.203, 24,106] **いい** [0.194, 250,831] **良い** [0.175, 164,231] **大事** [0.175, 32,455] **大変** [0.174, 64,740] **若い** [0.165, 35,550] **よい** [0.165, 187,255] **好き** [0.164, 76,141] **悪い** [0.164, 77,238] **素晴らしい** [0.156, 23,158] **すごい** [0.155, 42,113] **大好き** [0.153, 16,964] **面白い** [0.148, 49,827] **新しい** [0.147, 89,405]
豊か	0.208	18,257	**素敵** [0.197, 12,734] **優しい** [0.17, 13,917] **明るい** [0.152, 15,179] **静か** [0.151, 11,972]
喜び	0.201	11,407	**感動** [0.171, 19,102] **楽しみ** [0.163, 19,017] **満足** [0.149, 22,705]
人生	0.185	40,625	**生き方** [0.162, 7,735] **生活** [0.156, 118,812] **暮らし** [0.153, 9,273]
成功	0.183	38,802	**価値** [0.164, 49,844] **成長** [0.16, 38,725]
嬉しい	0.180	25,126	**悲しい** [0.179, 11,710] **嫌** [0.173, 13,454] **忙しい** [0.161, 15,859] **辛い** [0.152, 11,085] **寂しい** [0.15, 7,978] **怖い** [0.147, 17,277]



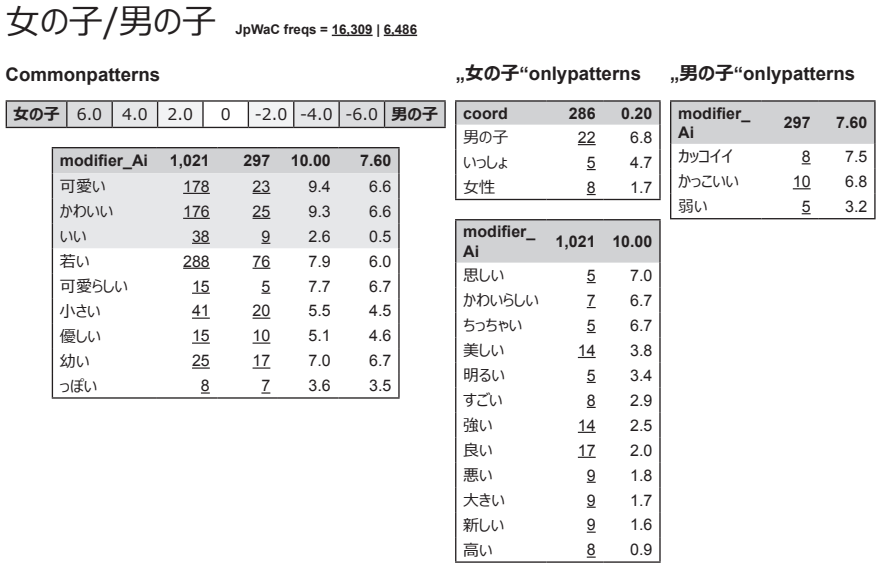
Slika 6: *Primer iskanja besede* shiawase *»sreča« v Tezavru.*

2.5.3 Primerjalne skice: funkcija prikaza podobnosti in razlik

Primerjalne skice prikažejo podobnosti in razlike v obnašanju besed. Tudi ta funkcija uporablja koncept skupnih trojic. Dve besedi se primerjata s pomočjo statističnega izračuna podobnega oz. drugačnega obnašanja teh besed glede na njihove kolokacijske odnose. V primeru podobnih besed, npr. sinonimov, se pojavi visoka

izpostavljenost teh besed v enakih skupnih trojicah, medtem ko ima nasprotna beseda zelo nizko izpostavljenost. Na ta način je mogoče raziskovati ne samo razlike v sopomenkah, ampak tudi podobnosti in razlike med besedami z dvema zapisoma ali med besedami z različnimi slovničnimi vlogami, kot so npr. prehodni in neprehodni glagoli.

Slika 7 prikazuje podobnosti in razlike, ki se pojavijo med kolokacijskimi besedami za besedi *onna no ko* »dekle« (女の子) in *otoko no ko* »fant« (男の子). Opazimo lahko kolokacijske trende, pri čemer se na besedo *onna no ko* vežejo pridevniki *kawaii* »srčkan«, *utsukushii* »lep«, na besedo *otoko no ko* pa pridevniki *hansamu* »čeden« in *kakkoii* »kul«. Zanimiva primera z visoko frekvenco sta kolokacijski zvezi *tsuyoi onna no ko* »močno dekle« (強い女の子) in *yowai otokonoko* »šibek fant« (弱い男の子). Zveza je namreč v nasprotju s splošnimi družbeno sprejetimi idejami in stereotipi, ki se kažejo kot *onna no ko wa yowai, otoko no ko wa tsuyoi* «dekle je šibko, fant je močen« (女の子は弱い・男の子は強い). Poleg tega je frekvenca besede dekle (16.309) 2,5-krat višja od frekvence besede fant (6.474). Tako lahko sklepamo, da je beseda dekle pogostejša zaradi tega, ker je primerna za širši razpon starosti subjekta in jo lahko zato uporabljamo v več kontekstih, možno pa je tudi, da so dekleta večkrat tema razprav na spletu kot fantje.



Slika 7: *Primerjalne skice: primer iskanja* onna no ko *»dekle« in* otoko no ko *»fant«.*

2.6 Povzetek in nadaljnja vprašanja

Rezultat pričujoče raziskave je izdelava japonske različice korpusnega iskalnega sistema SkE. Japonska različica je bila izdelana v dveh glavnih korakih: 1) oblikovanje spletnega korpusa JpWaC (400 milijonov besed), ki je morfološko analiziran z orodjem ChaSen, in 2) izdelava japonske slovnične datoteke, ki temelji na definiranju slovničnih pravil japonskega jezika s pomočjo regularnih izrazov in ChaSenovih besednih vrst. V sistemu SkE lahko uporabljamo tradicionalne iskalne metode, konkordance, s funkcijo Besedne skice pa lahko iz obsežnega korpusa v le nekaj sekundah pridobimo jezikoslovne podatke o kolokacijskih in slovničnih povezavah ključnih besed. Pred izdelavo japonske slovnične datoteke (Srdanović et al. 2008) je bil v raziskavah kolokacij v japonščini poudarek na sklonskih členkih, pripravljenih pa je bilo nekaj konkordančnikov za raziskovanje sklonskih členkov. S slovnično datoteko je omogočeno iskanje kolokacij za samostalnike, glagole, pridevnike na *-i*, pridevnike na *-na* in prislove, kar pomeni več kot 50 vrst kolokacijskih povezav. S tem predlagamo nov način za iskanje kolokacijskih povezav in gesel v japonščini. Kolokacijskih pravil je 22, z njimi pa vsaka besedna vrsta zajema številna kolokacijska gesla. Na primer, pri glagolih so prikazani poleg sklonskih členkov *ga* (が), *wo* (を), *to* (と), *ni* (に), *de* (で), *made* (まで), *kara* (から) in *he* (へ) in samostalnikov, ki se vežejo na glagole, tudi prislovi, odvisni glagoli, glagolske pripone in njihove kolokacije, drugi neodvisni glagoli in njihove paralelne povezave.

V pričujočem poglavju smo najprej razložili postopek izdelave japonske različice sistema SkE. Nato smo predstavili posamezne funkcije orodij Besedne skice, Primerjalne skice, Tezaver in Konkordančnik in podali nekaj primerov iskanja različnih japonskih besed v korpusu JpWaC s pomočjo izdelane japonske različice sistema. Na podlagi rezultatov primerjave časopisnih podatkov in spletnega korpusa, nameščenega v SkE, smo ocenili, da so spletni podatki manj specifični kot časopisni podatki in da v precejšnji meri pokažejo splošno rabo jezika. Japonsko različico tega sistema, podobno kot angleško, je mogoče aplicirati na različna področja jezikoslovnega izobraževanja, raziskovanja in leksikografije. Možnosti apliciranja in vrednotenja pridobivanja kolokacijskih povezav s sistemom so opisane v naslednjem poglavju.

Prednosti in pomankljivosti japonske različice orodja so opisane podrobno v Srdanović et al. (2011). Na tem mestu izpostavimo naslednje tri segmente, ki so potrebni izboljšave, in sicer bi bilo treba:

1. nadalje očistiti podatke spletnega korpusa in dodati raznovrstne podatke, vključno s kategorizacijo.

2. nadalje dodati in izboljšati slovnične odnose ter vnesti še druge izboljšave v sistem morfološke analize.
3. dodati dodatne informacije za naravne govorce in učence, kot so npr. funkcija prikazovanja oznak v japonščini v sistemu ChaSen, imena slovničnih izrazov v japonščini, *furiganu* (zapis branja pismenk) v rezultatih kolokacij in v primerih.

3 Aplikacija in vrednotenje japonske različice sistema za luščenje kolokacij[41]

3.1 Uvod

Za aplikacijo korpusa pri jezikovnem izobraževanju, leksikografiji in raziskovanju korpus ni dovolj, ampak je nujno tudi orodje za iskanje po korpusih. Orodje SkE, je eden od primerov korpusnih iskalnih sistemov trenutno najnaprednejše četrte generacije (McEnery in Hardie, 2012).

Po prihodu elektronskih korpusov so se pojavile naslednje značilnosti v razvoju leksikografije:

1. V 80. letih je bil Cobuildov angleški slovar prvi, pri katerem so bili uporabljeni elektronski korpus in konkordance.
2. V 90. letih sta Church in Hanks (1989) predlagala metodo »vzajemne informacije« (Mutual Information, MI) za statistično luščenje kolokacij iz korpusov, ta metoda pa se je začela uporabljati v slovaropisju.
3. Od leta 2000 so se pojavili naprednejši sistemi, ki omogočajo sistematičen pregled obnašanja besed (npr. besedne skice), in se začeli uporabljati v urejanju slovarjev.

Najprej je bilo orodje SkE uporabljeno za urejanje Macmillanovega angleškega slovarja (Rundell, 2002) s pridobljenimi podatki iz angleškega nacionalnega korpusa BNC (British National Corpus). Kronologijo projekta sta podrobno razložila Kilgarriff in Rundell (2002). Ugotovila sta, da so za metodologijo korpusne leksikografije uporabnejše besedne skice kakor pa obstoječe konkordance. Prvi cilj besednih skic je bil povečati zmogljivost metode skeniranja vrstic, tako da lahko dobimo karseda sistematične informacije o visokokakovostnih kolokacijah. Pomemben cilj je bil tudi skrajšati čas urejanja in izdelati kakovosten slovar. Pri projektu Macmillanovega slovarja je bilo mogoče ponovno potrditi pomembno vlogo konkordanc za izdelavo slovarja, po drugi strani pa za strokovnjake leksikografije besedne skice niso bile le izvor kolokacij, temveč je njihova vloga postala pomembna pri semantični analizi. Številne funkcije sistema SkE skupaj s seznamom kolokacij, povzetih po seznamu slovničnih povezav, kažejo glavne značilnosti obnašanja besed, kar pa se je izkazalo za prispevek k razumevanju besednih pomenov. Besedne skice tako veljajo za novo metodo sestavljanja slovarjev, ki temeljijo na korpusih.

41 Del rezultatov raziskave, predstavljene v tem poglavju, je objavljen v Kilgarriff et al. (2010) ter Srdanović & Nishina (2008).

SkE se je ravno tako uporabljal za gradnjo drugih slovarjev (Oxford University Press, Chambers Harrap in Collins). Uporablja se ga tudi pri poučevanju jezika kot vir pri sestavljanju različnih tipov testov, kakršen je npr. test stavkov z dopolnjevanjem (Smith et al., 2007, 2008; Smrž, 2004). Tako obstajajo številni primeri apliciranja orodja za razvoj jezikovnih tehnologij (Gatt in Van Deemter, 2006; Chantree et al., 2005).

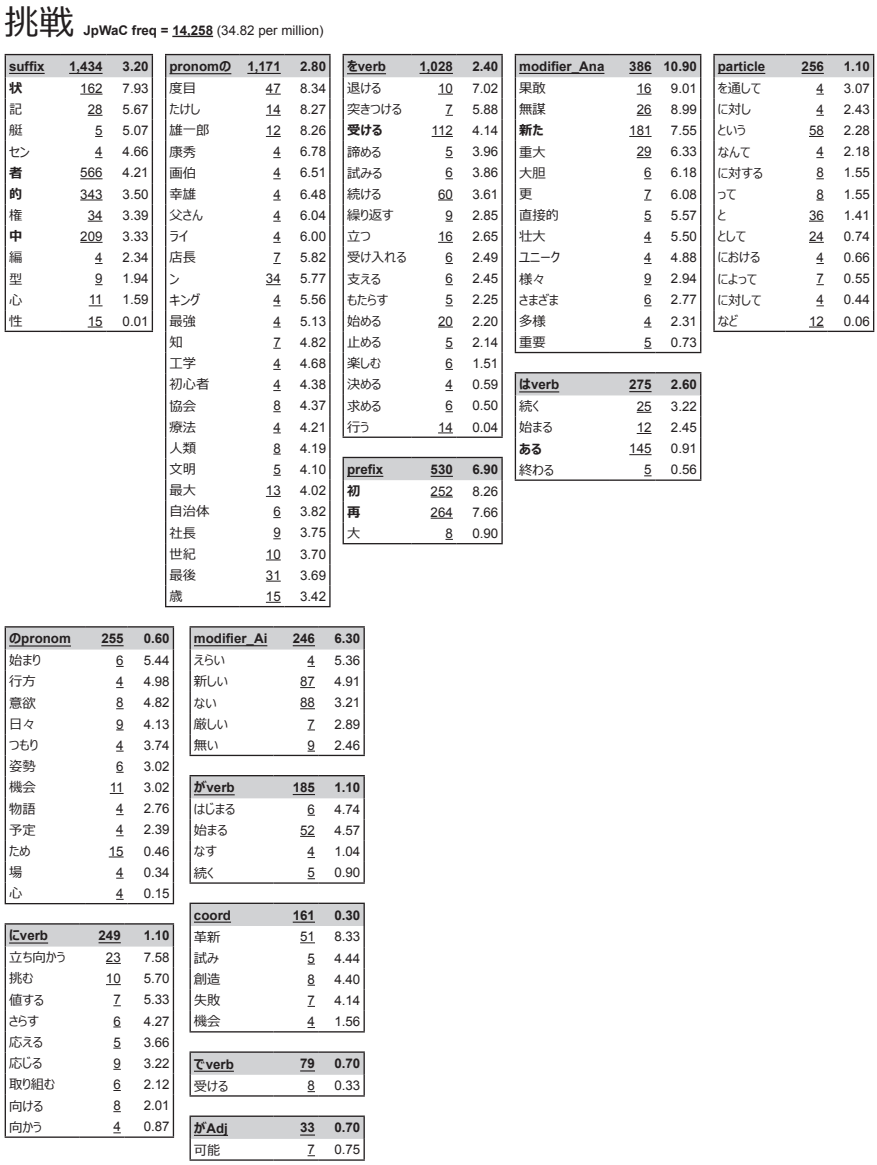
Angleška različica SkE je posebej prispevala k področju jezikovnega poučevanja in leksikografije. Tudi japonska različica bi se lahko v prihodnje uporabljala na podoben način.

V tem poglavju najprej razpravljamo o različnih možnostih uporabe sistema pri poučevanju japonščine, še posebno pa pri izdelavi slovarja japonščine. Beseda *chousen* »izziv« (挑戦) je analizirana z leksikalno-semantičnega vidika, predstavljenih je tudi nekaj primerov iskanja različnih slovničnih vzorcev in še druge možnosti uporabe. Sledita dve vrednotenji sistema SkE. Pri prvem vrednotenju so rezultati iz sistema SkE primerjani z naključno izbranimi primeri iz slovarja kolokacij *Nihongo hyougen katsuyou jiten* »Slovar uporabe japonskih izrazov v praksi« (日本語表現活用辞典) (Himeno, 2004). Pri drugem vrednotenju se ocenjuje, ali so rezultati kolokacij v sistemu SkE primerni za ureditev obsežnega slovarja japonskih kolokacij, kar je del projekta Sketch-Eval, ki je usmerjen k večjezičnemu vrednotenju sistema SkE.

3.2 Leksikalno-semantična analiza: primer besede *chousen* »izziv« (挑戦)

V tem poglavju smo s funkcijo Besedne skice poiskali japonsko besedo *chousen* »izziv« (挑戦), ki odgovarja angleški besedi *challenge* (za analizo besede v angleščini, gl. Kilgarriff in Rundell, 2002). Kolokacijske rezultate preučujemo z leksikalno-semantičnega vidika. Slika 8 prikazuje rezultate iskanja besede *chousen*.

Stolpca »modifier_Ana« in »modifier_Ai« na Sliki 8 kažeta pridevnike na *-na* in pridevnike na *-i*, ki modificirajo samostalnik *chousen*. Stolpci »は verb«, »を verb«, »が verb« in »に verb« kažejo, s katerimi glagoli se samostalnik *chousen* največkrat sopojavlja. V eni slovnični zvezi lahko pomensko razločimo več kolokacij, vidimo pa lahko tudi pomenska zveza med besedami, ki se pojavijo v različnih slovničnih povezavah. Če najprej pogledamo slovnične povezave na Sliki 8 – *atarashii chousen* »nov izziv« (新しい挑戦), *aratana chousen* »nov izziv« (新たな挑戦), *chousen wa hajimaru* »izziv/spopad se začne« (挑戦は始まる), *chousen wo hajimeru* »začeti izziv, spopasti se« (挑戦を始める), *chousen ga hajimaru* »izziv se začne« (挑戦が



Slika 8: *Besedne skice: rezultati iskanja besede* chousen »*izziv*« (挑戦).

始まる), *hajimeru* »začeti« (始める) in primere *chousen wo kokoromiru/motarasu/tsuzukeru* »poizkusiti/prinesti, sprožiti/nadaljevati izziv« (挑戦を試みる・もたらす・続ける) opazimo voljo, aktivno naravnanost in usmerjenost naprej. Beseda v teh kolokacijah izraža pomen »z lastno voljo se aktivno lotiti novih in na videz težkih stvari«.

Imamo pa še eno rabo, ki se povezuje z aktivnostjo, to je *chousen wo tanoshimu/motomeru* »uživati v izzivu/želeti si izziva, iskati izziv«（挑戦を楽しむ・求める）in *chousen ni atai suru* »biti vreden izziva«（挑戦に値する）.

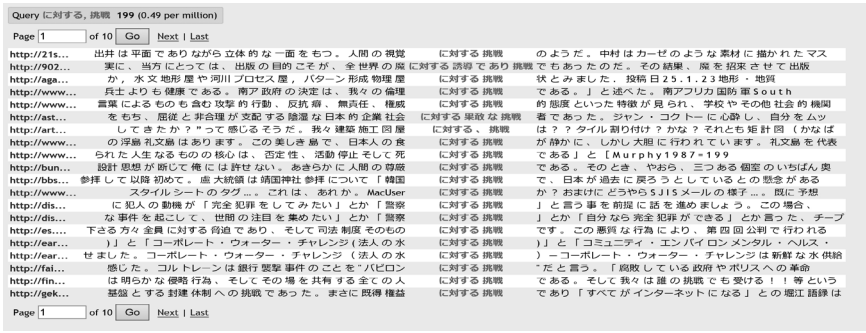
Po drugi strani pa se z besedo *chousen* sopojavljajo tudi besede *sai-* »ponovno«（再-）, *shippai* «neuspeh« (失敗) in *akirameru* «obupati, vdati se, opustiti« (諦める) ipd. in pomensko indicirajo neuspeh izziva. To so npr. *kibishii chousen* »težek izziv«（厳しい挑戦）, *mubouna chousen* »nepremišljen izziv«（無謀な挑戦）, *chousen wo kurikaesu/shirizokeru/akirameru/tomeru* »ponoviti/zavrniti/opustiti/ustaviti izziv«（挑戦を繰り返す・退ける・諦める・止める）. Obstajajo tudi izrazi, ki kažejo na pomembnost in izkazujejo močno voljo, npr. *juudaina/juuyouna chousen* »resen/pomemben izziv«（重大な・重要な挑戦）ter na pomembnost in težavnost izziva *chokusetsutekina/kakanna/daitanna chousen* »neposreden/drzen/neustrašen« (直接的な・果敢な・大胆な挑戦). Glagoli, ki se sopojavijo večkrat, pomenijo »sprejeti« oz. »prevzeti« (npr. *chousen wo ukeru*), so pa tudi primeri, ki pomenijo »zoperstaviti se« oz. »soočiti se« (*chousen ni tachimukau*). Takšna raba besede *chousen* pomeni »soočiti se s problemom po najboljših močeh in stremeti k prenovi« in »zoperstaviti se nečemu/nekomu«. Iz primerov je razvidno, da ne gre le za svojo lastno voljo oz. namen, ampak lahko tudi za voljo neke druge osebe.

Izraz *challenge to something/somebody* najdemo pri angleški analizi, medtem ko se v primeru japonščine ne pokaže s takšno stavčno strukturo. Zato smo vzorec poiskali s konkordancami. Kot kaže Slika 9, smo v okvirček CQL v konkordančnem oknu spodaj vnesli regularni izraz [word="に対する"] []{0,3} [word="挑戦"]. Pri tem interval {0,3} pomeni, da so med izrazoma に対する »proti« in 挑戦 »izziv« od 0 do trije znaki. Del rezultatov je prikazan na Sliki 10, nadalje pa jih lahko s sortiranjem razdelimo v naslednje pomenske sklope:

- Izziv proti avtoriteti (policija, avtoriteta, avtorske pravice, moč)
- Izziv proti skupini (demokratična družba, mednarodna skupnost, država, Amerika (アメリカ), ZDA（米国）, Južna Koreja)
- Izziv proti ideologiji (politika, principi, svoboda, mednarodna varnost, neoboroževanje).

Rezultati kažejo, da je zunanji pritisk na izzivalca velik, zato se pokaže struktura izzivalčeve močne volje.

Slika 9: *Iskanje* ~ ni tai suru + chousen (～に対する +挑戦).



Slika 10: *Prikaz s konkordancami za* ni tai suru chousen (に対する挑戦).

Sklenemo lahko, da besedne skice ne prikažejo le slovničnih razmerij in seznama preprostih kolokacij, ampak jih je moč uporabljati tudi kot vir leksikalno-semantične analize v slovarjih. Sistem bi lahko bil koristen tudi za primerjanje jezikov ali pa za izdelovanje dvojezičnih oz. večjezičnih slovarjev. Lahko bi ga na primer uporabili za urejanje obstoječega japonsko-slovenskega slovarja jaSlo (Erjavec et al., 2006).

3.3 Metodologija iskanja stavčnih vzorcev

Poleg opisane leksikalno-semantične analize, ki je razširjena v leksikografiji, se tudi v drugih jezikoslovnih študijah širi uporaba in metodologija korpusnih virov. SkE se tako lahko uporabi kot empirična metodologija za raziskovanje o jeziku. Čeprav se besedne skice, tezaver in primerjalne skice osredotočajo v glavnem na leksikalno-semantične podatke, so z vidika korpusnega jezikoslovja zanimivi tudi drugi podatki. V tem poglavju je nekaj primerov iskanja določenih struktur s funkcijo zloženega iskanja konkordanc, kjer lahko različne večenotne vzorce iščemo kot enote.

3.3.1 Morfološka produkcija in izpeljava

V obstoječih morfosintaktičnih oznakah v korpusu obstajajo naslednje morfološke izpeljave:

- suffix »pripona«（接尾辞）, prefix »predpona«（接頭辞）
- suffix_base »priponska osnova«（接尾辞が付く語幹）, prefix_base »predponska osnova«（接頭辞が付く語幹）
- bound_V »vezava_V« (odvisni in neodvisni glagoli in glagolske pripone, vezani na ključno besedo glagola)
- V_bound »V_vezava« (glagoli, na katere se pogosto veže ključna beseda, ki je neodvisni ali odvisni glagol ali glagolska zapona)

Na primer, predpona h glagolu *kiku* »slišati, vprašati«（聞く）je spoštljivi *o-*（お聞きしたい), obstajajo pa tudi primeri, kot so vezava *kumu*（組む）»sestaviti, združiti« v *torikumu*（取り組む）»spoprijeti se«, ali pa vezave *miru* »videti«（みる），*kureru* »dobiti«（くれる），*kudasaru* »dobiti«（くださる）v primerih *kiite miru, kiite kurenai ka*（聞いてみる »poskusiti poslušati«, 聞いてくれないか »me ne bi poslušal?«).

Ko iščemo besedo *hon* »knjiga«（本），se ta v nekaterih primerih pojavi kot predpona s pomenom »ta« (*honkenkyuu* 本研究 »ta študija«, *honsaabisu* 本サービス »ta usluga«), se pa pojavlja tudi kot samostalnik s pomenom »knjiga« skupaj s predpono (*gohon* ご本 »knjiga [spoštljivo]«, *gohon* 御本 »knjiga [spoštljivo]«, *kakuhon* 各本 »vsaka knjiga«). Klasifikacija v kategorijo števnika (助数詞) pa da rezultat tudi med kolokacijami (*-satsu no hon* ～冊の本 [število fizičnih knjig/zvezkov]).

S primerjalnimi skicami lahko prepoznamo priponi *sei*（性）za lastnosti oz. *sa*（さ）za posamostaljenje pridevnikov, njuni osnovi in razlike v kolokacijah. Pripona *sei*（性）se dodaja k osnovam pridevnikov na *-na* in samostalnikom (npr. *kanou* »verjetno« 可能 za *kanousei* »verjetnost« 可能性, *houkou* »smer« 方向 za *houkousei* »usmerjenost, usmeritev« 方向性, *seisan* »proizvod« 生産 za *seisansei* »produktivnost« 生産性), pripono *sa*（さ）se dodaja pridevnikom na *-i* (*ookii* »velik« 大きい za *ookisa* »velikost« 大きさ, *takai* »visok« 高い za *takasa* »višina« 高さ, *nagai* »dolg« 長い za *nagasa* »dolžina« 長さ). Tovrstne pripone se poučuje na osnovni ravni in z napredovanjem na višjo raven postane pomembno poznavanje besednih oblik, ki se ne pojavljajo skupaj. Pridevniku *yutaka* »bogat«（豊か）se ne dodaja pripone *sei*（性），medtem ko se pridevniku *benri* »priročen«（便利）doda pripono sa（さ). Pri besedah, kjer se lahko uporablja obe kolokaciji (*seikakusa/sei* »natančnost« 正確さ/性 in *fukuzatsusa/sei* »zapletenost« 複雑さ/性), so učenci pogosto v dilemi. S pomočjo orodja lahko hitro in natančno preverijo rabo tovrstnih kombinacij.

V slovarjih kombinacije samostalnika in sklona in informacije o sopojavljanju pomožnih glagolov in glagolskih pripon (npr. pasivni in kavzativni stavki) težko najdemo, medtem ko so z uporabo besednih skic in konkordančnika te lažje dosegljive. Glagoli, pridevniki na *-na* in pridevniki na *-i* so v rezultatih besednih skic prikazani kot leme, zato bi bilo pričakovano neposredno pridobiti tudi podatke o njihovem spreganju.

3.3.2 Iskanje vzorcev

V okviru kolokacijskih podatkov se pojavijo slovnični podatki o tem, kaj je glagolski osebek ali kaj je predmet oz. tema. Z uporabo konkordanc lahko prav tako iščemo različne preproste ali zapletene vzorce. Metodologija iskanja s konkordancami

je opisana v razdelku 2.3.3., na tem mestu pa je podanih nekaj primerov iskanja z uporabo funkcije CQL v konkordancah.

1. Primeri stavčnih vzorcev s … *kara /de tsukurareru* »biti narejeno od«(～から／で作られる)

Skupne točke in razlike podobnih stavčnih izrazov … *kara tsukurareru*（～から作られる）in … *de tsukurareru*（～で作られる）lahko poiščemo s konkordancami. Za iskanje posameznega vzorca je treba v okvirček CQL vnesti spodnja znakovna niza (ang. *character string*, jp. *mojiretsu* 文字列)：

```
[word="から"][word="作ら"][lemma="れる"]
[word="で"][word="作ら"][lemma="れる"]
```

V znakovne nize vnesemo lemo *reru* れる zato, da bo rezultat vključeval pregibne oblike, kot so npr. povedkovna oblika (*shuushikei* 終止形), prilastkova oblika (*rentaikei* 連体形), nedovršna oblika (*mizenkei* 未然形) in vezna oblika na *-masu* (*renyoukei* 連用形).

V rezultatih frekvenc se izraz *kara* »od, iz«（から）pojavi 432-krat, izraz *de* »od, iz«(で) pa 975-krat. Ko posamezen rezultat iščemo s funkcijo kolokacijskega kandidata, dobimo naslednje rezultate.

Kara（から）se sopojavlja s spodnjimi besedami:

- Naravni materiali oz. deli (*shokubutsu* »rastline« 植物, *kome* »riž« 米, *ki* »drevo, les« 木, *budou* »grozdje« ブドウ, *satoukibi* »sladkorni trs« サトウキビ, *sozai* »materiali« 素材, *shokuzai* »hrana, jestvine« 食材, *genryou* »surovine« 原料, *koshi* »odpadni papir« 古紙, *dojou* »prst, zemlja« 土壌, *sekiyu* »nafta« 石油)
- Abstraktni in specifični deli (*kotoba* »beseda« 言葉, *hansei* »refleksija« 反省, *imi* »pomen« 意味)
- Dobe in vidik mišljenja (*shiten* »stališče« 視点, *jidai* »doba« 時代, *koro* »čas, obdobje« 頃, *toshi* »leto« 年, *ikou* »odkar, po« 以降, *kanten* »vidik« 観点)

De（で）se sopojavlja s spodnjimi besedami:

- Lokacija, ustvarjalec (*koujou* »tovarna« 工場, <lokacija>+*no naka* »v«, *Nihon* »Japonska« 日本, *Amerika* »Amerika« アメリカ, <lokacija>+*no ue* »nad«, *Doitsu* »Nemčija« ドイツ, *Chuugoku* »Kitajska« 中国, *kuni* »država« 国, *tainai* »v telesu« 体内, *katei* »dom« 家庭, *chiiki* »regija« 地域 ipd.), ustvarjalec (*jibun* »jaz« 自分, *sutaffu* »osebje« スタッフipd.)
- Cilj (*mokuhyou* »cilj« 目標, *youso* »element« 要素, *nerai* »cilj« 狙い, *tsumori* »namen« つもり), mišljenje (*konseputo* »koncept« コンセプト, *hassou* »ideja« 発想, *imeeji* »podoba« イメージ, *kangae* »mišljenje« 考えipd.), način (*gijutsu*

»umetnost, spretnost, tehnika« 技術, *houhou* »metoda, način« 方法, *zentei* »predpostavka« 前提, *gihou* »tehnika« 技法, *kyouryoku* »sodelovanje« 協力, *tejun* »proces« 手順, *shudou* »voditi« 主導, *jinkou* »umeten« 人工), stanje (*dankai* »faza« 段階, *reberu* »raven« レベル; *tanki* »kratko obdobje« 短期; *kosuto* »stroški« コスト, *yosan* »proračun« 予算)

- Material (*ki* »les« 木, *sozai* »surovine« 素材, *zairyou* »sestavine« 材料, *ishi* »kamen« 石, *kami* »papir« 紙, *busshitsu* »snov« 物質, *kinzoku* »kovina« 金属, *ha* »list (na rastlini)« 葉, *garasu* »steklo« ガラス, *mokuzai* »hlodovina, les« 木材, *take* »bambus« 竹, *goukin* »zlitina« 合金, *kin* »zlato« 金, *shokuzai* »hrana, jestvine« 食材, *biizu* »kroglica za nakit« ビーズ, *purasuchikku* »plastika« プラスチック, *tamago* »jajce« 卵, *dairiseki* »marmor« 大理石)

Če se omejimo na kolokacije materiala, se *kara* in *de* sopojavljata npr. v kontekstih, vezanih na les in hrano, *kara* pa v primerih, ko gre za naravne stvari in primarne surovine, po drugi strani pa je za *de* mogoče opaziti trend sopojavljanja s človeškimi izdelki in s stvarmi, ki so narejene iz določenega materiala.

2. Primeri stavčnih vzorcev za kavzativ (glagol + *sasete ageru* »bom ti dal, da …« させてあげる)

V nadaljevanju pokažemo še en primer načina iskanja kavzalnih izrazov. Najprej v okvirček CQL vnesemo niz [tag="V.*"][word="せ|させ"][word="て"][lemma="あげる"]. Kavzalni pomožni glagol (*se* せ oz. *sase* させ) se veže na vse vrste glagolov pred njim. Poleg tega pa je mogoče najti izraze, sestavljene iz serij pregibnih oblik, ki sledijo oblikama *te* て in *ageru* あげる. Če iščemo v tem formatu, znaša frekvenca v spletnem korpusu 1170. Če še naprej iščemo rezultate s funkcijo kolokacijskega kandidata, so osnove glagola, ki se pojavijo več kot 10-krat, naslednje:

- *su* す (osnova od *suru* »delati«), *tabe* 食べ (osnova od *taberu* »jesti«), *ki* 聞（き）(osnova od *kiku* »slišati/spraševati«), *yasu* 休 (osnova od *yasumu* »počivati«), *ya* や (osnova od *yaru* »delati«), *mo* 持 (osnova od *motsu* »nositi, imeti«), *yoroko* 喜 (osnova od *yorokobu* »veseliti se«), *shi* 知 (osnova od *shiru* »znati, vedeti«), *no* 飲 (osnova od *nomu* »piti«), *a* 会 (osnova od *au* »srečati se«), *tanoshi* 楽し(osnova od *tanoshimu* »veseliti se«), *i* 行 (osnova od *iku* »iti«), *yo* 読 (osnova od *yomu* »brati«), 気づ (osnova od *kizuku* »opaziti«), *aso* 遊 (osnova od *asobu* »zabavati se/igrati se«), *ka* 勝 (osnova od *katsu* »zmagati«), *amae*甘え (osnova od *amaeru* »razvajati«).

3. Primeri za besedno zvezo *ki ni suru* »skrbeti«（気にする）

S konkordancami se lahko išče tudi kolokacije besednih zvez. Na primer, če uporabimo iskalno poizvedbo [word="気"] [word="に"] [lemma="する"], dobimo

10.845 zadetkov. Z uporabo funkcije kolokacijskega kandidata se znotraj primera ～を + 気にする število primerov približa na 4000. Tabela 9 prikazuje leksikalno-semantične tipe, ki se pojavijo največkrat.

Tabela 9: *Leksikalno-semantični tipi kolokacij za zvezo* wo ki ni suru »*skrbeti za*« (を気にする).

Leksikalno-semantični tipi	Kolokacije
mnenje, ocena, napoved druge osebe	*tanin* »druga oseba« 他人, *sekentei* »spodobnost« 世間体, *hannou* »odziv« 反応, *hyouka* »vrednotenje« 評価, *kinjo* »soseščina« 近所, *mawari* »okolica« 回り, *mawari* »okolica« 周り, *hyouban* »ugled« 評判, *yoron* »javno mnenje« 世論, *hito* »človek«人, *shuui* »okolica« 周囲, *omowaku* »pričakovanja« 思惑, *mesen* »zorni kot« 目線, *me* »očesno zrklo« 眼, *me* »oko« 目, *shisen* »pogled« 視線, *gaiken* »videz« 外見, *uranai* »prerokovanje« 占い, *meishin* »vraževerje« 迷信
lasten videz/stanje oz. videz/stanje nekoga drugega	*daietto* »dieta« ダイエット, *sutairu* »stil« スタイル, *teisai* »videz« 体裁, *nenrei* »starost« 年齢, *fasshon* »moda« ファッション, *fukusou* »oblačilo« 服装, *yogore* »umazanija, nečistost« 汚れ, *kamigata* »pričeska« 髪型, *taijuu* »telesna teža« 体重, *futori* »debelost« 太り, *mitame* »videz« 見た目, *youshi* »videz« 容姿, *mibae* »videz« 見栄え, *hiyake* »sončna opeklina« 日焼け
vreme	*ame* »dež« 雨, *tenki* »vreme« 天気, *youhou* »vremenska napoved« 予報, *tenkou* »vreme« 天候, *shigaisen* »ultra-vijolični žarki« 紫外線
bolezen, bolečina	*yamai* »bolezen« 病, *itami* »bolečina« 痛み, *guai* »zdravstveno stanje« 具合, *kizu* »poškodba« 傷, *jibyou* »kronična bolezen« 持病, *kenkou* »zdravje« 健康
finančno stanje	*yosan* »proračun« 予算, *kabuka* »cene delnic« 株価, *ryoukin* »stroški« 料金, *nedan* »cena« 値段, *saifu* »denarnica« 財布, *nenpi* »poraba goriva« 燃費
čas	*jikan* »čas« 時間, *tokei* »ura (aparat)« 時計, *jisa* »časovna razlika« 時差, *shuuden* »zadnji vlak« 終電, *okure* »zamuda« 遅れ
zaznavni dražljaji	*nioi* »vonj« 匂い, *souon* »hrup« 騒音, *kemuri* »dim« 煙, *kusai/nioi* »(neprijeten) vonj« 臭い, *zatsuon* »hrup« 雑音

rezultati	*seiseki* »uspešnost, (šolska) ocena« 成績, *tensuu* »število točk« 点数, *kekka* »rezultat« 結果, *machigai* »napaka« 間違い, *haisen* »poraz« 敗戦
vsebina	*nakami* »vsebina« 中身, *kashi* »besedilo pesmi« 歌詞, *bunpou* »slovnica« 文法, *saibu* »podrobnost« 細部, *gashitsu* »kakovost slike« 画質, *ontei* »(glasbeni) interval« 音程
drugo (vrstni red, količina, razdalja itd.)	*jun«i* »pozicija« 順位, *junban* »zaporedje« 順番, *suuchi* »številčna vrednost« 数値, *ritsu* »delež« 率, *kazu* »število« 数, *youryou* »kapaciteta, volumen« 容量, *kyori* »razdalja« 距離, *doukou* »trend« 動向

Da bi raziskali leksikalno-semantične tipe besed, ki se pojavljajo pred zvezo … *no koto wo ki ni suru* »skrbi me za …« (〜のことを気にする), iščemo naslednji vzorec: [word="の"][word="こと|事"][word="を"][word="気"][word="に"][lemma="する"]. Vidimo lahko, da se pogosto pojavljajo besede, ki se nanašajo na ljudi, npr. *kodomo* »otrok«（子供）, *kare* »on, tisti«（彼）, *anata* »ti«（あなた）, *hito* »človek«（人）, *watashi* »jaz«（私）, *-san* »gospod, gospa«（〜さん）, *-mono/sha* »oseba«（〜者）itd.

Iz podatkov, ki so označeni v korpusu in so v slovnični datoteki zaznamovani kot primerni vzorci, je s funkcijo Besedne skice mogoče hitro pridobiti kakovostne rezultate. V prihodnje bo z dodajanjem dodatnih korpusnih oznak za različne druge stavčne izraze, besedne zveze in modalnost mogoče pridobiti bogatejše jezikovne informacije iz rezultatov besednih skic. Raziskava pridobivanja kolokacij med prislovi ugibanja in modalno obliko na koncu stavka je opisana v 4. in 5. poglavju.

3.4 Področje poučevanja japonščine in možna raba sistema

V zadnjih letih so se začele jezikovne tehnologije uporabljati tudi na področju učenja drugega tujega jezika. V ta namen se uporabljajo različice sistema SkE v več jezikih, njihova uporabnost pa je pogosto predmet raziskav. Japonska različica je prav tako lahko koristen vir za učenje japonščine na Japonskem kot tudi v tujini in je lahko v pomoč tako učencem kot tudi učiteljem japonščine. V nadaljevanju je navedenih nekaj možnosti za uporabo sistema SkE.

3.4.1 Z vidika uporabnika

Najprej poglejmo možnosti uporabe SkE z vidika uporabnikov, in sicer a) učitelja japonščine in b) učenca japonščine. Učitelji japonščine so lahko materni govorci

ali tuji govorci in oboji pri uporabi SkE naj ne bi imeli večjih težav. Če je učitelj tuji govorec, se v nekaterih primerih pri bolj delikatnih jezikovnih zadregah ne bo mogel enostavno opredeliti. Pri učencih pa se možnosti uporabe orodja razlikuje glede na njihovo znanje japonščine. Medtem ko naj bi ga učenci na srednji in višji stopnji znali uporabljati brez večjih težav, je zelo verjetno, da bo za učence nižje stopnje neposredna raba predstavljala težavo.

Če je uporabnik učenec, se lahko pojavijo težave pri učenju jezika med uporabo orodja; na primer, ko se pojavi neznana beseda ali pismenka, ki je učenec ne zna prebrati. Orodje namreč ne vsebuje informacije o težavnostni stopnji, kar otežuje pedagoško rabo. Na voljo je nekaj računalniških podpornih sistemov, ki stopnjo vendarle upoštevajo. Obstaja npr. orodje za pomoč pri učenju pisanja spisov Natsume, ki pri izbiri vzorčnih stavkov iz številnih korpusov uporablja podatke o ravni znanja (Nishina in Yoshihashi, 2007; Nishina, 2008; Hodošček in Nishina, 2012). Prav tako obstaja orodje za iskanje primerov iz korpusa JpWaC-L (Hmeljak in Erjavec, 2012) glede na stopnje težavnosti s seznama besedišča JLPT. Opravljene so bile tudi raziskave sortiranja stavčnih primerov v konkordančnih vrsticah glede na njihovo zloženost oz. težavnost (Smrž, 2004). Če bo v prihodnje v SkE vdelana tovrstna funkcionalnost, bo sistem verjetno postal lažji za uporabo, prav tako pa bi bilo pri japonski različici dodati tudi furigano (zapis branja pismenk).

3.4.2 Z vidika štirih jezikovnih spretnosti

Gledano z vidika štirih jezikovnih spretnosti, to so branje, pisanje, poslušanje in govor, je cilj orodja SkE podpora za pisanje. Hkrati pa lahko z branjem primerov iz korpusa posredno pomaga izboljšati bralno razumevanje.

3.4.3 Z vidika učnih ciljev

Možnosti aplikacije sistema lahko razdelimo glede na učne cilje.

a. učenje različnih jezikoslovnih vedenj (npr. poznavanje stavčnih vzorcev in besednih pomenov)
b. ocenjevanje jezikovnih sposobnosti (sestavljanje testov)
c. sestavljanje učnih virov (npr. učbenikov)
d. izdelovanje računalniškega sistema za podporo učenju

Posamezna jezikoslovna znanja učnih vsebin, kot so različna leksikalno-semantična ali morfo-sintaksična vprašanja, ter primeri slovničnih vzorcev so opisani v razdelkih 3.2. in 3.3. Glede na izkušnje z uporabo SkE v izobraževanju (Smrž, 2004) je bila pri poučevanju drugega tujega jezika poleg kolokacijskih izrazov

najbolj koristna funkcija Primerjalne skice, s katero lahko primerjamo podobnosti in razlike med sopomenkami in blizupomenkami. Z uporabo tezavra je moč oceniti znanje, povezano z besediščem, ki izraža specifičen pomen. Smrž (2004) meni, da SkE omogoča izdelavo samodejnega testiranja, s katerim lahko preverimo, ali razumemo razlike med besedami. Pri izdelavi učnega gradiva, npr. učbenika, lahko uporabimo različne jezikoslovne podatke, ki so pridobljeni iz sistema SkE (npr. jezikovne, morfološke in slovnične vzorce ter tipične stavčne primere). Potem lahko generiramo seznam besed za izdelavo učnih virov, iz katerega se lahko učenec osredotoči na besede z visoko frekvenco. Prav tako lahko orodje uporabimo za razvoj drugih računalniško podpornih sistemov s funkcijo prikazovanja visokofrekvenčnih kolokacij, sopomenk in protipomenk.

3.5 Drugi načini uporabe

Poleg omenjenih načinov aplikacije je SkE moč uporabljati tudi na drugih področjih raziskovanja. Z vidika sociolingvistike lahko raziskujemo raznovrstne pojave, kot so npr. razlike v rabi besed in kolokacij o moškem in ženskem spolu (gl. primer *onna no ko* in *otoko no ko* na Sliki 7). Prek besed in kolokacij lahko opazimo družbeno sprejete ideje in stereotipe. Z uporabo sedanjih obsežnih spletnih korpusov je z vidika kulturoloških študij mogoče analizirati tudi spletne medije. S primerjavo različnih korpusov se lahko preučuje tudi posebnosti posameznega korpusa. Zanimivo je tudi primerjati rezultate obsežnih korpusov z rezultati mnenjskih raziskav. Na primer, mogoče je primerjati mape besed, ki prikazujejo povezave besed in njihovih asociacij, izluščenih z anketami (Joyce, 2005), z rezultati besednih skic in tezavra. Rezultati tovrstnih analiz so pokazali, da se besedne kombinacije, pridobljene iz korpusov, v precejšnji meri prekrivajo z besednimi kombinacijami, pridobljenimi z asociacijami, mogoče pa je opaziti tudi razlike v kombinacijah (Joyce in Srdanović, 2008). Ker lahko v rezultatih iskanja s SkE odkrijemo različne pomanjkljivosti analize morfemov z orodjem ChaSen, lahko rezultate uporabimo za izboljšavo natančnosti morfosintaktične analize japonskega jezika. Mogoče je uporabiti tudi funkcijo Corpus Builder, s katero sami zgradimo in/ali naložimo korpus v sistem SkE. Z uporabo funkcije WebBootCat lahko sestavimo specializirani korpus in ga apliciramo na želeno področje, npr. jezikovno izobraževanje, turizem in podobno (Baroni et al., 2006; Smith et al, 2008).

3.6 Vrednotenje 1: primerjava s slovarjem kolokacij

V tem poglavju primerjamo japonsko različico sistema SkE s Slovarjem uporabe japonskih izrazov v praksi (Himeno, 2004) ter ovrednotimo potencialno rabo

sistema za luščenje kolokacij. To je prvi slovar kolokacij za učence japonščine. Za razliko od dotedanjih slovarjev japonščine, pri katerih je bil v središču »opis besede«, ta slovar izpostavi »povezavo z besedo«. Vsebuje številne stavčne primere in kolokacijske podatke, zato je koristen vir za učence japonščine. V času izdelave obsežni korpusi japonščine še niso obstajali, zato so bili uporabljeni raznovrstni jezikovni viri (npr. drugi slovarji, literarna dela, časopisi ipd.).

Da bi ugotovili, ali je pri tovrstnem slovarju uporaba besednih skic prednost, smo iz slovarja naključno izbrali naslednjih 10 gesel in jih poskušali primerjati z rezultati, pridobljenimi iz orodja SkE z besednimi skicami:

utsumuku »pogledati dol/povesiti glavo«（俯く）, *kasuka* »nejasen, bežen«（微か）, *kurushimu* »trpeti«（苦しむ）, *shimeru* »zapreti«（閉める）, *taberu* »jesti«（食べる）, *tomeru* »prespati«（泊める）, *hakobu* »prenašati«（運ぶ）, *betsubetsu* »posamezno«（別々）, *meiryou* »jasen, razločen«（明瞭）, *waru* »razdeliti, razbiti«（割る）.

V nadaljevanju navajamo nekaj primerov rezultatov primerjave in diskutiramo o štirih točkah možne rabe sistema za luščenje kolokacij: 1) za številne slovnične povezave 2) kot metoda za izbiranje kolokacij 3) za izbiranje stavčnih primerov 4) za leksikalno-semantične podatke.

3.6.1 Uporaba za številne slovnične povezave

V slovarju japonskih kolokacij so opisane raznovrstne kolokacije. Prikazani so a) kolokacijski podatki za glagole – posamezni sklonski členki *ga*, *wo*, *to*, *ni* in veza s samostalniki ter prislovi, ki modificirajo glagole, b) kolokacijski podatki za pridevnike na *-na* – členki *na* in *no* in z njimi povezani samostalniki ter *ni*, *te* in druge oblike s prislovno funkcijo. Po drugi strani imamo 22 pravil slovničnih povezav v orodju SkE, ki so prikazane z binarno povezavo in pokrivajo več kot 50 različnih slovničnih povezav, kar zajema veliko več kolokacijskih informacij od slovarja. Poleg glagolov in pridevnikov na *-na* lahko iščemo tudi druge besedne vrste kolokacijskih gesel (še posebno za samostalnike, pridevnike na *-i* in prislove). Vsebuje tudi veliko kolokacijskih povezav posameznih besednih vrst. Na primer, pri glagolskih geslih so prikazani tudi sklonski členki *ga*, *wo*, *to*, *ni*, *de*, *made*, *kara* in *he*, tematski členek *wa* ter samostalniki, ki se vežejo nanje. Poleg tega vsebuje prislove, odvisne glagole, glagolske pripone in njihove kolokacije ter druge neodvisne glagole in paralelne povezave. Gledano z vidika prostora je v slovarju nemogoče zajeti vsa kolokacijska gesla, zato je to prepuščeno leksikografski strategiji. Po drugi strani pa je z orodjem SkE mogoče izračunati kolokacijsko

frekvenco in statistično prioriteto, na podlagi teh podatkov pa lahko vzpostavimo različne kolokacijske kategorije ter njihova najpomembnejša gesla, kar je lahko velik prispevek k izdelavi slovarjev. Tu je prednost orodja SkE očitna, čeprav je primerjava s slovarjem japonskih kolokacij pokazala pomanjkljivo obravnavo povezave <osnova pridevnika na *-na* + *ni* + glagol> in tako nakazala mesto, na katerem je bilo mogoče sistem še izboljšati. Ta povezava je bila potem naknadno dodana v sistem.

3.6.2 Metoda za izbiranje kolokacij

V nadaljevanju opišemo različne kolokacije, ki se pojavijo v posameznih slovničnih povezavah.

V slovarju so naštete številne kolokacije znotraj določenega tipa kolokacije, toda ne pojavijo se vedno tiste, ki imajo v korpusu visoko frekvenco. Na primer, beseda *kasuka* »nejasen, bežen« (微か) se v korpusu večkrat pojavi v kontekstu *kasukana kioku* »nejasen spomin« (微かな記憶), medtem ko te kombinacije v slovarju ne najdemo. Z vidika števila naštetih kolokacij lahko opazimo, da slovar vsebuje gesla z visokim in nizkim številom kolokacij. Imamo tudi gesla, ki se pojavijo le kot kolokacije v stavčnih primerih. Glede na to lahko vidimo uporabnost besednih skic pri izbiri kolokacij za slovar glede na frekvenco in pomembnost kolokacij v korpusu.

Po drugi strani pa se včasih kolokacije, ki obstajajo v slovarju, ne pokažejo med rezultati besednih skic. Možen razlog je uporaba spletnih podatkov sodobne japonščine v sistemu SkE, medtem ko je omenjeni slovar sestavljen predvsem z uporabo literarnih del, ki pogosto vsebujejo besedila modernega japonskega jezika. Obe vrsti gradiva sta do določene mere omejeni, in nujno je, da v orodje SkE vnesemo tudi druge korpuse in rezultate primerjamo z obsežnimi zbirkami podatkov ter da se za tovrstni slovar uporabi širše gradivo.

3.6.3 Uporaba za izbiranje stavčnih primerov

Čeprav so kolokacijski izrazi v slovarju ponazorjeni s primeri, je treba pazljivo izločiti stavčne primere, da pokažemo najpomembnejše kolokacije in stavčne vzorce. Potem ko dobimo rezultate visokih frekvenc med različnimi kolokacijskimi izrazi, so besedne skice koristne pri izbiranju stavčnih primerov. Na primer, v slovarskem geslu *utsumuku* najdemo naslednje tipične primere: *akai kao wo shi, utsumuita* »zardela je in sklonila glavo« 赤い顔をし、うつむいた; *shibaraku utsumuite kangaete ita* »nekaj časa je gledala dol in premišljevala« しばらくうつむいて考えていた; *hana ga utsumuite iru* »navzdol obrnjeni cvet« 花がうつむいている; *utsumuite*

shimatta »pogledal je dol/sklonil je glavo«うつむいてしまった; *utsumuite aruku* »hoditi s sklonjeno glavo« うつむいて歩く). Med stavčnimi primeri, uporabljenimi v podatkih besednih skic (Slika 11), so tudi primeri, ki jih ni v slovarju, in rečemo lahko, da so med njimi tudi kolokacijski kandidati (*hazukashisouni utsumuita* »sramežljivo je gledal dol/sklanjal glavo« はずかしそうにうつむいていた; *utsumuite hissori to nakidashita* »začel je tiho jokati s sklonjeno glavo« うつむいてひっそりと泣き出した; *shita wo utsumuitamama* »gledajoč navzdol«下をうつむいたまま; *mugon de utsumuku* »gledati dol brez besed« 無言でうつむく; »sklonjeno, povešavo cveteti« *utsumuki kagen ni saku*うつむき加減に咲く; *utsumukigachi de atta* »bil je nagnjen k sklanjanju« うつむきがちであった; *chotto utsumuki kagen no atama* »z rahlo sklonjeno glavo« ちょっとうつむきかげんの頭; *sukoshi utsumukinagara* »rahlo upognjeno« 少しうつむきながら). Z raziskovanjem besednih skic in konkordanc je mogoče izbrati tipične stavčne primere v slovarju, ki še bolj podrobno orišejo rabo določenih kolokacij.

うつむく **JpWaC freq = 583** (1.42 per million)

coord	350	4.00
ふり向く	4	8.22
かげる	3	7.29
うなだれる	3	6.97
考え込む	3	5.72
微笑む	4	4.78
咲く	5	3.74
閉じる	4	3.54
黙る	4	3.47
泣く	7	3.13
歩く	19	2.98
座る	7	2.89
見つめる	3	2.53

nounは	127	8.00
彼女	13	2.79
彼	10	1.54

nounを	104	1.60
顔	21	3.18
肩	3	2.98
頭	5	1.32
下	7	1.29

modifier_Adv	63	8.30
やや	6	6.55
終始	3	6.03
しばらく	5	4.27
少し	10	3.91
いつも	4	3.67
ずっと	3	3.23
ちょっと	7	2.99
また	3	2.43

nounに	41	1.20
げ	3	2.79

nounで	34	2.00
無言	4	6.22

nounが	24	0.90
花	3	1.68

suffix	18	0.30
かげん	7	9.85
がち	5	3.66

Slika 11: *Besedne skice: rezultati iskanja za besedo* utsumuku *»pogledati dol/povesiti«.*

3.6.4 Uporaba za leksikalno-semantične podatke

V procesu izdelave slovarja, kot je omenjeni slovar kolokacij, lahko z uporabo besednih skic, tezavra in primerjalnih skice pridobimo podrobnejše pomenske informacije, kot so sopomenke, protipomenke ter razlike med njimi. Slovar je zasnovan tako, da če ima geslo A podobne kolokacije geslu B, so kolokacije prikazane le z geslom A, geslo B pa ima samo referenco na geslo A. To je recimo primer za gesla *shimeru* »zapreti«（閉める）in *shimaru* »zapreti se«（閉まる）, ter *tomeru* »dati prenočišče, prenočiti (prehodno)«（泊める）in *tomaru* »ostati, prenočiti (neprehodno)«（泊まる）. Če napravimo besedne skice za *shimeru* in *shimaru*, se poleg osnovne razlike v prehodnem in neprehodnem glagolu pokažejo tudi nekateri drugi zanimivi rezultati.

Najprej so to kolokacije s priponami: npr. *shimeru* se pogosto veže na *-rareru*, *-saseru* in *-ppanashi* (*Doa ga kichinto shimerarete imasu* »Vrata so dobro zaprta.« ドアがきちんと閉められています; *Amado o shimesaseru* »Naročiti, da zapre polkna« 雨戸を閉めさせる; *Kaaten o shimeppanashi ni suru* »Pustiti zavese odprte« カーテンを閉めっぱなしにする). Pokaže tudi, da se *shimeru* veže na členek *de* »s, z«, ki izraža način delovanja ali orodje (kot na primer, *ushirode de* »z rokami za hrbtom« 後ろ手で, *te de* »z rokami« 手で, *kagi de* »s ključem« 鍵で). Prikaže se tudi sestavljen členek *tame ni* »zaradi« (ために), npr. *Anzen no tameni mado wo shimete iru* »Okno je zaprto zaradi varnosti« (安全のために窓を閉めている). Obstajajo pa tudi kolokacije z neodvisnimi in odvisnimi glagoli ter glagolskimi priponami, ki se vežejo na glagol, na primer *kiru, naosu, kureru/itadaku/morau*: *shimekiru*, *shimenaosu*, *shimete kureru/itadaku/morau* »do konca zapreti, ponovno zapreti, zapreti [+ glagoli dajanja in sprejemanja]« (閉めきる, 閉め直す, 閉めてくれる・いただく・もらう).

Po drugi strani pa pri glagolu *shimaru* vidimo kolokacije, ki tvorijo zložene glagole, kot so *shimarikakeru* »zapreti« (閉まりかける), *shimatteoru* »biti zaprto« (閉まっておる) in *shimarihajimeru* »začeti zapirati se« (閉まり始める). Samostalniki *doa* »vrata« (ドア), *mado* »okno« (窓), *kaaten* »zavese« (カーテン) in *shattaa* »roleta« (シャッター) se pogosto vežejo na oba glagola (*shimeru* in *shimaru*), medtem ko se samostalniki *resutoran* »restavracija« (レストラン), *shouten* »trgovina« (商店) in *toshokan* »knjižnica« (図書館) vežejo le na glagol *shimaru*.

V slovarju z glagoloma *shimeru* in *shimaru* ne najdemo kolokacij s samostalniki *amado* »polkna« (雨戸), *tobira* »vrata« (扉), *futa* »pokrov« (蓋), *motosen* »ventil« (元栓), *jaguchi* »pipa« (蛇口), *barubu* »žarnica« (バルブ). Zanimivo je, da se kolokaciji *shouji* »drsno (papirnato) okno ili vrata« (障子) in *fusuma* »drsna vrata« (襖), ki sta v slovarju, v spletnem korpusu nista izpostavljeni. Spletni podatki odražajo sodobno japonščino in rečemo lahko, da se vzporedno z družbenimi spremembami pokažejo tudi spremembe v jeziku.

Japonska različica besednih skic torej lahko priskrbi raznovrstne kolokacijske podatke, primerne za izdelavo kolokacijskega slovarja, doda podatke o pomenih besed ter pomaga pri izboru stavčnih primerov. Poleg tega lahko besedne skice razumemo kot pomoč ne le pri izdelavi slovarja kolokacij, temveč tudi pri izdelavi raznih slovarjev (japonski jezikoslovni slovar, dvojezični slovarji, tezaver, slovar sopomenk in protipomenk ter slovar stavčnih vzorcev). Čeprav o času, ki je potreben za urejanje določenega slovarja, še nimamo podatkov, lahko po angleški različici sklepamo, da bi uporaba besednih skic morala tudi skrajšati čas izdelave slovarja. Po drugi strani pa sklepamo, da lahko z uporabo različnih jezikovnih virov, vključno s slovarjem japonskih kolokacij, slovnično datoteko v orodju SkE še izboljšamo.

3.7 Vrednotenje 2: vrednotenje izdelave kolokacijskega slovarja

V tem podpoglavju ovrednotimo orodje SkE za uporabo za urejanje obsežnega japonskega slovarja kolokacij. Vrednotenje je del projekta SketchEval, ki je usmerjen na večjezično vrednotenje sistema. Iz posameznega korpusa smo naključno izbrali 42 enot (samostalnike, pridevnike na *-i* in glagole) ter z rezultati orodja SkE ocenili, ali so primerni za izdelavo obširnega kolokacijskega slovarja. Za tovrstno vrednotenje smo najprej sestavili seznam besed z visoko, srednjo in nizko frekvenco za vsako besedno vrsto. Če je v seznamu beseda, ki bi morala biti zaradi morfoloških razlogov izpuščena iz vrednotenja, smo z rezervnega seznama izbrali primerno besedo in jo zamenjali. Slika 12 prikazuje seznam besed in rezervni seznam, ki sta bila pridobljena iz korpusa JpWaC. Gesla, ki so bila izključena iz vrednotenja, so prečrtana, gesla, ki so bila uporabljena namesto izbrisanih, pa so poudarjena. Na primer, pridevnik na *-i -ppoi* (っぽい) je pripona za izražanje podobnosti in zaradi tega je bil zamenjan s pridevnikom *omoi* »težak« (重い), ta pa je bil potem uporabljen za vrednotenje.

	Sample lists			Reserve lists		
	Nouns	**Verbs**	**Adjectives**	**Nouns**	**Verbs**	**Adjectives**
Common (100-2999) minfr = 10067 maxfr = 481866	N = 1963 急 研究 完成 男性 緑 評価	N = 367 生まれる 扱う 支払う 忘れる	N = 74 よろしい ~っぽい~ 素晴らしい 大きい	心配 箱 積極 プロセス 地区 建設	ちゃう 知れる 語る 受け入れる	重い 長い 忙しい 無い
Mid (3000-9999) minfr = 1931 maxfr = 10061	N = 5326 欠席 蓄積 マスター 俳句 情勢 有力	N = 840 まつ ~づく~ 資する 溜まる	N = 105 黒い おとなしい こい 親しい	フォント 刑事 澄 蝶 包装 メス	溢れる 拒む 隠る しむ	柔らかい むずかしい きつい とんでもない
Low (10,000-30,000) minfr = 324 maxfr = 1930	N = 15076 ~クレイ~ 方角 近鉄 走り ~苑~ 人妻	N = 2114 駆け込む やせる 書き留める 滅ぶ	N = 299 むつかしい 仲良い ~くすい~ 腹立たしい	水槽 グローバリゼーション 青島 懇話 射精 モグラ	対する 振舞う 吸い上げる くぼむ	若々しい 面倒くさい 耐え難い 香ばしい

Slika 12: *Seznam naključno izbranih besed iz korpusa JpWaC (samostalnik vsebuje vezavo na* -sa *(N.Vs) in osnovo pridevnika na* -na [42]*).*

Slika 13 kaže kolokacijska gesla samostalnika *hyouka* »vrednotenje« (評価), pridobljena z orodjem SkE v oknu SketchEval. Vsako geslo je prikazano z dvajsetimi

42 To je v skladu z morfološkim orodjem *ChaSen* (http://chasen.naist.jp/stable/ipadic/, dostop 11.4.2015), ki je bil uporabljen za analizo korpusa JpWaC.

Rubric: **G** = Good; **Gb** = Good but wrong grammatical relation; **M** = Maybe (not striking collocate); **Ms** = Maybe (specialized vocab); **B** = Bad

GramRel	Collocation	Rating					Freq
		G	Gb	M	Ms	B	
modifier_Ai	高い	❍	❍	❍	❍	❍	3388
modifier_Ai	正しい	❍	❍	❍	❍	❍	208
modifier_Ana	多元的	❍	❍	❍	❍	❍	12
modifier_Ana	定性的	❍	❍	❍	❍	❍	12
modifier_Ana	正当	❍	❍	❍	❍	❍	107
modifier_Ana	適正	❍	❍	❍	❍	❍	68
modifier_Ana	厳正	❍	❍	❍	❍	❍	10
particle	に当たって	❍	❍	❍	❍	❍	73
prefix	再	❍	❍	❍	❍	❍	938
pronomの	読者	❍	❍	❍	❍	❍	253
pronomの	一定	❍	❍	❍	❍	❍	178
suffix	損	❍	❍	❍	❍	❍	134
suffix	額	❍	❍	❍	❍	❍	591
suffix	益	❍	❍	❍	❍	❍	62
suffix	替え	❍	❍	❍	❍	❍	43
にverb	値する	❍	❍	❍	❍	❍	104
のpronom	対象	❍	❍	❍	❍	❍	290
のpronom	実施	❍	❍	❍	❍	❍	136
はAdj	無効	❍	❍	❍	❍	❍	251
をverb	下す	❍	❍	❍	❍	❍	181

Slika 13: *Kolokacijska gesla za ovrednotenje samostalnika* hyouka *»vrednotenje« (SketchEval).*

kolokacijami, ovrednotili pa so jih strokovnjaki za japonski jezik. Tabeli 10 in 11 prikazujeta rezultate vrednotenja glede na ocenjevalca in glede na seznam visoko-, srednje- in nizkofrekvenčnih kolokacij.

Vrednotenje so opravili trije ljudje, pri čemer je bil rezultat dveh oseb podoben. Gesel, ki so ocenjena kot dobre kolokacije (*yoi kyouki*良い共起), je 83,40 % oz.

81,98 %. Gesel, ki so ocenjeni z oznako *morda* (morda bi bilo geslo dobro vključiti v slovar), je 9,69 % in 3,72 %. Pri tretjem ocenjevalcu je razlika znatna, saj je dobrih kolokacij za 38,93 %, potencialno dobrih (morda) pa 40,23 %. Razlog je lahko v presoji, katera gesla so primerna za slovar, predvsem pa je problematična kategorija *morda*. Med ocenjevanjem se lahko presoja le na osnovi prikazanega rezultata, mogoče pa je pogledati dejanske primere konkordanc in uporabiti druge funkcije orodja ter tako priti do bolj natančnih rezultatov. Možno je, da je ravno to vzrok za razliko pri oceni kolokacij. Na primer, beseda *shiharau* »plačati«（支払う）se sopojavlja z besedo *ryou* »plačilo«（料）, toda kolokacija kot takšna je nepopolna. S pregledom konkordanc lahko pridemo do relevantnih kombinacij z morfemom *ryou*, denimo *shiyouryou wo shiharau* »plačati najemnino« (使用料を支払う) in *jugyouryou wo shiharau* »plačati šolnino« (授業料を支払う).

Najvišji delež gesel, ki so bila ocenjena kot slabe kolokacije (*yokunai kyouki*良くない共起), je 17,74 %, najnižji pa 3,30 %.

Če gledamo z vidika visoko-, srednje- in nizkofrekvenčnih kolokacij, je delež visokofrekvenčnih kolokacij najvišji (76,37 %). Pri besedah s srednjo in nižjo frekvenco ni opaznejših razlik (65,78 % in 64,42 %).

Tabela 10: *Rezultat vrednotenja kolokacijskih gesel (s strani ocenjevalca).*

Izbira	Odgovor			
	Ocenjevalec A	Ocenjevalec B	Ocenjevalec C	Povprečje
Dobro	**83,40 %**	**81,98 %**	**38,93 %**	**68,10 %**
Dobro (z napačno slovnično zvezo)	2,27 %	0,46 %	2,50 %	1,74 %
Morda (ne povsem ustrezna kolokacija)	9,69 %	3,72 %	**40,23 %**	**17,88 %**
Morda (specializirano besedišče)	0,31 %	1,51 %	0,60 %	0,81 %
Slabo	3,30 %	12,33 %	17,74 %	**11,12 %**
Ni na voljo	1,03 %	0 %	0 %	0,34 %

Tabela 11: *Rezultati vrednotenja kolokacijskih gesel (od nizkofrekvenčnih do visoko frekvenčnih).*

Izbira	Odgovor		
	Visoka frekvenca	Srednja frekvenca	Nizka frekvenca
Dobro	**76,37 %**	**65,78 %**	**64,42 %**
Dobro (z napačno slovnično zvezo)	1,32 %	1,44 %	2,56 %
Morda (ne povsem ustrezna kolokacija)	**15,49 %**	**16,67 %**	**20,12 %**
Morda (specializirano besedišče)	1,10 %	0,67 %	0,58 %
Slabo	**4,84 %**	**15,22 %**	**12,32 %**
Ni na voljo	0,88 %	0,22 %	0,00 %

Rezultate raziskave lahko strnemo v naslednje točke.

1. Morda (ne povsem ustrezna kolokacija) se pogosto pojavi v naslednjih primerih:
- Različna pisava (pismenke in hiragana) (npr. *mezasu* »stremeti, ciljati« 目指す/めざす/目ざす)
- Nizkofrekvenčna kolokacija, npr. večkratna pojavitev zaradi ponovitve istega stavka v korpusu ali ponovitev iste kolokacije v okviru ene spletne strani
- Kolokacije besednih derivatov (npr. geslo *kansei* 完成 »zaključek, končanje, dovršitev« in kolokacija *gakkyoku no kanseisha* »oseba, ki konča skladbo« 楽曲の完成者)
- Lastna imena, imena krajev (*Ikego no Midori* 池子の緑, *Kintetsu no Nakamura/Yamaguchi* 近鉄の中村・山口)

2. Slaba kolokacija se pogosto pojavi v naslednjih primerih:
- Napaka pri morfološki analizi (npr. *yoroshikuu* »prosim (pozdrav)« よろしくぅ se analizira kot *yoroshii* よろしい »v redu, dobro« in *kui* くい »jedenje, ugriz«, posledično se kot kolokacija za *yoroshii* よろしい pokaže *kui* くい). Poleg tega je izraz *isogaba mawaru* 急がば回れ »počasi se daleč pride« analiziran kot *iso+ga+ba+mawareru* (急 + が + ば + 回れる) in se pojavi kot neobstoječa kolokacija **iso ga mawareru* *急が回れる.
- Ponovitev istih stavčnih primerov v korpusu.

Komentarje ocenjevalca lahko strnemo v naslednje točke.

- SkE bo zagotovo prišel prav pri sestavljanju slovarja, predvsem zato, ker lahko tako odkrijemo povezave, ki se jih ne zavedamo glede na našo intuicijo.
- Pri nekaterih primerih besed lahko čutimo vpliv aktualnih tem na spletu v času izdelave korpusa.
- Dobro bi bilo izboljšati orodje ChaSen ali imeti alternativno boljše orodje za morfološko analizo. Še posebno je problematično, da orodje pridevnike na *-na* in določene glagole prepozna kot samostalnike.
- Za dolge odseke in povedi bi bilo koristno orodje, ki bi lahko pravilno določilo, od kod do kod seže vezljivost pri atributivnih odnosih.
- Koristno bi bilo orodje za analizo, ki bi vzelo v obzir problematiko zapisovanja japonščine.
- Z vidika poučevanja japonščine je nujno analizirati jezikovne pojave ločeno po zvrsteh besedila: specifičnosti formalnih samostalnikov (npr. formalni samostalnik *uchi ni* »dokler, preden« うちに v stavku *wasurenai uchi ni kakitomeru* »zapišite, preden pozabite« 忘れないうちに書き留める), ter pisani ali govorni jezik (npr. pogovorni izraz *-tte* »pravi, je rekel« って).
- V nadaljevanju so navedeni štirje problematični primeri za »dobro, morda in slabo«. 1. nepopolnost (beseda in njena kolokacija ne tvorita popolne kolokacije, ker določen del manjka), 2. kolokacije so sicer slovnično pravilne, a pomensko šibke, 3. kolokacijske besede so sintaktično in slovnično predaleč, 4. lastna imena.

Ker je približno 68 % kolokacij ovrednotenih z oznako »dobro«, 18 % z »morda« ter 11 % s »slabo«, lahko sklepamo, da je japonska različica SkE učinkovito orodje za sestavljanje slovarjev. Toda glede na rezultate vrednotenja je izboljšava sistema nujna, še posebno na področju morfološke analize. Kilgarriff et al. (2010) podrobno popišejo potek in rezultate projekta SketchEval. Sodeč po prispevku je vrednotenje japonske različice dalo najboljše rezultate v primerjavi z različicami v drugih jezikih. Srdanović et al. (2011) podrobneje predstavijo rezultate japonskega vrednotenja ter prednosti in omejitve sistema SkE.

3.8 Povzetek in nadaljnja vprašanja

V pričujočem poglavju smo raziskovali možnost aplikacije orodja SkE na področje izobraževanja japonščine in raziskovanja jezika. Posebno smo se posvetili apliciranju na izdelovanje slovarjev za učence japonščine ter ovrednotili japonsko različico SkE. Rezultate smo primerjali z japonskim slovarjem kolokacij ter ovrednotili japonsko različico SkE kot pomoč pri izdelavi japonskega kolokacijskega slovarja v okviru večjezičnega projekta SketchEval. Glede na rezultate vrednotenja

je SkE učinkovito orodje za izdelavo kolokacijskega slovarja. Trenutno nimamo slovarja japonščine, ki bi pokrival raznovrstne kolokacijske povezave na podlagi uravnoteženega korpusa, zato je uporaba orodja SkE v kombinaciji s tovrstnim korpusom zelo zaželjena za izdelavo pomembnih učnih materialov. Obenem se je v japonski različici orodja pokazalo nekaj slabosti, pri katerih so nujne izboljšave. To velja predvsem za pomanjkljivosti, ki izhajajo iz preozke morfosintaktične analize besedil.

Kolokacijske povezave, ki jih lahko iščemo s tem orodjem, so bile izluščene iz velike količine podatkov, ki so jih tvorili naravni govorci, in tako odsevajo tipične kolokacije. Izziv za v prihodnje ostaja dodajanje mnenja informantov k pridobljenim rezultatom, npr. tako da verjetnost pojavitve kolokacij primerjamo z intuicijo naravnih govorcev. S tovrstnim vrednotenjem lahko s psihološkega vidika potrdimo zanesljivost kolokacijskih gesel. Njihovo zanesljivost je prav tako možno potrditi z uporabo in primerjanjem več različnih korpusov. Na primer, pred kratkim opravljena raziskava kolokacij pridevnikov in samostalnikov (Srdanović et al., 2013) s pomočjo korpusa BCCWJ (Maekawa et al., 2013) in obsežnega spletnega korpusa JpTenTen (Srdanović et al., 2012) kaže, da so ti korpusi zanesljivo gradivo za izdelavo slovarja.

Ker rezultati kolokacijskih povezav temeljijo na morfološki analizi orodja ChaSen in ker je treba odpraviti njegove šibke točke, je nujno idiomom ter prepodrobno razčlenjenim morfemom ponovno dodati oznako za eno enoto. Izziv za v prihodnje predstavlja tudi primerjava rezultatov orodja ChaSen z rezultati, ki temeljijo na drugih orodjih, kot so UniDic in MeCab, ter raziskave aplikacij drugih orodij za morfološko analizo. Slovar UniDic (Den et al., 2007), ki je nastal po zaključku analize korpusa BCCWJ, je do določene mere prispeval k izboljšanju morfološke analize korpusov japonskega jezika in je kot takšen uporabljen pri analizi novega spletnega korpusa JpTenTen (Srdanović et al., 2013).

Če upoštevamo različne možnosti uporabe pri poučevanju japonščine, je za izboljšavo japonske različice sistema zaželeno, da sistem prikaže kokolacije in stavčne primere glede na stopnjo znanja japonščine ter dopiše furigano. Še eden od izzivov, ki ostaja za prihodnost, je vnos drugih korpusov v sistem in možnost iskanja glede na zvrst besedila.

4 Kolokacijski odnos na daljavo med prislovi in modalno obliko na koncu stavka

4.1 Uvod

Kolokacija na daljavo je pojav, pri katerem se dve besedi ali beseda in določeni stavčni element istočasno nahajata na določeni razdalji v besedišču. Raziskave o tem pojavu v japonščini in drugih jezikih so omejene in jim je v jezikoslovju namenjeno premalo pozornosti. V svetu korpusnega jezikoslovja je že v času razmaha zanimanja za raziskave kolokacijskih odnosov postalo splošno sprejeto, da je razdalja med kolokacijami največ pet besed (Sinclair, 1991). V tradicionalni japonščini se odnos med prislovom in modalno obliko na koncu stavka kaže kot ujemalni odnos, s strani korpusnega jezikoslovja pa je lahko obravnavan kot kolokacija. Glede na to, da se ta odnos pojavlja tudi na daljši razdalji med prislovom in modalno obliko, ga lahko imenujemo kolokacija na daljavo.

Primer 4-1:

Tabun ashita ame ga furu deshou.

»Verjetno bo jutri padal dež.«

（たぶん明日雨が降るでしょう。）

V primeru 4-1 se izraz *tabun* »verjetno« (たぶん) imenuje povedni prislov, saj modificira povedni način in ima ustaljeno ujemanje s povedjo (Sakakura, 1988). Modalnost uravnava pomen govornega dejanja in tako v osnovi izraža govorčev način izrekanja v času govora (Moriyama et al., 2000). Na primer, *deshou* »verjetno« (でしょう) na koncu povedi v zgornjem stavku izraža modalnost prepričanja. Odnos med *tabun* in *deshou*, kot se pojavlja v omenjenem primeru, v japonski slovnici velja za ujemalni odnos. Ujemanju povednih prislovov in modalnosti je bilo v dosedanjih raziskavah namenjeno že kar nekaj pozornosti (Minami, 1974; Kudō, 2000; Bekeš, 2006). V pričujoči raziskavi bomo ta odnos obravnavali kot kolokacijo s stališča korpusnega jezikoslovja in ga podrobneje raziskali.

Glede na raziskavo analize japonskih pogovorov (Bekeš, 2006) se kolokacije povednih prislovov in modalnih oblik na koncu stavka, v primerjavi z le modalno obliko v stavku, pojavijo v približno enem od desetih primerov. S stališča jezikovne teorije je pojav kolokacij v takem razmerju zaznamovan (Halliday, 1991), ima pa pomembno mesto pri izražanju modalnosti. Gledano s stališča povednih prislovov

pa je primerov kolokacij med povednimi prislovi in modalno obliko na koncu stavka precej veliko, tako da lahko tudi ta pojav sistematično opazujemo.

Kudō (2000) je s pomočjo korpusov raziskoval ujemalni odnos med povednimi prislovi in modalnostjo. Razkril je stopnjo vezave med kolokacijami različnih prislovov in določenih modalnih oblik, ter najmočnejše vezi vzel za ujemalni in kolokacijski odnos. Kudō pravi, da lahko povedne prislove razdelimo v štiri skupine, in sicer nujnost, pričakovanje, domneva in možnost, stopnja prepričanja pa se kaže v tem zaporedju. Ujemalni odnos med povednimi prislovi in modalnostjo pri korpusih lahko razumemo kot statistično pomembno tendenco kolokacij.

V tem poglavju raziščemo razpršenost povednih prislovov, kakor tudi kolokacijsko razmerje med povednimi prislovi in modalnostjo, na podlagi različnih korpusov. Najprej preverimo, kako se spreminja razpršenost prislovov glede na korpus. Zatem raziščemo, ali se kolokacijski odnos povednih prislovov in modalnosti pojavlja z močno stopnjo vezave, kateri prislov se tipično veže s katero modalno obliko in s katerim modalnim tipom ter ali se tendenca kolokacij spreminja glede na vrsto korpusa. V pričujočem segmentu obenem izpostavimo pomembnost kolokacijskih odnosov na daljavo.

4.2 Razpršenost prislovov in značilnosti korpusov

V tem razdelku primerjamo razpršenost povednih prislovov v trinajstih različnih korpusih in ugotavljamo, kako na to vplivajo značilnosti posameznega korpusa. Pridobljeni podatki so s pomočjo metode razvrščanja (ang. *cluster analysis*) razdeljeni v skupine. Nadalje so z entropijo izračunana tudi odstopanja v pojavnosti povednih prislovov.

Uporabljeni so naslednji (pod)korpusi (podrobneje so predstavljeni v Tabeli 1):

(1) Vladna poročila, uradne bele knjige (KokkenOW), del BCCWJ-evega mini korpusa, (2) Članki s področja računalniške obdelave naravnih jezikov (NLP), (3) 16 japonskih naravoslovnih učbenikov za študente (16K), (4) Korpus neformalnih razgovorov (NUJCC), (5) Korpus formalnih razgovorov (Oikawa), (6) Spletni korpus Yahoo! Chiebukuro (KokkenOC), (7) Veliki japonski spletni korpus (JpWaC), (8) Korpus publikacij (KokkenBK), del BCCWJ-evega mini korpusa, (9) Enoletni podatki časopisa Mainichi shimbun iz leta 2002 (Mai2002), (10) Učbeniki japonskega jezika za osnovno šolo (KokugoK), (11) Učbeniki za srednjo šolo (KokkenK), del BCCWJ-evega mini korpusa, (12) Učbeniki japonskega jezika, priključeni h KokkenK (KKK), (13) Mešani tipi besedil časopisov, starejše literature itn. (Kudō).

Predmet raziskave so naslednji povedni prislovi: *angai* »precej, bolj kot sem pričakoval« (あんがい), *aruiwa* »morda, ali pa« (あるいは), *doumo* »precej, bolj ali manj« (どうも), *douyara* »verjetno, potem takem« (どうやら), *hyotto shitara/hyotto suru to* »morda, obstaja le možnost« (ひょっとしたら/ひょっとすると), *kanarazu(shimo)* »gotovo, ne vedno + neg.« (かならず[しも]), *kitto* »gotovo, nedvomno, absolutno, v vsakem primeru« (きっと), *koto ni yoru to/koto ni yoreba* »morda, lahko, obstaja možnost« (ことによると/ことによれば), *moshika shitara/moshika suru to/moshika sureba* »morda« (もしかしたら/もしかすると/もしかすれば), *ookata* »večinoma« (おおかた), *osoraku* »verjetno« (おそらく), *sazo* »gotovo« (さぞ), *tabun* »verjetno« (たぶん), *taigai* »verjetno, v večini primerov« (たいがい), *taitei* »običajno« (たいてい), *zettai(ni)* »absolutno« (ぜったい[に]), in *yohodo/yoppodo* »precej« (よほど/よっぽど).

4.2.1 Porazdelitev korpusov

Tabela 12 prikazuje relativno frekvenco razpršenosti povednih prislovov glede na posamezni korpus oz. podkorpus v odstotkih. V primeru, ko se določeni prislovi pojavijo relativno velikokrat, se to zazna kot podatki z odstopanjem. Večja odstopanja so opazna pri naslednjih štirih korpusih: uradne bele knjige (KokkenOW), članki s področja računalniške obdelave naravnih jezikov (NLP), korpus neformalnih razgovorov (NUJCC) ter 16 naravoslovnih učbenikov (16K). Na drugi strani najbolj uravnoteženo razpršenost prislovov izkazujeta korpus publikacij (KokkenBK) ter veliki japonski spletni korpus (JpWaC), kjer podatkov z večjim odstopanjem ni.[43]

43 Ali je celoten korpus brez odstopanja, lahko preverimo tudi z analizo podatkov, ki niso prislovi. BCCWJ-ev korpus publikacij ter spletni korpus JpWaC sta bila pripravljena zato, da bi vsebovala uravnotežene podatke (s čim manj odstopanj). Rezultati primerjave korpusa JpWaC s korpusom časopisov pa kažejo, da je pri časopisih odstopanja več (več o tem v razdelku 2.3).

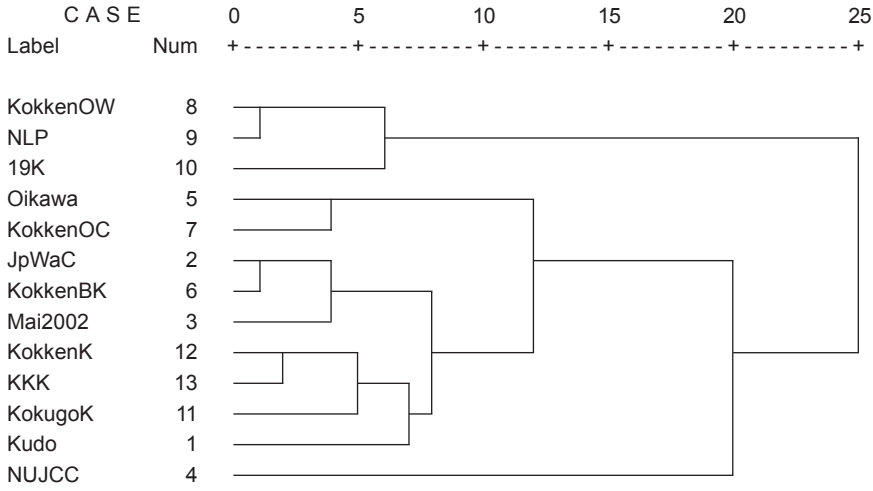
Tabela 12: *Razpršenost povednih prislovov v različnih korpusih.*[44]

PRISLOV / KORPUS	KokkenOW	NLP	16K	NUJCC	Oikawa	KokkenOC	JpWaC	KokkenBK	Mai2002	KokugoK	KokkenK	KKK	Kudô
kanarazu	5 %	23 %	42 %	7 %	14 %	4 %	8 %	15 %	25 %	12 %	28 %	16 %	4 %
zettai	2 %			52 %		14 %	9 %	6 %	11 %	3 %	2 %	4 %	5 %
zettai ni	2 %		4 %			11 %	6 %	8 %	12 %	3 %	9 %	2 %	
kanarazushimo	**84 %**	**66 %**	39 %	1 %	5 %	2 %	6 %	6 %	8 %	0 %	10 %	6 %	
yohodo	0 %				1 %	2 %	2 %	3 %	2 %	3 %	1 %	2 %	4 %
yoppodo				2 %		2 %	1 %	1 %	1 %	1 %			
taigai				2 %	8 %	1 %	1 %	1 %	0 %				1 %
taitei	1 %	6 %	4 %	1 %		5 %	4 %	2 %	3 %	4 %	6 %	12 %	1 %
kitto		3 %		15 %	8 %	15 %	12 %	14 %	10 %	38 %	26 %	26 %	28 %
ookata					1 %	0 %	0 %	1 %	0 %	1 %	1 %	2 %	3 %
osoraku	1 %	3 %	7 %	1 %	8 %	1 %	13 %	12 %	9 %	2 %	5 %	10 %	19 %
sazo						0 %	0 %	1 %	1 %	4 %	1 %	2 %	5 %
tabun	2 %		3 %	3 %	39 %	26 %	16 %	11 %	6 %	3 %	4 %	8 %	10 %
doumo	0 %		1 %	6 %	7 %	6 %	8 %	7 %	5 %	15 %	2 %	4 %	5 %
douyara				2 %		3 %	5 %	5 %	3 %	3 %			5 %
angai	0 %		1 %	3 %	1 %	0 %	2 %	1 %	1 %	1 %	1 %		2 %
hyotto shitara				1 %	1 %	1 %	1 %	1 %	1 %	1 %	1 %	2 %	3 %
hyotto suru to							0 %						
koto ni yoreba						0 %	0 %	0 %	0 %				
koto ni yoru to	3 %		1 %				0 %	0 %	1 %				1 %
moshi kashitara				5 %	8 %	5 %	3 %	3 %	1 %	2 %	2 %	2 %	5 %
moshika sureba						1 %	0 %						
moshika suru to							1 %	1 %	0 %	1 %	1 %	2 %	
	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %

44 Besedi *yohodo*, *yoppodo* sta bili analizirani kot identična beseda v korpusih Kudô in Oikawa. *Zettai* in *zettai ni* sta bili analizirani kot identična beseda v korpusih Kudô in NUJCC. *Moshikashitara* in *moshikasuruto* sta bili analizirani kot identična beseda v korpusu NUJCC. Čeprav je pri poskusni analizi korpusa JpWac *doumo* dobil 37 %, *ookata* 44 % in *zettai* 17 %, so bile iz teh rezultatov izključene besede, ki ne sodijo med prislove.

4.2.2 Razvrščanje korpusov v skupine

Na podlagi razpršenosti prislovov v korpusih smo poskusili razdeliti korpuse s pomočjo metode razvrščanja v skupine (Slika 14). Analiza združi korpuse s skupnimi točkami v clusterje oz. skupine.



Slika 14: *Razvrščanje 13 korpusov glede na razpršenost prislovov.*

Rezultati razvrščanja v skupine se skladajo z relativno frekvenco razpršenosti prislovov, prikazano v Tabeli 12. Prvo skupino sestavljajo korpusi s specifičnim namenom, to so uradne bele knjige (KokkenOW), članki o računalniški obdelavi naravnih jezikov (NLP) ter naravoslovni učbeniki (16K). Podobnost med KokkenOW in NLP je bila še največja, kar kaže na njihovo zelo specifično naravo. Tudi korpus formalnih razgovorov (Oikawa) ter korpus Chiebukuro (KokkenOC) tvorita skupino, in sicer kot korpusa s formalnim slogom pogovora. Nadalje spadata h korpusu z uravnoteženimi podatki spletni korpus (JpWaC) in korpus publikacij (KokkenBK), dodan pa je še korpus časopisov (Mai2002), kar predstavlja skupino raznolikih pisnih korpusov. Podatki Kudō spadajo v isto skupino z učbeniki, na podlagi tega pa lahko sklepamo, da so najbolj podobni tistim iz učbenikov. Nazadnje imamo še korpus neformalnih razgovorov (NUJCC), ki se v precejšnji meri razlikuje od drugih. Dobljeni podatki potrjujejo veljavnost metode, s katero so bili korpusi razdeljeni na različne tipe korpusov.

V nadaljevanju sledi tudi analiza, kako lahko tako razvrščene korpuse uporabimo pri poučevanju japonščine za tuje učence.

4.2.3 Odstopanja v razpršenosti prislovov v korpusu glede na vrednost entropije

Če odstopanja v razpršenosti povednih prislovov izračunamo z entropijo,[45] so rezultati podobni tistim v Tabeli 12. Bolj uravnoteženo je pojavljanje besed v korpusu, višja je entropija; bolj so besede v korpusih specifične in bolj je vidno odstopanje pri razpršenosti prislovov v korpusu, nižja je entropija.

Verjetnost pojavljanja določenega povednega prislova je označena s $P(A_j)$, entropija določenega korpusa pa je sestavljena iz sledeče formule.

$$I=-\sum jP(A_j)\log(1/P(A_j))$$

Če se pojavi le en povedni prislov, se entropija tega korpusa približa vrednosti 0. Rezultati izračuna so v Tabeli 13 (v raziskavi smo za najnižjo vzeli vrednost 2). Vrednost je najvišja v korpusu publikacij (KokkenBK), sledi spletni korpus (JpWaC), najnižja pa je v korpusu uradnih belih knjig (KokkenOW). Gledano s strani informacijske entropije, je ta v BCCWJ-jevih publikacijah in spletnih korpusih visoka in nima odstopanja podatkov, medtem ko ga podatki uradnih belih knjig imajo.

Tabela 13: *Razpršenost prislovov v korpusu glede na vrednost entropije.*

Korpus	Entropija	
KokkenBK	3,714070283	uravnotežen
JpWaC	3,700987843	
Mai2002	3,498045606	
KokkenOC	3,343810596	
KKK	3,324255803	
Kudō	3,32284281	
KokugoK	3,13598565	neuravnotežen
KokkenK	3,070358948	
Oikawa	2,816762276	
NUJCC	2,485516638	
16K	1,961601061	
NLP	1,413799449	
KokkenOW	1,072310389	

45 Entropija se uporablja kot mera za izračun odstopanja v razpršenosti določenih enot.

O primerjavi med korpusom publikacij (KokkenBK) in spletnim korpusom JpWaC, ki imata tendenco uravnoteženosti oz. podatkov brez večjega odstopanja, je več napisanega v razdelku 5.3.5. Opisane so skupne točke in razlike v kolokacijskih odnosih med povednimi prislovi in modalnostjo na koncu stavka v omenjenih korpusih.

4.3 Tendenca kolokacijskih odnosov prislovov in modalne oblike na koncu stavka

To poglavje razišče tendenco kolokacij prislovov in modalne oblike na koncu stavka v posameznem korpusu ter stopnjo razlikovanja teh tendenc glede na tip korpusa. Poleg frekvence, s katero se pojavljajo kolokacije, razišče tudi, ali jih lahko s pomočjo metode razvrščanja v skupine razdelimo glede na tendenco kolokacij prislovov in modalne oblike na koncu stavka.

4.3.1 Vrste kolokacij med prislovi in modalnostjo ter predmet raziskave

Poznamo nekaj različnih vrst kolokacij med prislovi in modalno obliko na koncu stavka. V pričujoči raziskavi so analizirane naslednje vrste.

- *Tabun, issho ni kita deshou ne.* [A-P-M]
  »Verjetno sta prišla skupaj, kajne.« (たぶん、一緒にきたでしょうね。)
- *Tabun, daijoubu* [A-P-Ø]
  »Verjetno je v redu.« (…たぶん、大丈夫)

[A-P-M] označuje primer, ko se v stavku pojavijo prislov (A), predikat (P) in modalnost (M). [A-P-Ø] pa označuje primer, ko se v stavku pojavita prislov (A) ter predikat (P), predikat pa ni zaznamovan z dodatno modalnostjo (Ø).

Obstajajo tudi druge vrste stavčnih vzorcev, s katerimi pa se pričujoča raziskava ne ukvarja. Na primer, ko se modalnost pojavi pred prislovom [P-M-A], ko prislova sploh ni [Ø-P-M], ko imamo samo prislov [A] ali pa samo modalnost [M] (primeri so iz korpusa NUJCC v Bekeš, 2006).

- *Kare mo issho ni kita deshou ne, tabun* [P-M-A]
  »Tudi on je morda prišel z njimi, verjetno.« (彼もいっしょに来たでしょうね、たぶん)
- *Juuhachi de wa nai n' janai ka...* [Ø-P-M]
  »Kaj nisi še osemnajst (let star) ...?« (十八ではないんじゃないか…)
- *Tabun, ne* [A]
  »Verjetno, ne« (たぶん、ね)

- *Deshou, ne* [M]
  »Morda, ja« (でしょうね)

4.3.2 Primerjava med Kudōjevimi podatki in spletnim korpusom

To poglavje naprej predstavi rezultate korpusne analize, ki jo je opravil Kudō (2000), in dobljene podatke razvrsti s pomočjo metode razvrščanja v skupine. Nato prikaže rezultate analize spletnega korpusa, ki je bila opravljena v sklopu pričujoče raziskave ter primerja podatke s Kudōjevimi.

4.3.2.1 Kolokacijski odnos med prislovi in modalno obliko na koncu stavka pri Kudōju

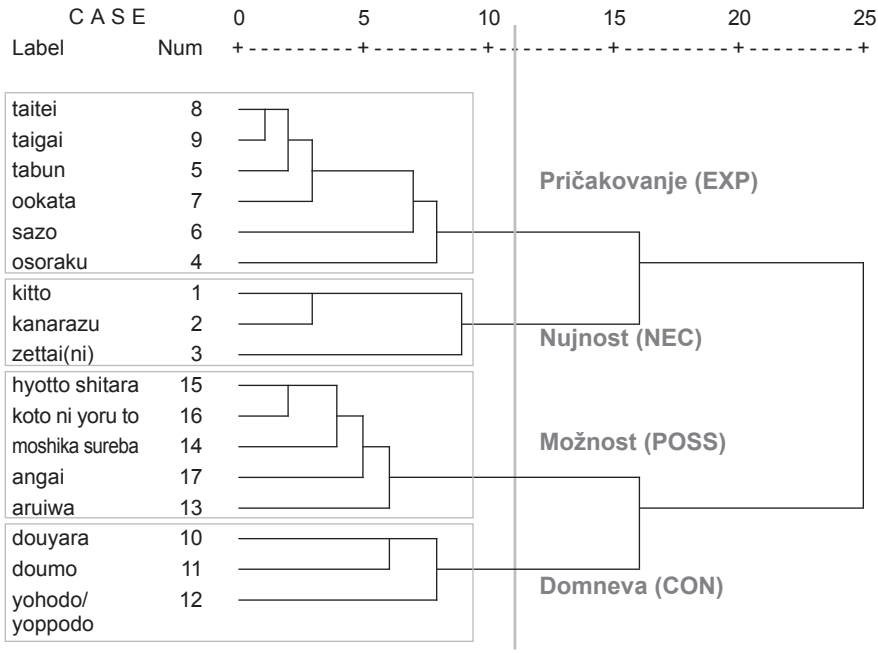
Korpus, ki ga uporablja Kudō, je sestavljen iz mešanih tipov besedil, med katerimi so časopisne vsebine, moderna japonska literatura (romani, ki so starejši več kot 50 let) itn. Kot že omenjeno, je njegova analiza korpusa pokazala, da imajo posamezni prislovi večjo tendenco vezave z določeno vrsto modalnosti. Tabela 14 prikazuje štiri skupine povednih prislovov (prislovi nujnosti, pričakovanja, domneve in možnosti, s stopnjo prepričanja v tem zaporedju) ter določeno vrsto modalnosti, s katero ima vsaka skupina večjo tendenco sopojavljanja.

Če Kudōjeve rezultate analize razvrstimo s pomočjo metode razvrščanja v skupine, dobimo podobne rezultate (Slika 15). Tako pri rezultatih analize, prikazane v grafu, kot pri rezultatih grupiranja, imajo *kitto* »gotovo, nedvomno, absolutno, v vsakem primeru«, *kanarazu* »gotovo« in *zettai(ni)* »absolutno« visoko tendenco sopojavljanja z vrsto modalnosti, ki izraža nujnost (*suru no da*, *ni chigai nai* itn.), zato tvorijo eno skupino. Nadalje so *osoraku* »verjetno«, *tabun* »verjetno«, *sazo* »gotovo«, *ookata* »večinoma«, *taitei* »običajno«, *taigai* »verjetno, v večini primerov« prislovi, ki se pogosto sopojavljajo z modalno obliko za pričakovanje (*to omowareru*, *no de wa nai darou* itn.). V skupino, v kateri modalnost izraža domnevanje, spadajo *douyara* »verjetno«, *doumo* »precej, bolj ali manj« ter *yohodo* »precej«. Nazadnje se v skupini, kjer modalnost izraža možnost, opazi pojavljanje prislovov *aruiwa* »morda«, *moshika shitara* »morda«, *hyotto shitara* »morda«, *koto ni yoru to* »morda« ter *angai* »precej, bolj kot sem pričakoval«.

Tabela 14: *Analiza prislovov in modalnosti na koncu stavka pri Kudōju (2000).*

PRISLOV	*suru no da* (NEC)	*ni chigainai*(NEC)	*ni kimatteiru*(NEC)	*hazu da(NEC)*	*darou/mai*(EXP)	*to omowareru(EXP)*	*no dewa nai darou ka*(EXP)	*rashii(CON)*	*to mieru* (CON)	*you da/mitai da*(CON)	*shisou da*(CON)	*kamoshirenai*(POSS)	*darouka* (POSS)	*senu tomo kagiranu* (POSS)	*suru fushi ga aru* (POSS)	SKUPNO	OSTALO (ni modalnost)
kitto	**139**	**38**	**8**	**3**	66	12				1	4	8				279	85
kanarazu	**17**	**5**	**2**	**1**	11											36	146
zettai(ni)	**48**															48	38
osoraku	31	18		1	**112**	**5**	**10**	2		1		2				182	--
tabun	19	1		2	**74**		**1**	1			2	3				103	--
sazo					**52**		**1**				1					54	--
ookata	2	1			**24**		**1**									28	13
taitei	3			1	**7**											11	80
taigai	2				**4**											6	33
douyara	5						1	**29**		**10**					1	46	39
doumo	13	1					6	**24**				1				45	385
yohodo	6	2			7		2	**12**	**9**	**3**			2			43	150
aruiwa					3	2	4					**53**	**3**	**1**		66	69
moshika sureba	2			1	1	1	11					**30**				46	--
hyotto shitara	2						7					**16**	**1**			26	--
koto ni yoru to	1						4					**7**	**1**	**1**		14	--
angai		1			1		3	1			1	**8**				15	81

Legenda (štiri vrste modalnosti): *NEC – nujnost, EXP – pričakovanje, POSS – možnost, CON – domneva.*



Slika 15: *Podatki raziskave Kudōja (2000) razvrščeni v skupine.*

4.3.2.2 Kolokacije povednih prislovov in modalnih oblik na koncu stavka v spletnem korpusu

V tem poglavju analiziramo 100 primerov posameznih prislovov iz spletnega korpusa JpWaC (rezultati so prikazani v Tabeli 15). Tudi pri analizi spletnega korpusa opazujemo štiri skupine prislovov kot tudi štiri skupine vrst modalnosti. Pridobljeni podatki so razdeljeni s pomočjo metode razvrščanja v skupine glede na tendenco sopojavljanja prislovov z vrsto modalnosti (Slika 16). Pridobljeni rezultati sovpadajo s Kudōjevimi podatki. Poleg tega pogledamo tudi razlike, ki se pojavijo pri tendencah kolokacij med prislovi in modalnimi oblikami.

Pri analizi podatkov s spleta se je modalna oblika pojavila skupaj s prislovom ne le v istem odstavku, temveč pogosto tudi v naslednjem. Prav tako se modalna oblika ni vedno pojavila samostojno, saj je bilo nekaj primerov, ko se je pojavila v nizu z drugo modalno obliko (primera 4-2 in 4-3).

Primer 4-2:

> *Tabun, sore-tte ima ni hajimatta koto dewa nai no da to omoimasu ga, saikin wa toku ni, »rikai fukanou« na jiken ga fueta youna ki ga shimasu.* (JpWaC)

»Mislim, da to verjetno ni nekaj, kar se je pričelo šele sedaj; zdi pa se mi, da je še posebno v zadnjem času 'nerazumljivih' dogodkov vse več.«

たぶん 、 それ って 今 に 始まっ た こと で は ない の だ と 思い ます が 、 最近 は 特に 、 〃 理解 不可能 〃 な 事件 が 増え た よう な 気 が し ます 。

Primer 4-3:

Eguchi seifu iin go shiteki no ten wa, osoraku Toukyou-to no chousa ni narareta bun nado wo kiso to shite osshatte oru no dewa nai ka to suitei shite oru wake de gozaimasu ga, ... (JpWaC)

»Ocenjuje se, da to, na kar opozarja član vlade Eguchi, gotovo temelji na raziskavi tokijske uprave in na drugih dokumentih ...«

江口 政府 委員 御 指摘 の 点 は 、 おそらく 東京 都 の 調査 に なら れ た 分 等 を 基礎 として おっしゃっ て おる のでは ない か と 推定 し て おる わけ で ござい ます が 、 ・・・

Tabela 15: *Kolokacije prislovov in modalne oblike na koncu stavka v JpWaC.*

Prislov/ Modalna oblika	NEC_no_da	NEC_to_iu_koto_da/mono_da	NEC_koto_da	NEC_koto_ni_naru/to_naru	NEC_mono_da	NEC_nakereba_naranai	NEC_ni_chigai_nai	NEC_hazu_da	NEC_hitsuyou_ga_aru/nai	NEC_beki_da/darou	NEC_to_wa_kagiranai	NEC_wake_da	NEC_wake_de_wa_nai	EXP_to_suru/sareru	EXP_darou/deshou/mai	EXP_no_darou	EXP_koto_to_omou/zonjiru	EXP_de_wa_nai_ka	EXP_to_omou	EXP_no_de_wa_nai_ka	EXP_no_darou_ka	EXP_no_de_wa_nai
kanarazu	4	0	2	1	4	6	0	2	1	2	0	0	1	0	2	0	0	0	4	0	0	0
kanarazushimo	3	3	6	0	4	0	0	0	2	0	6	1	9	0	1	0	0	0	4	0	0	3
zettai ni	3	0	1	0	0	4	0	2	13	1	0	0	2	0	1	0	0	0	12	0	0	0
zettai	4	0	2	0	1	3	0	2	2	2	0	1	0	0	1	1	0	0	0	2	0	0
taigai	14	0	2	0	3	1	0	0	0	0	0	1	0	0	2	0	0	0	4	1	1	0
taitei	0	1	4	3	2	1	1	0	0	0	0	3	2	2	0	0	0	0	5	0	0	0
aruiwa	0	1	0	0	0	0	0	1	0	2	0	2	0	0	1	1	0	0	0	0	0	0
yohodo	5	0	0	1	3	3	1	2	1	0	0	2	0	0	7	5	0	1	11	0	0	0
tabun	2	0	0	0	1	0	0	4	0	0	0	0	0	0	15	15	0	4	15	1	0	0
kitto	4	0	0	0	2	0	1	2	0	0	0	0	0	0	23	11	0	0	14	4	0	0
osoraku	2	0	0	0	0	0	1	0	0	0	0	1	0	0	28	17	0	5	14	8	0	0
sazo	0	0	0	0	0	0	2	1	0	0	0	0	0	0	36	44	9	1	0	1	0	1
ookata	4	0	0	0	0	0	0	0	0	0	0	0	0	0	7	5	0	0	3	2	0	0
doumo	3	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	0	0	3	2	0	0
douyara	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
angai	6	1	0	0	4	0	0	0	0	0	0	0	0	0	1	1	0	5	0	2	0	0
hyotto shitara	2	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	1	1	14	0	0
hyotto suru to	1	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	10	2	0
koto ni yoreba/yoru to	5	0	2	0	0	0	0	0	0	0	0	1	0	1	2	1	0	1	2	11	0	0
moshika suru to	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	1	10	1	0
moshika shitara	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	1	5	1	0	1

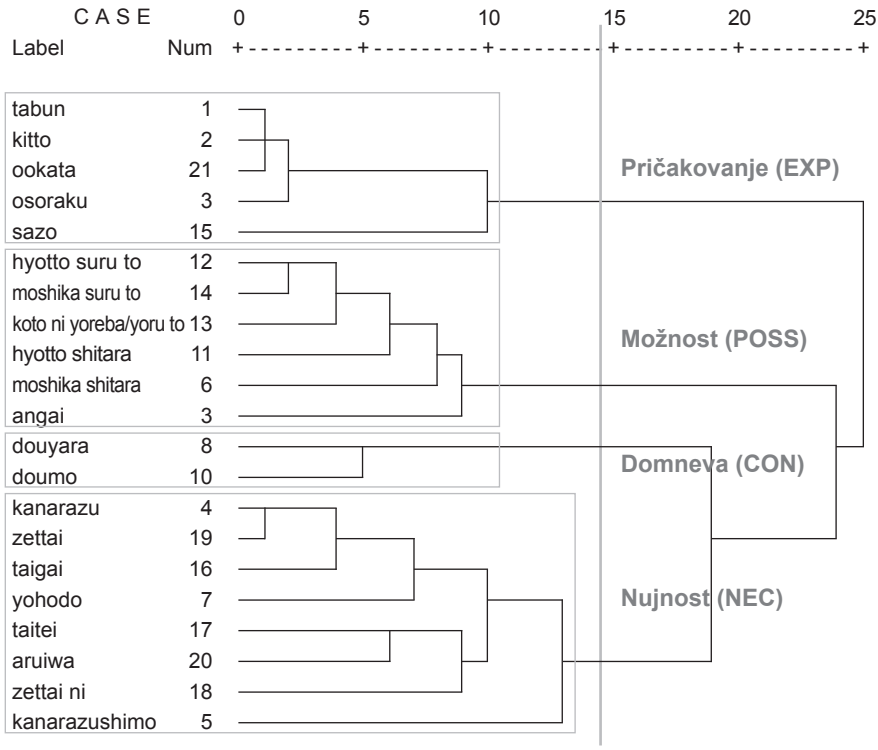
Prislov/ Modalna oblika	CON_you/mitai_da	CON_rashii/poi	CON_sou_da	CON_sou_ni_nai	CON_ni/to_mieru	POSS_koto_mo_aru	POSS_kamoshirenai	POSS_no_kamoshirenai	POSS_kana	POSS_kanousei_ga_aru/nai	POSS_darou_ka	POSS_ienai	POSS_to_wa_kagiranai	POSS_no ka?/kana/kashira	UNM_da	UNM_suru	UNM_de wa nai	UNM_natte_inai/iru	UNM_ni_naru/suru	UNM_da ka?	UNM_shinai	UNM_suru ka?	OTHER
kanarazu	0	1	0	0	0	3	0	0	0	0	0	0	0	0	5	39	0	1	0	0	0	0	0
kanarazushimo	5	0	2	0	0	1	2	0	0	0	2	8	6	0	0	0	11	1	3	0	18	0	3
zettai ni	0	0	2	0	0	3	0	0	0	0	0	2	0	0	5	7	3	0	1	0	15	0	13
zettai	0	0	1	0	0	0	0	0	0	0	0	0	0	0	9	10	2	0	0	0	11	0	18
taigai	3	1	1	0	0	2	2	0	0	0	2	0	0	0	16	26	1	0	1	0	3	0	0
taitei	2	0	0	0	0	5	0	0	0	1	0	0	0	0	15	40	3	1	0	0	3	0	0
aruiwa	0	0	0	0	0	0	4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
yohodo	6	0	2	1	4	0	0	0	0	0	2	0	0	4	15	15	1	0	0	1	4	1	0
tabun	3	0	0	0	0	0	2	0	1	0	1	0	0	0	4	14	0	0	0	1	3	0	0
kitto	0	0	0	0	0	0	3	0	0	0	2	0	0	0	9	10	0	0	0	0	3	0	0
osoraku	3	1	0	0	0	0	5	0	0	0	0	0	0	1	2	9	0	0	0	0	1	0	0
sazo	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
ookata	1	1	0	0	0	0	0	0	2	0	0	0	0	0	7	22	0	1	1	0	0	0	0
doumo	17	5	2	0	0	1	0	0	1	0	0	0	0	1	5	10	1	0	0	0	4	1	0
douyara	41	30	6	1	0	0	1	0	0	0	0	0	0	0	8	8	0	0	0	1	0	0	0
angai	5	0	1	0	0	2	20	4	0	0	0	0	0	2	24	15	0	0	0	0	2	0	0
hyotto shitara	1	7	0	0	0	0	52	3	1	2	0	0	0	1	1	1	0	0	0	1	1	0	0
hyotto suru to	0	0	0	0	0	0	45	17	5	0	0	0	0	4	2	3	0	0	1	3	1	0	0
koto ni yoreba/yoru to	1	1	0	0	0	3	34	19	0	0	0	1	0	1	1	9	0	0	0	0	0	0	1
moshika suru to	3	0	0	0	0	0	43	16	0	5	2	0	0	5	1	0	1	0	0	3	0	0	0
moshika shitara	0	0	1	0	0	0	55	14	2	0	4	0	0	0	4	1	0	0	0	3	0	0	0

Legenda: *NEC – nujnost, EXP – pričakovanje, POSS – možnost, CON – domneva.*

4.3.2.3 Primerjava Kudōjevih in spletnih podatkov

Z analizo prislovov in modalnosti na koncu stavka pri spletnih podatkih lahko potrdimo Kudōjev predlagani sistem. Tudi pri spletnih podatkih ima posamezen prislov tendenco močne vezi z določeno vrsto modalnosti in opazimo lahko štiri omenjene skupine povednih prislovov. V obeh korpusih lahko najdemo tudi določene razlike.

Prvič, pri spletnih podatkih je veliko različnih vrst modalnosti. S Kudōjevo metodo se jih lahko v spletnem korpusu razvrsti na 40 različnih vrst. Za primerjavo: Kudōjevi podatki imajo 15 različnih vrst modalnosti. Rečemo lahko, da je pri spletnih podatkih vrst modalnih oblik več kot pri Kudōju zaradi raznolikosti podatkov. Znaten del spletnih podatkov je bil pridobljen z blogov in podobnih strani, kjer najdemo več primerov prvoosebne rabe. Ker modalnost izraža govorčevo vedênje, je pričakovan večji obseg ter več oblik modalnosti pri spletnih podatkih, v primerjavi s podatki iz časopisov in romanov.



Slika 16: *Spletni podatki razvrščeni v skupine.*

Pri analizi obeh, Kudōjevih in spletnih podatkov, tendence sopojavljanja s podobnimi modalnimi oblikami kažejo naslednji prislovi: *kanarazu* »gotovo«, *zettai(ni)* »absolutno«, *tabun* »verjetno«, *osoraku* »verjetno«, *sazo* »gotovo«, *ookata* »večinoma«, *doumo* »precej, bolj ali manj«, *douyara* »verjetno, potem takem«, *hyotto shitara* »morda, obstaja le možnost«, *koto ni yoreba* »morda, lahko, obstaja možnost« ter *moshika shitara* »morda«. Na drugi strani, pa prislovi *taigai* »verjetno, v večini primerov«, *kitto* »gotovo, nedvomno, absolutno, v vsem primeru«, *yohodo* »precej«, *taitei* »običajno« ter *aruiwa* »morda« kažejo tendenco tvorjenja kolokacij z drugimi modalnimi oblikami. Tendenca kolokacije prislova *taigai* se glede na Kudōjeve podatke pojavlja z obliko pričakovanja, pri spletnih podatkih pa prestopi k obliki nujnosti. Ravno obratno pa je s prislovom *kitto*. Prislova *yohodo* in *taitei* pogosto kažeta tendenco kolokacije z nezaznamovano modalnostjo (ang. *unmarked modality*), medtem ko se *aruiwa* veže v paru z drugimi prislovi. Iz teh razlogov so rezultati kategorizacije zgornjih treh prislovov vprašljivi.

Pri frekvencah kolokacij s posamezno modalno obliko obstajajo razlike. Razlog je v tem, da je pri Kudōjevih podatkih uporabljen formalni knjižni način, medtem ko ta pri spletnih podatkih variira. Prav tako Kudōjevi podatki vsebujejo tudi nekoliko starejšo japonščino, kar lahko zaradi pomenskih sprememb in sprememb v rabi skozi čas do določene mere vpliva na rezultate.

4.3.3 Primerjava korpusov formalnega in neformalnega govorjenega jezika

Za analizo v pričujočem delu sta uporabljena korpusa japonskega govornega jezika. Prvi je korpus Oikawa, ki je sestavljen iz podatkov, pridobljenih v intervjujih s 50 ljudmi. Uporabljen je formalni način govora, saj se sodelujoči med seboj ne poznajo in imajo različen družbeni status (učitelj in študent). Drugi je korpus NUJCC, ki ga sestavlja 100 neformalnih pogovorov, govorci pa se med seboj poznajo.

Ker korpusa nista preobsežna, je bila pogostost opazovanja nizka, vseeno pa je iz podatkov možno predvideti tendenco sopojavljanja nekaterih tipov prislovov z določeno vrstjo modalnosti. Razlika v stilu (formalno, neformalno) ima vpliv na rabo prislovov ter tendenco kolokacij med prislovi in modalno obliko na koncu stavka.

V korpusu Oikawa se z visoko frekvenco pojavljajo povedni prislovi *tabun*, *kitto*, *kanarazu* in *moshika shitara*. *Tabun* je pogosto v kolokaciji s *to omou*, opazna pa je tendenca izražanja pričakovanja. Enako je tudi s prislovom *kitto*, ki je pogosto v kolokaciji s *to omou*, *(n)darou*, *(n)deshou* in pogosto izraža pričakovanje. *Kanarazu* je v kolokaciji s *suru* in ima tendenco izražanja nujnosti. *Moshika shitara* in *kamo shirenai* se nagibata k izražanju možnosti.

V korpusu NUJCC se z visoko frekvenco pojavljajo prislovi *kitto*, *kanarazu*, *zettai(ni)*, *doumo* in *moshika shitara*. *Kanarazu* in *moshika shitara* se obnašata podobno kot v korpusu Oikawa. *Doumo* se najpogosteje sopojavlja z *rashii* in *you da* ter nakazuje na možnost. *Kitto* in *zettai(ni)* se pojavljata v kolokacijah tako z modalnimi oblikami za pričakovanje kot tudi za nujnost. Za *zettai(ni)* lahko rečemo, da je značilen za neformalen način izražanja, saj pojavlja le v korpusu NUJCC.

4.3.4 Primerjava med različnimi korpusi učbenikov

To poglavje primerja podatke, dobljene iz naslednjih treh korpusov učbenikov: korpusa japonskih naravoslovnih učbenikov za študente (16K), korpusa učbenikov japonskega jezika za osnovno šolo (Kk) ter korpusa učbenikov za srednjo šolo (KokkenK), ki je del BCCWJ-ja. Poleg celotnega korpusa KokkenK se ukvarjamo tudi

z značilnostmi podatkov iz učbenikov japonskega jezika, dodanih k omenjenemu korpusu (Kkk).

4.3.4.1 Razpršenost povednih prislovov pri korpusih učbenikov

Iz razpršenosti prislovov v učbenikih lahko opazimo njihove skupne točke. Najbolj pomenljiva skupna točka med korpusom naravoslovnih učbenikov za študente in korpusom učbenikov za srednjo šolo, ki vsebujeta podatke z različnih področij, je visoka pojavnost prislovov *kanarazu* ter *kanarazushimo*.[46] Razlike so vidne predvsem pri prislovu *kitto*. V univerzitetnih znanstvenih besedilih ga ni mogoče najti, medtem ko se pogosto pojavlja na številnih področjih v osnovnošolskih učbenikih. Nasprotno pa prislov *osoraku* pogosto opazimo v univerzitetnih znanstvenih besedilih, kar kaže na njegovo bolj formalno in zahtevno rabo v primerjavi s *kitto*. Pri primerjavi korpusov učbenikov japonskega jezika, se v obeh najpogosteje pojavljata prislova *kitto* in *kanarazu*. Razlika se vidi pri prislovu *doumo*, ki se pogosteje pojavlja v učbenikih za osnovnošolce kot za srednješolce. Nadalje je pri primerjavi prej omenjenih korpusov učbenikov japonskega jezika s korpusom srednješolskih učbenikov z različnih področij, opaziti skupno točko pri prislovih *kitto* in *kanarazu*, ki se povsod pojavljata dokaj pogosto in ju lahko razumemo kot specifična za to vrsto besedila, namenjenega učencem. Razlika je v tem, da se v korpusu srednješolskih učbenikov prislov *kanarazushimo* pojavlja pogosteje, *doumo* in *taitei* pa redkeje. Prav tako se v učbenikih japonskega jezika redkeje pojavlja *kanarazushimo*. To kaže na dejstvo, da so učbeniki jezika nekoliko drugačni od drugih učbenikov, obenem pa lahko sklepamo, da so razlike povezane tudi s tem, da učbeniki maternega jezika vsebujejo literarna besedila kot čtivo.

4.3.4.2 Povedni prislovi in modalne oblike v korpusih učbenikov

Tendenca kombinacij med povednimi prislovi in vrsto modalnosti oziroma njeno obliko v srednješolskih učbenikih je prikazana v Tabeli 16. Podatki učbenikov japonskega jezika, vsebovani v korpusu srednješkolskih učbenikov, so skopi, vendar kolokacije prislovov in modalnih oblik na koncu stavka izkazujejo enako tendenco kot celotni korpus. Prislovi *aruiwa*, *moshika shitara*, *moshika suru to* imajo tendenco kolokacij z modalno obliko za možnost (POSS), *kanarazu* z obliko za nujnost (NES), *kitto* z obliko za nujnost in pričakovanje (NESS&EXP), *tabun* in *osoraku* pa le z obliko za pričakovanje (EXP). Če primerjamo korpusa naravoslovnih učbenikov za študente in učbenikov japonskega jezika za osnovno šolo, lahko pri obeh opazimo podobnost tendenc kolokacij z vrsto modalnosti pri prislovih, ki se pojavijo v obeh

46 *Kanarazu* se pogosto pojavlja v vseh vrstah korpusov, *kanarazushimo* pa se pogosto pojavlja v specializiranih korpusih.

korpusih. Na primer, prislov *kanarazu*, za katerega je rečeno, da se v celotnem korpusu veže z modalno obliko za nujnost (NEC), se v naravoslovnih besedilih pogosto pojavlja tudi z oblikami *suru koto* in *wake de wa nai*, ki izražajo nujnost. Nadalje se v učbenikih za japonski jezik za osnovno šolo pogosto veže z *no da* in *ni chigai nai*, v učbenikih za srednjo šolo pa lahko opazimo vezavo s *hazu da* in *mono da*, tako da se pri več podkorpusih potrjuje sopojavnost z nujnostno modalnostjo.

Tabela 16: *Kolokacijski odnos med prislovi in modalno obliko v korpusu srednješolskih učbenikih (KokkenK).*

	NEChazu_da	NECmono_da	NECmono_de_wa_nai	NECni_chigainai	NECno_da	EXPdarou	EXPno_darou	EXPto_omou/-wareru	CONsou_da	POSdarouka	POSkamo shirenai	POSkoto_mo_aru	POSmono_de_mo_nai	POSno_kamo shirenai	UNMzero	OSTALO	SKUPNO
angai															1		1
aruiwa						1					**3**						4
doumo								1	1								2
hyotto shitara													1				1
kanarazu	**1**	**1**													19	**5**	26
kanarazushimo			2									1			7		10
kitto	**4**	**1**		**3**		**3**	**2**	**1**			1				8		23
moshika shitara											**1**			**1**			2
moshika suru to														**1**			1
ookata															1		1
osoraku					1	**2**	**1**	**1**									5
sazo						1											1
tabun					1	**1**		**1**		1							4
taitei															6		6
yohodo															1		1
zettai															2		2
zettai ni															7	2	9
SKUPNO	5	2	2	3	2	8	3	4	1	1	5	1	1	2	52	7	99

4.3.5 Značilnosti posameznih korpusov z vidika znanstvenih besedil

V pričujoči raziskavi analiziramo dva korpusa, ki sta sestavljena iz besedil s področja znanosti in tehnologije. 16K je korpus učbenikov naravoslovja, NLP pa sestavljajo članki o računalniški obdelavi naravnih jezikov. Ta dva korpusa sta si zvrstno podobna. Podobnosti lahko vidimo tako v razvrstitvi prislovov kakor tudi v tendenci kolokacij med prislovi in modalno obliko na koncu stavka. Korpusoma je na primer skupna visoka pojavnost prislovov *kanarazu* in *kanarazushimo*. Vendarle pa lahko najdemo tudi razlike. V naravoslovnih učbenikih se prislov *kitto* najpogosteje pojavlja v kolokaciji z modalno obliko za nujnost *suru koto*, medtem ko se v člankih iz NLP taka vrsta modalnosti sploh ne pojavi.

Če ta korpusa primerjamo z učbeniki za japonski jezik, se v slednjih prislovi *kitto*, *doumo*, *kanarazu*, *sazo*, *taiten* in *douyara* pojavljajo bolj pogosto kot v besedilih s področja znanosti in tehnologije (Tabeli 17 in 18).

Tabela 17: *Analiza korpusov NLP in 16K.*

Prislov/ modalna oblika	NEC_hitsuyou_ga_nai	NEC_koto/mono_de_wa_nai	NEC_nai_koto_ga_wakaru	NEC_no_de_wa_nai	NEC_suru_koto	NEC_wake_de_wa_nai	EXP_darou	EXP_ienai	EXP_no_de_wa_nai_darou_ka	EXP_to_kangaerareru	POSS_dekiru/nai	POSS_sarenai_koto_ga_aru	POSS_to_wa_kagiranai	UNM_da	UNM_de_wa_nai	UNM_shinai	UNM_suru	UNM_to_naru	UNM_V-te_inai	UNM_V-te_iru/kuru/oku	OTHER	total
16K_kanarazu	0	0	0	0	**13**	1	2	0	0	0	5	0	0	3	0	1	21	3	0	3	4	56
16K_ kanarazushimo	**1**	**4**	**6**	**1**	**0**	**3**	0	3	0	0	0	0	6		**18**	**11**	0	0	1	0	1	55
16K_osoraku	0	0	0	0	0	0	**3**	**0**	**1**	**3**	0	0	0	0	0	0	0	0	0	1	1	9
NLP_kanarazu	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3	0	0	3	0	7
NLP_ kanarazushimo	**0**	**2**	**0**	**0**	**0**	**6**	0	2	0	0	0	1	1	0	**6**	**3**	0	0	2	0	0	23
total	1	6	6	1	13	10	6	5	1	3	5	1	7	3	24	15	24	3	3	7	5	150

Tabela 18: *Analiza učbenikov japonskega jezika.*

Prislov/ modalna oblika	NEChazu_da	NECmono_da	NECni_chigainai	NEC_no_da	EXPdarou	EXPno_darou	EXPno_de_wa_nai_ka	EXPto_omou/-wareru	CONrashii	CONyou_da	CONsou_da	POSSkamoshirenai	POSSdarouka	UNMzero	OTHER	total
angai	0	0	0	0	0	0	1	0	0	0	0	3	0	0	0	4
aruiwa	0	0	0	0	0	0	0	0	0	0	0	**5**	0	0	0	5
doumo	0	1	0	3	0	0	0	0	**7**	**1**	**1**	0	0	22	7	42
douyara	0	0	0	1	0	0	0	0	**4**	**3**	0	0	0	2	0	10
hyotto shitara	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2
kanarazu	0	0	**1**	**3**	0	0	0	0	0	0	0	0	0	**16**	1	21
kanarazu_to	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2
kanarazushimo	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
kesshite	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
kitto	**41**	**2**	**10**	**16**	**15**	**4**	**0**	**3**	0	0	0	0	0	10	37	101
moshika shitara	0	0	0	0	0	0	1	0	0	0	0	**8**	**1**	0	2	12
ookata	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	2
osoraku	0	0	1	0	**3**	0	0	**1**	0	0	0	0	0	0	0	5
sazo	0	0	0	1	**14**	0	0	0	0	0	0	0	0	0	0	15
tabun	0	0	0	0	**4**	**1**	0	0	0	0	0	0	0	1	0	6
taitei	0	1	0	1	0	0	0	0	0	0	0	0	0	**9**	1	12
yoppodo	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	2
zettai	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
zettai ni	0	0	0	1	0	0	0	1	0	0	0	0	0	4	0	7

4.3.6 Tendenca z vidika korpusa belih knjig

To poglavje obravnava kolokacije med prislovi in modalno obliko na koncu stavka v korpusu KokkenOW (del BCCWJ-ja), v katerem so zbrane uradne bele knjige, objavljene v zadnjih 30 letih.

4.3.6.1 Razpršenost povednih prislovov v korpusu belih knjig

V primerjavi z drugimi korpusi je v korpusu KokkenOW razpršenost bistveno drugačna, kar kaže na specifičnost podatkov v tem korpusu. Kot je prikazano v Tabeli 12, se KokkenOW razlikuje od KokkenBK in KokkenOC, saj se tu najpogosteje pojavljajo povedni prislovi *kanarazushimo*, *kanarazu* in *koto ni yoru to*.

Ostali prislovi, obravnavani v tem poglavju, se pojavljajo redko ali pa sploh ne. Iz take porazdelitve prislovov lahko sklepamo, da spadajo med besedila s posebnim namenom, in sicer uradne spise.

V podatkih iz člankov NLP in naravoslovnih učbenikov 16K se tako kot v KokkenOW pogosto pojavljata *kanarazushimo* in *kanarazu*, vendar pa se v KokkenOW *osoraku* ne pojavlja v enaki meri, medtem ko se *koto ni yoru to* pojavlja pogosto. Iz tega lahko sklepamo, da je visoka frekvenca *kanarazushimo* in *kanarazu* skupna točka vseh izbranih besedil s posebnim namenom, medtem ko *koto ni yoru to*, *osoraku* in podobni prislovi izkazujejo posebne značilnosti teh besedil. Visoka pojavnost prislova *koto ni yoru to* je značilna za podatke iz uradnih belih knjig.

4.3.6.2 Kolokacije prislovov in modalnih oblik v korpusu belih knjig

Tabeli 19 in 20 prikazujeta tendenco kombinacij prislovov in modalnih oblik na koncu stavka. Z izjemo prislovov *kanarazushimo* in *koto ni yoru to* so kolokacije prislovov z drugimi modalnimi oblikami redke, zato je identifikacija močnih kolokacij s posameznimi modalnimi oblikami otežena. *Koto ni yoru to* ima močno vez z vrsto modalnosti za možnost (POSS), pogosto pa se pojavlja z modalno obliko *kangaerareru*, kar kaže na določeno zadržanost pri izražanju mnenj v uradnih spisih. *Kanarazushimo* se pogosto pojavlja z vsemi vrstami modalnih oblik. Poleg nezaznamovane modalnosti (*de wa nai, shinai*) se močne kolokacije kažejo s *to wa ienai*, *mono de wa nai*, *de wa nai koto*, *shinai koto/mono* in *wake de wa nai*. V nekaterih primerih se ustaljena modalna oblika poveže z drugo modalno obliko in lahko tako tvorijo plasti. Na primer, oblika *kanarazushimo... to wa ienai~* se izkazuje v kombinaciji z različnimi oblikami, kot so *kanarazushimo… to wa ienai baai ga aru/ men ga aru/ men ga mirareru/ joukyou ni aru/ koto/ to kangaerareru*. Tudi v drugih korpusih tvori *kanarazushimo* močne kolokacije. V JpWaC je to s *to wa kagiranai*, *wake de wa nai*, *ienai* in *koto/mono da*, v NLP z *wake de wa nai*, ter v 16K z *nai koto ga wakaru*, *to wa kagiranai*, *koto/mono de wa nai*.

Tabela 19: *Kolokacije med prislovi in modalno obliko v korpusu belih knjig (vrsta modalnosti za možnost).*

Prislov/ Modalna oblika	POSbaai_ga_ooi	POSde_wa_nai/shinai(~)	POSde_wa_nai_koto(~)	POShitsuyou_ga_aru/nai	POShitsuyou_to_shinai	POSkamoshirenai	POSkangaerareru	POSkangaerareru~	POSkoto/mono_de_aru	POSkoto_ga_aru/yosou_sareru	POSmirareru	POSmono_de_wa_nai	POSmono_mo_aru	POSmono_to_kangaerareru	POSmono_to_wa_ienai	POSnai_kanousei_ga_aru	POSnai_koto/mono	POSnai_to_suru	POSno_de_wa_nai	POSomowareru	POSto_(wa)_ienai(~)	POSto_(wa)_iinikui(~)	POSto_wa_kagiranai	POSto_wa_kangaerarenai_mono	POSwake_de_wa_nai
angai																									
aruiwa	2					1			2				1												
doumo																									
kanarazu										1															
kanarazushimo		**13**	**14**	**5**	**1**							**19**			**3**	**2**	**13**	**2**	**1**		**38**	**9**	**5**	**1**	**14**
koto ni yoru to							**7**	**1**			**2**									**1**					
osoraku														1											
tabun(ni)																									
taitei(wa)																									
yohodo																									
zettai										1															
zettai(N)																									
zettai ni																									

Kar se tiče modalnih oblik, s katerimi se sopojavlja *kanarazushimo* v belih knjigah, je pogosta nezaznamovana modalnost, pa vendar kolokacijo največkrat tvori z vrsto modalnosti za možnost (POSS). *Kanarazushimo* je v ujemalnem odnosu z nikalno obliko, kar se lahko razume kot delno zanikana oblika za modalnost nujnosti (NEC). V primerjavi z drugimi korpusi je v podatkih iz belih knjig znatno več modalnih oblik, ki so v kolokaciji s *kanarazushimo*. V kontekstu tako imenovanih političnih taktik, ki podatke posredujejo na specifičen način, tj. z olepševanjem, lahko pogosto opazimo izraze za možnostno modalnost. Uporaba kolokacij med prislovi in modalnostjo na koncu stavka vsebuje implikacije pri pogovoru in je kot takšna relevantna za poznavanje strategij v komunikaciji, zato je pomembno, da pojav podobno analiziramo tudi s strani pragmatične rabe.

Tabela 20: *Kolokacije med prislovi in modalno obliko v korpusu belih knjig (vse ostale vrste modalnosti).*

Prislov/ Modalna oblika	NEChazu_da	NECnakereba_naranai	NECsuru_koto/mono	NECto_suru_mono	NECto_wa_mirarenai	NECto_wa_natte_inai_koto	NECto_wa_shinai	EXPdarou	EXPde_wa_nai_no_darou_ka	EXPde_wa_nai~	EXPnai_darou	EXPnai_mono_to_kangaerareru	EXPnai_no_de_wa_nai_ka	EXPnai_to_mirareru	EXPno_darou	EXPto_ieru	EXPto_mirareru	CONnai_you_ni_omowareru
angai									1									
aruiwa	1	1														1		
doumo																		
kanarazu	1	2	1												2			
kanarazushimo					1	1	1			4	2	2	1					1
koto ni yoru to																		
osoraku														1				
tabun(ni)			1					1									1	
taitei(wa)																		
yohodo															1			
zettai			1	1														
zettai(N)																		
zettai ni		2	1															

4.4 Povzetek

V tem poglavju smo prikazali razpršenost povednih prislovov v 13 različnih korpusih, ki smo jih kasneje na osnovi teh podatkov poskušali razvrstiti v skupine (clusterje). Glede na vrsto korpusa smo pokazali podobnosti in razlike v razpršenosti prislovov ter opazovali odnos med tipično modalnostjo prislova in vrsto korpusa. Enake podatke smo izračunali s formulo entropije ter dobili podobne rezultate glede odstopanja korpusov. Pri podkorpusu publikacij korpusa BCCWJ (KokkenBK) in pri spletnem korpusu JpWaC je bilo potrjeno, da so podatki v precejšnji meri uravnoteženi in nimajo večjih odstopanj glede na razpršenost prislovov v primerjavi z ostalimi korpusi oz. podkorpusi.

V pričujoči raziskavi smo z uporabo različnih korpusov ugotovili, da ima vsak prislov tendenco močne vezave z določeno vrsto modalnosti. To se sklada tudi z rezultati Kudōjeve analize. Glede na tendenco kolokacij med prislovi in modalno

obliko smo lahko vrsto modalnosti razdelili z grupiranjem v štiri skupine, ki jih predlaga Kudō. Ugotovili smo tudi, da se tipična vrsta modalnosti za posamezen prislov v Kudōjevih podatkih ne sklada nujno z rezultati analize spletnih korpusov in da lahko pričakujemo razlike v različnih zvrstah besedila. To je bilo potrjeno pri primerjavi tendence kolokacij med prislovi in modalno obliko na koncu stavka v različnih korpusih, kjer smo ugotovili, da se tako razpršenosti prislovov kot tudi tendence glede na vrsto korpusa razlikujejo. V zvezi s tem bo v bodoče pomembno razmisliti, kako lahko rezultate analize različnih vrstah korpusov ter različnih jezikovnih zvrstah v teh korpusih praktično apliciramo na učenje japonščine. Tako bi določili standard, v kakšnem številu poučevati prislove ter kakšen bi bil tempo učenja kombinacij med prislovi in modalno obliko na koncu stavka, glede na stopnjo znanja japonščine. Pomembno je preučiti tudi informacije glede na vrste besedil in namembnost slovarjev. Tudi pri urejanju slovarjev je treba upoštevati pomembne informacije, kot so posebna jezikovna raba v akademskem pisanju in drugih registrih, ter razlike v rabi jezika pri pisani in govorjeni japonščini. Pri uresničevanju teh ciljev igrajo pomembno vlogo korpusi japonskega jezika, kot so BBCWJ in drugi. Treba je omeniti tudi to, da na Nacionalnem inštitutu za japonski jezik in jezikoslovje trenutno potekajo dela na projektu uravnoteženega govornega korpusa, ki bo vključeval spontane dialoge v različnih situacijah.

Kar se tiče razdalje pri oddaljenih kolokacijah, pri katerih smo kot primer obravnavali kolokacije med prislovom in modalnostjo na koncu stavka, je lahko med besedama, ki sestavljata kolokacijo, tudi več kot 10 besed. Opazili pa smo, da se ta razdalja razlikuje glede na prislov in glede na zvrst. Pri pojavu kolokacij na daljavo kot tudi pri problemu ugotavljanja dejanske tipične razdalje in razpona med povezanimi besedami bosta v prihodnosti potrebna sistematičen pristop in podrobna jezikovna analiza.

5 Luščenje kolokacij na daljavo med prislovi in modalnimi oblikami na koncu stavka ter predlog za uporabo v učnem gradivu za japonščino[47]

5.1 Uvod

V prejšnjem poglavju smo opisali pojav kolokacij na daljavo in kot primer v japonskem jeziku podali kolokacije prislova in modalne oblike, ki smo jih analizirali v več različnih korpusih. Splošno gledano lahko kolokacije na daljavo, ki se pojavljajo kot dve enoti v oddaljenem položaju, izluščimo z orodjem SkE, če so te enote prepoznavne pri morfološki analizi. Za ta namen je pomembno določiti razdalje kolokacij s pomočjo regularnih izrazov ter ciljne enote v slovnični datoteki. Kolokacije prislovov, ki jih lahko izluščimo iz korpusov na tak način z uporabo obstoječe slovnične datoteke, opisane v 2. poglavju, so kolokacije prislova z glagolom, pridevnikom na *-i* ali *-na* ter končnim členkom. Vendar oblike, ki niso prepoznavne z orodjem za morfološko analizo kot ena enota z določeno besedno vrsto, ne morejo biti izluščene iz korpusa, saj so za to potrebni dodatni koraki. To velja tudi za oblike modalnosti, ki so večinoma v orodju razdeljene na več različnih enot in jih v ciljni obliki ni mogoče izluščiti z enostavnim modificiranjem slovnične datoteke.

Rezultat proučevanja številnih stavkov iz različnih korpusov je pokazal, da se modalne oblike kažejo v različnih variantah. Glede na to, da različne variante modalne oblike na koncu stavka niso prepoznane kot posamezne enote in jih ni možno pridobiti na enostaven način, je za to treba uporabiti dodatne jezikovne pripomočke, npr. obširen seznam modalnih oblik in njihovih besednih vrst ter pripomoček za prepoznavanje pravopisnih različic.

Pri uporabi splošnih orodij za računalniško obdelavo naravnih jezikov se pojavi problem, ko so modalne oblike *kamo shirenai*, *ni chigai nai* in druge, ki bi morale biti obravnavane kot pomenske enote, razdeljene v različne morfeme ali pa fraze. Kot pri *kamo shirenai* (かもしれない), *kamo shiremasen* (かも知れません) in *kamo* (かも) je raba modalne oblike odvisna od več dejavnikov, npr. stila pisanja, formalnosti in podobno, in se razlikuje glede na sestavne dele, izbiro pisave, število morfemov, izbor variant. Z uporabo pripomočka, ki določi in opiše modalno obliko kot enoto, postaneta samodejno luščenje ali prepoznavnost modalne oblike mogoča.

Slovar funkcionalnih izrazov Tsutsuji[48] (Matsuyoshi in Sato, 2007), ki ima hierarhično strukturo, je eden od tovrstnih jezikovnih pripomočkov. Modalno obliko

47 Določeni deli tega poglavja so objavljeni v Srdanović et al. (2009a) in Srdanović et al. (2009b).

48 Dosegljivo na strani http://kotoba.nuee.nagoya-u.ac.jp/tsutsuji/ (dostop 11.4.2015)

do določene mere tudi zabeleži, ker jo obravnava kot vrsto funkcionalnega izraza. Fujiike (2008) predlaga uporabo »daljše enote« (長単位) za sestavljene in zložene besede, ki temeljijo na frazah. Na primer, *kamo shiremasen* in *wake de wa nai* sta vsaka posamezno določena kot ena daljša enota. Ker tako slovar funkcionalnih izrazov kot tudi daljše enote niso specializirane v modalnih oblikah, ne vsebujejo vseh modalnih oblik, ki jih obravnavamo v monografiji.

Kida et al. (2002) vse prislove in izraze, ki se sopojavljajo s prislovi, ovrednotijo kot ujemalne izraze. Osredotočili so se na to, da je moč napovedati ujemalni izraz v stavkih, ki vsebujejo prislov. Zato so ročno ustvarili korpus, ki je vseboval označen (ang. *tagged*) ujemalni odnos med prislovom in modalnostjo, ter uredili slovar ujemalnih izrazov, osnovan na hierarhiji modalnosti. Ker pa je bilo delo opravljeno ročno, ne ponujajo sistema ali jezikovnih pripomočkov za samodejno prepoznavanje modalnosti.

V tem poglavju najprej opišemo luščenje kolokacij na daljavo med prislovi na eni strani in glagoli, pridevniki na *-i* in *-na* ter končnimi členki na drugi. Nato podrobno opišemo metodo in postopke za avtomatsko luščenje modalnih oblik. S ciljem, da omogočimo iskanje kolokacijskih odnosov med raznolikimi modalnimi oblikami in prislovi s pomočjo orodja SkE, pripravimo seznam modalnih oblik in njihovih variant, določimo enote modalnih oblik, njihove besedne vrste, ter ponovno označevanje korpusa s dodatno oznako »modalnost«. Na koncu s primeri ponazorimo možno aplikacijo dobljenih rezultatov na poučevanje japonščine.

5.2 Luščenje prislovov in njihovih kolokacij na daljavo

Kolokacije med prislovom in glagolom, ki so v bližnji povezavi, lahko poiščemo z besednimi skicami v japonski različici orodja SkE in hitro dobimo rezultate. Če so kolokacije na daljavo določene kot posamezne enote z uporabljenim morfološkim analizatorjem ChaSen, s preprostim dodajanjem novega pravila ali spreminjanjem faktorja razdalje pri obstoječih pravilih v slovnični datoteki orodja SkE omogočimo luščenje tudi teh kolokacij in njihov prikaz med besednimi skicami. Takšni kolokacijski odnosi so 1) (oddaljena) kolokacija med prislovom in glagolom (modifies_V), 2) (oddaljena) kolokacija med prislovom in pridevnikom na *-i* ali *-na* (modifies_Adj) ter 3) (oddaljena) kolokacija med prislovom in končnim členkom (particle_fin). Slika 17 kot primer prikazuje prislov *tabun* in glagole, pridevnike ter končne členke, ki z njim pogosto tvorijo kolokacije. Vključuje tudi oddaljene kolokacije.

modifies_V	2632	8.5
思う	47	11.87
勘違い	4	9.52
想定	4	9.43
想像	6	8.98
違う	18	8.88
言う	52	8.45
にる	12	8.37
なる	141	8.35
ある	116	7.84

modifies_Adj	732	
大丈夫	10	
よい	24	
いい	30	
無理	11	
すごい	12	
好き	15	
強い	14	
本気	6	
怖い	7	

particle_fin	201	12.0
ね	81	37.33
よ	61	34.05
なぁ	13	22.65
なあ	8	18.73
な	11	18.23
の	10	8.93

Slika 17: *Luščenje (oddaljenih) kolokacij s* tabun *z besednimi skicami: primer z glagolom, pridevnikom na* -i *in* -na *ter končnim členkom.*

V nasprotju z zgoraj omenjenimi besednimi vrstami kolokacij z modalnostjo na koncu stavka ni mogoče poiskati le s spremembo slovnične datoteke v orodju SkE. Eden od razlogov za to je, da orodje za analizo morfemov ChaSen ne prepozna modalne oblike na koncu stavka kot posamezne enote, drugi pa je ta, da nimamo seznama japonskih modalnih oblik in njihovih variant. V nadaljevanju razpravljamo o metodi, ki bi v orodju SkE omogočila prikaz oddaljenih kolokacij z modalno obliko na koncu stavka.

5.3 Luščenje modalnih oblik

5.3.1 Metoda pridobivanja modalnih oblik[49]

Analiza iz 4. poglavja je bila v pomoč pričujoči raziskavi, saj omogoča vpogled v različnost modalnih oblik. Obstoječim rezultatom smo dodali nove podatke in nadaljevali analizo modalnih oblik. Najprej smo s pomočjo korpusa JpWaC pripravili vzorčni korpus iz 20 milijonov morfemov. Iz tega korpusa smo nato izluščili stavke s povednimi prislovi in si ogledali, na kakšen način je izražena modalnost. Na podlagi teh izsledkov smo sestavljeni seznam modalnih oblik, na več morfemov deljene modalne oblike in ponovno dodeljene oblikoskladenjske oznake modalnih oblik uvozili v SkE.

Format korpusa JpWaC v orodju SkE uporablja niz oznak za vsako besedo: besedno obliko, lemo in besedno vrsto, in sicer iz rezultatov, pridobljenih iz analizatorja ChaSen. Kot smo že omenili, so nekatere modalne oblike sestavljene iz več morfemov, po drugi strani pa se ena beseda v SkE sklada z enim morfemom v analizatorju ChaSen. Iz tega razloga je pomembno, da modalno obliko sestavimo v eno

49 Pri izdelavi metode pridobivanja modalnih oblik je sodeloval Bor Hodošček. Podrobnosti te raziskave so predstavljene v japonski reviji *Shizen gengo shori* (*Journal of Natural Language Processing*) (gl. Srdanović et al. 2009b).

besedno enoto in jo ponovno označimo v korpusu. Iz rezultatov raziskave vzorčnega korpusa je postalo jasno, da ima pojavnost modalne oblike številne variante, kot so praktična uporaba, stil in zapis. Zato smo modalne oblike z enakim pomenom (*darou, deshou, de arou, daro*) združili v eno reprezentativno obliko in poenotili 31 gesel. Teh 31 reprezentativnih modalnih oblik je prikazanih v Tabeli 21.

Tabela 21: *Gesla modalnih oblik v vzorčnem korpusu JpWaC (povzeto po Srdanović et al. 2009b).*

kashira, kana, kamo shirenai, -zaru wo enai, sou da, darou, -te wa naranai, to ienai/ to ieru, to wa kagiranai, to omou, to kangaeru, nakute wa ikenai, nakute wa naranai, nakereba ikenai, nakereba naranai, ni chigai nai, no ka, no da/ no de wa nai, hazu da, beki da, mitai da, you da, you ni omou, rashii, wake da/ wake de wa nai, wake ni wa ikanai, ki ga suru/ ki ga shinai

Da bi lahko modalno obliko temeljiteje spoznali, smo raziskali tudi primere, v katerih se modalna oblika veže z drugo modalno obliko ali z besedo, ki ni modalna oblika. Opazili smo tendenco združevanja modalnih oblik z besedami in morfemi v določeno ustaljeno obliko. Na primer, modalna oblika *deshou* (でしょう) se pogosto pojavi samostojno, obstajajo pa primeri, ko se v stavku veže npr. na formalni samostalnik *koto* (こと) in postane nova modalna oblika *koto deshou* (ことでしょう). Na ta način se pomen določenih modalnih oblik razlikuje, odvisno od tega, ali se pojavijo samostojno ali v kombinaciji z drugimi besedami ali morfemi kot ustaljene kombinacije besed. Zato ustaljene izraze iz kombinacije modalnih oblik in besed ali morfemov obravnavamo kot nove modalne oblike. V Tabeli 22 je prikazanih 12 skupin besed, ki se lahko vežejo na modalne oblike iz Tabele 21.

Tabela 22: *Besedna/morfemska gesla, ki se lahko vežejo z modalno obliko (povzeto po Srdanović et al. 2009b).*

koto, mono, to iu, no, ka, kana, na⁺*, wa*⁺*, mo*⁺*, nano*⁺*, da, de wa nai*
⁺ kot pri *to wa ienai,* dopuščeno je vstavljanje le v modalno obliko

Jasna je tudi tendenca, da se modalna oblika pojavi v zaporedju z drugo modalno obliko. Tu se pomen določene modalne oblike spet razlikuje, če se ta pojavi samostojno ali pa kombinirano z drugo modalno obliko. Na primer, *de arou to omoimasu* v monografiji imenujemo »sestavljena modalna oblika«, po drugi strani pa obstaja samostojna modalna oblika *de arou*. V orodju SkE se sestavljene modalne oblike ne obravnavajo povsem ločeno, ampak več variant sestavljene modalne oblike obravnavamo kot eno novo modalno obliko. Tako modalna oblika *de arou to omoimasu*

postane del nove leme sestavljene modalne oblike *darou to omou*, kjer so zaporedni morfemi zbrani kot posamezne leme in v SkE postanejo ena enota.

Tabela 23: *Predelava leme v korpusu JpWaC, ko se ta v orodju ChaSen razlikuje: primer* omou *»misliti« (povzeto po Srdanović et al. 2009b).*

Pred predelavo				Po predelavi		
besedna oblika	lema	besedna vrsta	→	besedna oblika	lema	besedna vrsta
おもう *omou*	おもう	glagol – neodvisen		おもう	思う	glagol – neodvisen
思う *omou*	思う			思う		

Tabela 24: *Predelava leme in oznake besedne vrste v korpusu JpWaC, ko se ta v orodju ChaSen razlikujeta: primer* kanarazu *»obvezno« (povzeto po Srdanović et al. 2009b).*

Pred predelavo				Po predelavi		
besedna oblika	lema	besedna vrsta	→	besedna oblika	lema	besedna vrsta
必ず *kanarazu*	必ず	prislov – vezava s členkom		必ず	かならず	prislov – splošno
かならず *kanarazu*	かならず	prislov – splošno		かならず		

Kot je razvidno iz Tabel 23 in 24, lahko kljub enakemu pomenu in funkciji najdemo primere, v katerih se leme pri morfološki analizi razlikujejo glede na zapis s pismenkami ali s hiragano ali pa se pojavijo celo razlike pri oznakah besednih vrst. To ni omejeno le na modalne oblike, problem se pojavi tudi pri prislovih, ki so včasih sestavljeni iz več morfemov (*kanarazushimo*, *hyotto shitara*). Ponovno smo torej dodelili leme in besedne vrste modalnih oblik in prislovov znotraj korpusa, zato da bi vse modalne oblike in prislove poenotili na stopnji gesel in besednih vrst.[50] Vse oznake besednih vrst za modalno obliko smo na novo zapisali in označili (Mod). S tem smo omogočili, da jih v primeru razlik v zapisu modalnih oblik in prislovov zaradi sloga pisanja ali napačne analize v orodju ChaSen še vedno lahko sistematično poiščemo z orodjem SkE.

Če v pričujoči raziskavi na ta način izluščene modalne oblike izrazimo v številkah, je reprezentativnih modalnih oblik 31, tistih, ki se pojavljajo v kombinacijah, 596 in variant pojavne oblike 2641.

50 Zaradi tega je pomembno, da lahko v SkE iščemo poenotene izraze. Na primer, vse prislove iščemo po vpisih v hiragani in ne v pismenkah (多分 *tabun* → たぶん *tabun*).

Upoštevajoč zgornje točke, smo datoteki slovničnih odnosov v SkE, ki določa različne kolokacijske odnose, dodali pravilo, ki definira odnos med prislovi in modalnimi oblikami, kot prikazuje Tabela 25.

Tabela 25: *Pravila, ki določajo odnos med prislovi in modalno obliko.*

```
*DUAL
=Adv/modality
        2:"Adv.*" "P.advzer"? []* 1:"Mod."
```

Tudi za to pravilo smo uporabili t. i. skladnjo korpusne poizvedbe v SkE osnovani na pravilnih izrazih, besednih vrstah in besedah. "DUAL" je pravilo, ki lahko išče odnos med prislovom in modalno obliko iz obeh strani. "Adv" označuje prislov, "Mod" pa modalnost. S temi pravili ne iščemo kolokacij le z modalno obliko, ki se pojavi prva za prislovom, temveč tudi z ostalimi modalnimi oblikami, ki se pojavijo do konca povedi. Pravila, s katerimi luščimo kolokacije med prislovom in modalno obliko, ne upoštevajo stukturne odvisnosti, ki bi bila možna z metodo luščenja kolokacij z uporabo orodja CaboCha. Ta ima v primerjavi z metodo, predlagano v tem poglavju, kar nekaj slabosti. Ena je ta, da modalna oblika presega meje stavka, ki je v orodju enota strukturne odvisnosti. Druga pa je ta, da je težko izvesti natančno analizo strukturne odvisnosti kolokacij na daljavo v spletnih besedilih, kot na primer v JpWaC, saj se v precejšnji meri razlikuje od domene, ki jo je analizator CaboCha uporabil za učenje.

5.3.2 Rezultat luščenja modalnih oblik

Različnim modalnim oblikam in njihovim variantam, ki jih lahko vidimo v vzorčnem korpusu JpWaC z 20 milijoni besed, so bile dodane oznake za modalnost. Z vključitvijo novega pravila kolokacijskega odnosa v slovnično datoteko je postala mogoča poizvedba kolokacij med prislovi in modalno obliko na koncu stavka z besednimi skicami.

Na Sliki 18 je prikazan rezultat označevanja kolokacij med prislovi *tabun*, *douyara*, *moshika shitara*, *kanarazushimo* in modalno obliko. Vsak prislov tvori kolokacijo z različnimi modalnimi oblikami, tipične modalne oblike, s katerimi se prislovi sopojavljajo, pa se razlikujejo od prislova do prislova. *Tabun* pogosto tvori kolokacije z modalnimi oblikami *no darou*, *darou*, *to omou* in *no darou to omou*, ki izražajo modalnost pričakovanja. *Douyara* jih tvori z domnevnimi oblikami, kot so *rashii*, *you da*, *no you da, mitai da*. *Moshika shitara* se veže z oblikami *kamo shirenai, no*

たぶん

modality	959	18.2
のだろう	124	36.59
と思う	221	35.87
だろう	175	35.23
のだと思う	32	30.22
のではないかと	28	24.82
と思うのだ	24	22.54
ことだろう	15	18.74
はず	22	18.64
のだ	73	17.3
ではないだろう	9	16.82
気がする	16	16.39
のではないか	13	16.07
のかもしれない	13	14.97
かな	13	14.88
かもしれない	18	14.78
に違いない	7	14.34
そうだ	13	12.99
のではないだろうか	8	12.77

どうやら

modality	382	20.3
らしい	121	48.54
ようだ	95	40.61
のようだ	27	31.21
みたいだ	16	26.2
そうだ	21	21.46
らしいのだ	5	16.36
ことらしい	3	13.33
のだ	24	12.04
と考える	4	9.39
のだろうか	5	8.77
のかもしれない	4	8.53
かもしれない	5	8.06
のではないかと	3	7.59
と思う	8	6.98
かな	3	6.73
のだろう	4	6.36
のか	4	4.4

もしかしたら

modality	269	20.7
かもしれない	102	43.36
のかもしれない	59	39.67
のかな	10	19.6
のではないか	8	15.87
のではなきかと	5	12.04
に違いない	3	10.41
と思う	11	10.3
のか	9	10.15
のだ	13	8.87
だろう	6	7.36
気がする	3	7.21
のだろうか	3	6.47
のだろう	3	5.6

かならずしも

modality	249	18.1
とは限らない	53	51.05
わけではない	35	34.77
ものではない	11	21.89
ことではない	9	21.89
といえない	8	20.93
ではないと思う	4	16.02
ものではないと思う	3	15.83
と思う	17	14.12
と考える	6	13.69
のだ	23	13.66
ようだ	8	13.06
気がする	5	13.06
ではないのだ	3	10.74
かもしれない	6	10.74
だろう	9	10.47
のではないか	4	10.19
らしい	4	9.34

Slika 18: *Rezultat iskanja kolokacijskih izrazov med prislovi in modalno obliko v besednih skicah:primer s prislovi* tabun, douyara, moshika shitara, kanarazushimo.

kamoshiranai, no kana, ki izražajo možnost. *Kanarazushimo* pa pogosto tvori kolokacije z oblikami *to wa kagiranai, wake de wa nai* in *mono de wa nai*, ki sodijo v nujnostni tip modalnosti.

きっと

modality	876	18.8
のだろう	160	41.04
だろう	159	34.54
に違いない	34	33.09
ことだろう	44	32.13
はず	52	29.05
と思う	125	28.76
のだと思う	14	20.53
と思うのだ	15	17.9
のではないかと	14	17.61
かもしれない	23	17.49
のだろうと思う	6	17.03
のだ	66	16.83
のかもしれない	15	16.66
だろうと思う	7	15.37
気がする	12	14.11
のではないだろうか	9	14.08
のではないか	9	13.1
のか	20	11.75
のだろうか	10	11.27

Slika 19: *Besedne skice: primer kolokacij med* kitto *»gotovo« in modalno obliko.*

Slika 19 prikazuje besedno skico kolokacij med prislovom *kitto* »gotovo« in modalno obliko na koncu stavka. Seznam oblik je urejen v zaporedje glede na statistično izračunano izpostavljenost kolokacije (številke na desni strani), številke na levi strani pa izražajo frekvenco. Če kliknemo številko v drugi vrsti na sliki, si lahko ogledamo stavčne primere tudi v korpusu. Iz pridobljenih modalnih oblik razumemo, da *kitto* pogosto tvori kolokacije z modalno vrsto za nujnost in pričakovanje.

Tabela 26 prikazuje frekvenco izbranih pogostih variant modalnih oblik, ki se sopojavljajo s prislovom *kitto*. Pridobili smo jih z uporabo funkcije za prikaz stavčnih primerov, in sicer iz izbranih 1000 stavkov s prislovom *kitto*.

Tabela 26: *Izbrane pogoste variante modalnih oblik, ki se vežejo s* kitto *»gotovo«.*

Modalna oblika	Variante modalnih oblik
to omou 89	to omoimasu 57, to omou 18, to omoimasu yo 8, to omou no desu 2, to omou yo 2, to omotta no da 2
darou 73	darou 53, darou na 6, darou naa 5, darou to omou 3, darou to omoimasu 2, darou to omou no desu 2, darou ne 2
deshou 65	deshou 55, deshou ne 10
hazu 54	hazu desu 19, hazu 18, hazu da 17

Modalna oblika	Variante modalnih oblik
ni chigai nai 36	ni chigai nai 34, ni chigai arimasen 2
no darou 27	
no deshou 26	no deshou 21, no deshou ne 5
n'darou 21	n'darou na 9, n'darou naa 7, n'darou 5
n'deshou 18	n'deshou 9, n'deshou ne 9

Rezultati analize korpusa JpWaC izpostavljajo pomembnost kolokacijskega razmerja med prislovom in modalno obliko na koncu stavka ter omogočajo vpogled v reprezentativne modalne oblike. Pri učenju japonskega jezika je pomembno nameniti pozornost tovrstnim kolokacijskim razmerjem. V nadaljevanju nakažemo, kako lahko empirično izluščene rezultate apliciramo na sestavljanje besedišča v okviru priprave učnega načrta.

V Tabeli 27 je prikazana razpršenost prislovov v korpusih ter pet modalnih oblik, ki imajo najvišji indeks izpostavljenosti s posameznim prislovom.

Tabela 27: *Pogosti kolokacijski odnosi med prislovom in modalno obliko na koncu stavka v vzorcu korpusa JpWaC.*

Vrsta modalnosti	Prislov	Frekvenca prislova	Pet najpogostejših modalnih oblik (glede na frekvenco in izpostavljenost kolokacij)				
EXP	tabun	1527	no darou 124	to omou 221	darou 175	no da to omou 32	no de wa nai ka to 28
NEC/ EXP	kanarazu	1448	hazu 31	no da 95	to omou 41	darou 20	wake da 6
EXP	osoraku	1341	darou 269	no darou 128	koto darou 37	to omou 131	ni chigai nai 21
EXP/ NEC	kitto	1340	no darou 160	darou 159	ni chigai nai 34	koto darou 44	hazu 52
NEC/ EXP	zettai	1294	no da 87	darou 40	to omou 41	hazu 17	beki 13
CON	doumo	1022	you da 59	rashii 44	ki ga suru 28	no da 63	no you da 12
NEC	zettai ni	816	te wa naranai 21	no da 79	beki 15	to omou 29	hazu 12

Vrsta modalnosti	Prislov	Frekvenca prislova	Pet najpogostejših modalnih oblik (glede na frekvenco in izpostavljenost kolokacij)				
CON	douyara	548	rashii 121	you da 95	no you da 27	mitai da 16	sou da 21
NEC/EXP	taitei	497	no da 45	to omou 17	darou 14	hazu 6	you ni omou 3
EXP	yohodo	409	no darou 34	darou 22	to omou 25	no da 31	no ka 17
NEC	kanarazushimo	382	to wa kagiranai 53	wake de wa nai 35	mono de wa nai 11	koto de wa nai 9	to ienai 8
POSS	moshika shitara	316	kamo shirenai 102	no kamo shirenai 59	no kana 10	no de wa nai ka 8	no de wa nai ka to 5
POSS	angai	187	no kamo shirenai 11	ki ga suru 9	no da 18	kamo shirenai 8	darou ka to omou 2
POSS	hyotto shitara	81	kamo shirenai 25	no kamo shirenai 17	to kangaeru no kamo shirenai 1	no de wa nai ka to 2	kamo shirenai no da 1
NEC/EXP	taigai	72	no da 6	to omou 4	wake da 2	no kana to omou 1	darou 3
EXP	sazo	31	koto darou 6	darou 8	no darou 5	darou to omou 2	darou to kangaeru 1
EXP	ookata	24	mono to omou 1	mono darou 1	to omou 2	no kana 1	to kangaeru 1
CON	koto ni yoru to	2	rashii 1	beki 1			to kangaeru 1

Gledano z vidika razpršenosti prislovov se pogosto pojavljajo *tabun, kanarazu, osoraku, kitto* in *zettai*. Če pogledamo kolokacijski odnos med prislovi in modalnimi oblikami, postane jasno, da se najpogosteje pojavljajo modalne oblike *no darou, darou, to omou*, ki izražajo pričakovanje, ter *hazu, no da, ni chigai nai,* ki izražajo nujnost. Če imajo nekateri prislovi tendenco kolokacij s pričakovalno modalno obliko, je med kolokacijami s prislovi zaslediti tudi nujnostne modalne oblike. Primera tovrstnih prislovov sta denimo *kitto* in *osoraku,* ko tvorita kolokacije z *ni chigai nai* in drugimi modalnimi oblikami za nujnost. Obratno velja v primeru, ko prislov tvori

kolokacijo z nujnostno modalno obliko – takrat tvori ta prislov kolokacijo tudi s pričakovalno modalno obliko. *Kanarazu* in *zettai* denimo tvorita kolokacijo tudi z *darou* in *to omou*, tako imenovano pričakovalno modalno obliko na koncu stavka.

Rezultat primerjave razpršenosti prislovov in kolokacij med prislovi in modalno obliko na koncu stavka v vzorčnih podatkih iz korpusa JpWaC in korpusu publikacij znotraj uravnoteženega korpusa BCCWJ je pokazal podobno porazdelitev in tendenco kolokacijskih odnosov.

Ne le v besednih skicah, tudi v konkordančniku, primerjalnih skicah in tezavru lahko iščemo podatke o prislovih in modalnih oblikah. Ko s primerjalnimi skicami primerjamo *tabun* in *kitto*, opazimo, da se *kitto* pogosteje veže z *no darou, ni chigai nai, koto darou* in *hazu*, in vidimo lahko lastnosti, ki so blizu vrsti nujnostne modalnosti. Na drugi strani pa *tabun* tvori kolokacije s *to omou, no da to omou, no de wa nai ka to, to omou no da,* ki izražajo pričakovanje. Ker nujnost izraža večjo stopnjo verjetnosti kot pričakovanje, lahko sklepamo, da je stopnja verjetnosti nižja pri *tabun* kot pri *kitto*.

5.3.3 Vrednotenje izluščenih kolokacij prislovov in modalnih oblik

V tem poglavju ovrednotimo natančnost izluščenih kolokacij med prislovi in modalno obliko na koncu stavka. Obstajajo štiri reprezentativne vrste modalnosti, ki tvorijo kolokacije s prislovi (Kudō, 2000) in iz vsake za namene vrednotenja izberemo po en prislov, ki se veže najpogosteje. Iz rezultatov iskanja modalnosti, ki tvori kolokacijo s posameznim prislovom, izberemo štiri modalne oblike (prvi štirje vnosi z vrha s seznama prikazanih modalnih oblik v rezultatih besednih skic) ter iz konkordanc naključno izberemo po deset stavčnih primerov s kolokacijami izbranih prislovov in modalnih oblik. Tako skupno ovrednotimo 160 stavčnih primerov.

Izbrane stavčne primere sta ovrednotila strokovnjaka na področju poučevanja japonskega jezika (A, B) in strokovnjak na področju računalniške obdelave naravnih jezikov (C). Možne izbire so »pravilno« (kolokacija med prislovom in modalno obliko je pravilna), »delno« (kolokacija med prislovom in modalno obliko je delno pravilna) in »nepravilno« (kolokacija med prislovom in modalno obliko ni pravilna«. Strokovnjaki so imeli tudi možnost komentiranja posameznih primerov.

Tabela 28: *Rezultati ovrednotenja (povzeto po Srdanović et al. 2009b).*

IZBIRA	ODGOVOR		
	Ocenjevalec A	Ocenjevalec B	Ocenjevalec C
(pravilno) kolokacija med prislovom in modalno obliko je pravilna	93,12 %	93,75 %	96,87 %
(delno) kolokacija med prislovom in modalno obliko je delno pravilna	1,8 %	1,25 %	1,87 %
(nepravilno) kolokacija med prislovom in modalno obliko ni pravilna	5 %	5 %	0.6 %

Kot vidimo v Tabeli 28, je natančnost visoko ovrednotena. Iz stavčnih primerov, ki so ovrednoteni kot nepravilni, ter komentarjev ocenjevalcev lahko sklenemo naslednje:

Obstajajo stavčni primeri, v katerih prihaja do napačne analize delov modalnosti.

1. V nekaterih stavčnih primerih je japonščina uporabljena nepravilno.
2. V korpusih so podvojeni stavčni primeri.
3. V nekaterih stavčnih primerih prislov in modalna oblika ne tvorita kolokacije.
4. V nekaterih stavčnih primerih se kolokacija tvori z drugo modalno obliko in se razlikuje od prikazanega.

Pri točki 1 je odprava težave relativno preprosta, in sicer moramo preveriti, kako so bile oznake določene, nato pa določiti pravilno modalno obliko in jo dodati na seznam variant. Pri točki 2 je treba popraviti nepravilne oznake, pri točki 3 pa odstraniti podvojene stavke. Možna razlaga za točki 4 in 5 je ta, da je bilo več stavkov vnešenih v korpus kot en stavek zaradi napak pri računalniški obravnavi, lahko pa, da razlog leži v kompleksni stavčni strukturi, zaradi katere je prislovu samodejno dodeljena modalna oblika, s katero dejansko ne tvori kolokacije. V prihodnosti bo treba raziskati tudi možnost implementacije programa za skladenjsko analizo, ki prikaže ujemalni odnos v stavčni strukturi.

5.3.4 Primerjava s Kudōjevimi podatki

V prejšnjem delu smo govorili o vrednotenju in visoki natančnosti rezultatov luščenja kolokacij. Naprej te rezultate primerjamo z raziskavo o kolokacijskih odnosih med

prislovi in modalno obliko na koncu stavka v sklopu študij japonskega jezika (Kudō, 2000). Kudō pravi, da bi morali o ujemanju z modalno obliko razmišljati kot o stopnjevalnem pojavu, saj vsak prislov pripada določeni vrsti modalnosti in ima tendenco kolokacij s to modalno obliko. Za podrobnosti o podatkih Kudōjeve raziskave glej Tabelo 14 v poglavju 4.3.2. Na primer – *kitto* pripada vrsti modalnosti za nujnost in tvori kolokacije s *suru/no da* in *ni chigai nai*. *Osoraku* se pogosto veže s pričakovalnimi oblikami *darou/mai, to omowareru, no de wa nai darou ka*. *Douyara* se pogosto veže z domnevno obliko *rashii, moshika shitara* pa z možnostno obliko *kamo shirenai*.

Tudi v podatkih, pridobljenih s pričujočo raziskavo, lahko kot v Kudōjevih podatkih prepoznamo štiri vrste modalnosti, s katerimi prislovi tvorijo kolokacije. Tudi v teh podatkih lahko največkrat vidimo kolokacije s pričakovalno in nujnostno vrsto modalnosti. Mnoge modalne oblike, ki jih lahko zasledimo v Kudōjevih analizah, se pojavijo tudi v korpusu JpWaC, vendar lahko v zaporednosti frekvence prislovov in v tendenci kolokacij s posameznimi prislovi opazimo tudi razlike. Na primer, pri prislovu *kitto* se v pričujoči raziskavi, za razliko od rezultatov Kudōjeve analize, bolj pogosto pojavijo kolokacije z vrsto modalnosti za pričakovanje kot za nujnost. Pri Kudōju lahko najdemo tudi malce starejšo terminologijo, na primer *sazo* se tu pojavlja precej pogosteje kot v korpusu JpWaC. Razlog za te razlike leži v različnosti korpusov, kar je bilo poudarjeno že v petem poglavju z rezultati klasifikacije več vrst korpusov.

5.3.5 Primerjava z uravnoteženimi publikacijami

V tem poglavju preučujemo kolokacijske odnose med prislovi in modalno obliko na koncu stavka v BCCWJ-jevem korpusu publikacij (KokkenBK) in v korpusu JpWaC. V ta namen smo sestavili manjši podkorpus iz korpusa JpWaC, ki vsebuje 20 milijona naključno izbranih besed iz originalno 400-milijonskega korpusa.

Tabeli 27 in 29 prikazujeta razpršenost prislovov v JpWaC korpusu in korpusu publikacij. Prikazujeta tudi pet modalnih oblik, ki imajo najvišjo vrednost indeksa izpostavljenosti ali se najpogosteje vežejo v kolokacije s posameznimi prislovi. Gledano z vidika razpršenosti prislovov so si rezultati iz obeh korpusov precej podobni, največje razlike pa so opazne pri pogostih prislovih *kanarazu* in *tabun*. *Kanarazu* se pogosto pojavi v korpusih publikacij. Če vzamemo v poštev, da ga najdemo v korpusu časopisov, učbenikov in znanstvenih del, lahko rečemo, da je značilen za besedila, pisana v formalnem jeziku (gl. Tabelo 12 v razdelku 4.2.1). *Tabun*, ki je v JpWaC pogost, je prisoten tudi v korpusu Oikawa (formalni razgovori) in podatkih iz Chiebukuro. JpWaC v primerjavi s korpusi publikacij v večji meri izraža značaj formalnih pogovorov.

Tabela 29: *Kolokacijski odnos med prislovi in pogostimi modalnimi oblikami na koncu stavka v korpusih publikacij.*

Vrsta modalnosti	Prislov	Frekvenca	Pet najpogostejših modalnih oblik (glede na frekvenco kolokacij)
NEC/ EXP	kanarazu	4548	no da (334), hazu da (163), darou (151), to omou (76), ni chigai nai (44)
EXP	osoraku	4216	darou (961), no darou (625), to omou (225), ni chigai nai (157), no da (153)
EXP/ NEC	kitto	3547	darou (417), no darou (332), ni chigai nai (251), to omou (224), hazu da (157)
EXP	tabun	3241	darou (487), no darou (412), to omou (290), no da (111), no de wa nai ka (109)
CON	doumo	2320	no da (156), rashii (149), you da (126), ki ga suru (65), to omou (43)
NEC	zettai ni	2114	no da (141), hazu da (61), darou (59), to omou (51), beki da (25)
POSS	moshika shitara	1824	kamo shirenai (401), no kamo shirenai (314), no de wa nai ka (130), no ka (95), no da (70)
NEC	kanarazushimo	1591	wake de wa nai (127), no da (97), mono de wa nai (69), to wa ienai (78), to wa kagiranai (72)
NEC/ EXP	zettai	1581	no da (111) to omou (58), darou (38), hazu da (26), no ka (25)
CON	douyara	1512	rashii (504), you da (319), no you da (87), no da (42), darou (12)
NEC/ EXP	taitei	1217	no da (98), darou (29), koto de wa nai (18), you da (15), to omou (15)
EXP	yohodo	1047	no darou (95), no kamo shirenai (76), no da (60), darou (43), rashii (42)
POSS	hyotto shitara	967	kamo shirenai (218), no kamo shirenai (148), no de wa nai ka (111), no ka (36), to omou (36)
EXP	sazo	427	darou (128), koto darou (76), no darou (29), ni chigai nai (26), to omou (23)
POSS	angai	423	no da (53), no kamo shirenai (35), kamo shirenai (35), no de wa nai ka (8), de wa nai ka (7)
EXP	ookata	281	no darou (31), no da (21), darou (16), to omou (10), koto darou (6)
NEC/ EXP	taigai	172	no da (11), darou (2), darou ka (2), hazu da (2), no ka (2)
POSS	koto ni yoru to	62	no kamo shirenai (10), kamo shirenai (8), no de wa nai ka (4), no ka (3), mono de arou (2)

Če pogledamo kolokacijski odnos med prislovi in pogostimi modalnimi oblikami, lahko v obeh korpusih vidimo podobne rezultate. Iz tega lahko sklepamo, da so tendence modalnih oblik, ki tipično tvorijo kolokacijo z določenim prislovom, enake. Zelo zanimivo je, da modalne oblike, ki izražajo pričakovanje (*no darou, darou, to omou*), in tiste, ki izražajo nujnost (*hazu, no da, ni chigai nai*), veliko pogosteje tvorijo kolokacije. To kaže na to, da imata med štirimi tipi modalnosti, ki jih navaja Kudō (2000), tisti za pričakovanje in nujnost višjo funkcionalno prioriteto v komunikaciji japonskega jezika kot tisti za možnost in domnevo.

V obeh korpusih je v primeru, da je med pričakovalno modalno obliko in določenim prislovom kolokabilna vez, prisotna tudi kolokacija tega prislova z nujnostno modalno obliko. Na primer, prislova *kitto* in *osoraku* oba tvorita kolokacijo z *ni chigai nai*. Ko pa določen prislov tvori kolokacijo z nujnostno modalno obliko, isti prislov tvori kolokacijo tudi s pričakovalno modalno obliko, npr. *kanarazu, zettai,* ki tvorita kolokacijo z *darou* in *to omou*, t. i. pričakovalno obliko na koncu stavka.

5.4 Razvoj učnega gradiva za študente japonščine

5.4.1 Korpusno zasnovan predlog za izdelavo učnega načrta besedišča

Pri učenju japonskega jezika je treba nameniti pozornost kolokacijskim razmerjem, vključno s kolokacijami med prislovi in modalno obliko na koncu stavka, na pomembnost katerih so nakazali rezultati korpusne analize. V tem poglavju najprej pogledamo, v kolikšni meri so pokrite kolokacije med prislovi in modalnimi oblikami na koncu stavka v učbenikih za osnovno in srednjo stopnjo učenja japonščine. Nato predstavimo predloge, kako lahko praktično uporabimo empirično pridobljene rezultate analiz različnih korpusov pri izdelavi učnega načrta za besedišče.

Glede izdelave učnega načrta iz rezultatov na korpusih osnovane analize bo govor o (1) vrstah korpusov, zvrsti besedila in enot učenja v učnem načrtu za besedišče, (2) povezavi med frekvenco enot v posameznem korpusu in zaporedju uvajanja enot za učenje ter (3) o tem, kako lahko te ugotovitve apliciramo na izdelavo učnega načrta.

5.4.1.1 Obravnavanje prislovov in modalnih oblik v učbenikih japonščine

Kolokacijski odnos med prislovi in modalno obliko na koncu stavka je v učbenikih za učenje japonskega jezika prikazan v primerih slovnične razlage ali v besedilu, ni

pa razlage o samem kolokacijskem odnosu ali ujemanju. Analizirali smo šest učbenikov začetne in nadaljevalne stopnje za poučevanje japonščine in ugotovili, da se s prislovi najpodrobneje ukvarja učbenik za nadaljevalno japonščino *Bunka chuukyuu nihongo 1, 2* (gl. Tabelo 30 (*Bunkachuu*)). V vsakem poglavju tega učbenika lahko najdemo razdelek, posvečen prislovom, poleg tega sta za vsakega vključena tudi dva stavčna primera in nekaj vaj, obravnavana pa je približno tretjina prislovov, ki so predmet te raziskave. Učbenik vseeno ne vsebuje razlag o povezavah med prislovi in modalnimi oblikami. Modalne oblike, ki niso obravnavane pri prislovih, se v določenih primerih pojavijo v stavčnih vzorcih, ponekod pa skupaj s prislovom, a pri tem ne gre za nameren ali sistematičen postopek podajanja informacij.

Tabela 30: *Prislovi in modalne oblike na koncu stavka v učbenikih za učenje japonščine.*

	Začetna stopnja			Nadaljevalna stopnja			
Prislov	**Shinbun-ka**	**Shinga-kusho**	**Minnano**	**Bunkac-huu**	**Shinga-kuchuu**	**ShinNi-hongo**	**JpWaC**
aruiwa	/	/	/	/(veznik)	●darou ka	/	beki, no ka, no kamo shirenai
angai	/	/	/	/	○	/	no kamo shirenai, ki ga suru, no da
osoraku	/	/	/	※darou	○	/	darou, no darou, koto darou
kanarazu	○	◇	◆te ku-dasai	/	○	/	hazu, no da, na-kereba naranai
kitto	/	●no de-shou	◆to omoi-masu	▲◆da-rou, V-te ne, ※no darou	/	◆darou, deshou	no darou, darou, ni chigai nai
zettai ni	/	/	◆te ku-dasai	▲V-ou	/	/	te wa naranai, no da, te wa ikenai
taigai	/	/	/	/	○	/	no da, to omou, wake da
taitei	/	◇	△	△◇te iru	/	/	no da, to omou, darou
tabun	◆darou to omoima-suyo	▲deshou, darou to omoimasu	◆to omoi-masu	▲◆da-rou, to omoima-su, ※no darou	/	◆to omou, ja-nai (desu ka)	no darou, to omou, darou
doumo	/	/	/	▲◆rashii, sonna kanji da	○	/	you da, rashii, ki ga suru
yohodo	/	/	/	/	○	/	no darou, darou, to omou

Legenda: ○ v tekstu △ v zgledu ◇ v vajah/ ni v učbeniku

●▲◆★ enako kot zgoraj, vsebuje tudi modalno obliko

※ v primeru stavčnega vzorca za modalne oblike

Učbeniki, ki so uporabljeni v analizi so:

1. (Shinbunka) Bunka začetna stopnja japonščine 1, 2 [Shin bunka shokyuu nihongo I, II], Bunka Institute of Language, Bonjinsha 2000,
2. (Shingakusho) Začetna stopnja japonščine za študente [Shingaku suru hito no tame no nihongo shokyuu], International Students Institute, 1994,
3. (Minnano) Japonščina za vsakogar 1, 2 [Minna no nihongo I, II], 3A Network, 1998,
4. (Bunkachuu) Bunka nadaljevalna stopnja japonščine 1, 2 [Bunka chuukyuu nihongo I, II], Bunka Institute of Language, 2006/2005,
5. (Shingakuchuu) Nadaljevalna stopnja japonščine za študente [Shingaku suru hito no tame no nihongo chuukyuu], International Students Institute, 2000,
6. (ShinNihongo) Nadaljevalna stopnja sodobne japonščine [Shin nihongo no chuukyuu], Association for Overseas Technical Scholarship, 3A Network, 2000

Rezultati korpusa JpWaC so delno povzeti v zgornji Tabeli 30 (za podroben opis gl. Tabelo 26).

Ko se modalnost pojavi v zgledu, ni nujno, da je ena tistih, ki pogosto tvori kolokacije. V šestih učbenikih se obravnavani prislovi ter kolokacije med prislovi in modalnimi oblikami razlikujejo, čeprav so namenjeni učencem z enako stopnjo znanja (osnovna ali srednja). V učbenikih osnovne stopnje je obravnavanih pet prislovov. To so *tabun, kanarazu, kitto, taitei, zettai ni*, pri tem pa se samo prva dva pojavita v vseh učbenikih. Tudi če obe stopnji obravnavamo skupno, težko vidimo kakršne koli sledi sistematične obravnave. V učbeniku Shinbunka sta obravnavana le *tabun* in *kanarazu* in le *tabun* ima zraven še modalno obliko. V učbeniku Bunkachuu je *tabun* z modalno obliko omenjen le enkrat, omenjen je tudi prislov *kitto*. V učbeniku Shingakusho sta skupaj z modalno obliko obravnavana prislova *kitto* in *tabun*, v Shingakuchuu pa kljub različnim obravnavanim prislovom kolokacije z modalno obliko niso obravnavane. V desnem stolpcu Tabele 30 (JpWaC) so prikazane tri modalne oblike z najvišjo frekvenco tvorjenja kolokacij z določenim prislovom znotraj korpusa. Če učbenike primerjamo s podatki iz korpusa, lahko rečemo, da so prikazane razlike v rabi modalnih oblik skromne, reprezentativne oblike pa sploh niso obravnavane. Na primer, čeprav je modalna oblika, s katero tvorita kolokacije prislova *tabun* in *kitto*, omenjena, se najpogostejša modalna oblika *no darou* ne pojavi v nobenem od učbenikov. V učbeniku Bunkachuu se modalna oblika sicer včasih pojavi kot del zgleda pri stavčnih vzorcih, ne pri obravnavi prislovov. V

povezavi s prislovom *kitto* modalna oblika *ni chigai nai*, s katero se prislov pogosto sopojavlja, ni obravnavana. Glede na ostale prislove je tudi pokrivanje modalnih oblik v učbenikih precej nezadovoljivo.

5.4.1.2 Vrste korpusov in cilji učencev

Glede na vrsto korpusa se tako razpršenost prislovov kot tendenca kolokacije med prislovom in modalno obliko spreminjata. Vsak korpus ima različne lastnosti, vrste besedil in stil, ki se odražajo v obnašanju prislovov. V pričujočem poglavju razpravljamo o tem, kako lahko klasifikacijo korpusov apliciramo na poučevanje japonskega jezika.

Pri vseobsežnem učenju japonskega jezika so pomembni najbolj splošni in razširjeni jezikovni podatki. Tovrstne podatke zajemajo uravnoteženi korpusi, kot so BCCWJ-jev korpus publikacij in veliki japonski spletni korpus JpWaC. Tudi časopis Mainichi shimbun (Mai 2002) lahko postane referenčni vir, saj vsebuje veliko splošnih podatkov (gl. Tabeli 1 in 12).

Da bi bile pri celostnem učenju japonskega jezika pokrite vse štiri jezikovne veščine, je dobro, da za referenčno gradivo uporabimo podatke tako pisnega kot tudi govornega korpusa. V tej raziskavi se posvečamo korpusom formalnega (Oikawa in KokkenOC) ter neformalnega govora (NUJCC). O tem, ali naj se gesla iz korpusa neformalnega jezika obravnavajo kot učna snov, obstaja med strokovnjaki za poučevanje japonščine veliko različnih mnenj, zato je izbira odvisna od izobraževalnih ciljev.

Korpuse, ki se nagibajo na specifično področje, lahko uporabimo pri izdelavi učnega načrta s specifičnim namenom. Na primer, korpusa NLP in 16K, ki sta sestavljena iz naravoslovnih besedil, lahko uporabimo pri izdelavi učnega načrta za učence japonskega jezika, ki so naravoslovci. Korpusi, ki temeljijo na japonskih učbenikih za osnovno in srednjo šolo, so zasnovani glede na cilj učnega gradiva, notranjo politiko avtorjev in založnikov učbenika ter drugih faktorjev, tako da so drugačni od ostalih korpusov. Zaradi tega jih ne moremo obravnavati kot primarno jezikovno gradivo in je potrebna posebna previdnost, če jih uporabljamo za izdelavo učnega načrta.

5.4.1.3 Frekvenca v korpusu in zaporedje obravnave učne snovi

V pričujočem razdelku se posvečamo vprašanju, na kateri stopnji obravnavati učno snov, osnovano na korpusnih podatkih in kako to snov povezati s frekvenco v korpusih. Kot referenčno gradivo je priporočljivo vzeti korpus, ki je po tematiki soroden izobraževalnim ciljem, korpusna frekvenca in izpostavljenost ciljne snovi pa naj bi se odražala v učnem načrtu. V osnovi lahko denimo vzamemo pogostost

pojavljanja kot enega od kriterijev pri določanju zaporedja obravnave učne snovi. Sinclair in Renouf (1988) denimo izpostavljata, da je smiselno pri izdelavi učnega načrta za besedišče angleščine kot tujega jezika najprej predstaviti besede z visoko pojavnostjo.

V pričujoči raziskavi upoštevamo tudi ravnotežje števila besed na posamezni stopnji učenja. Medtem ko relativno pogostost znotraj uravnoteženih korpusov BCCWJ in JpWaC vzamemo kot osnovno referenco, določimo zaporedje na naslednji način. Prislove razporedimo po relativni frekvenci, od prislova *tabun*, ki preseže 10 %, do precej nizkofrekventnega prislova *moshika sureba*, ki dosega manj kot 5 %. Zaporedje obravnave, če so ostali pogoji ustaljeni, tako kot pri zgoraj omenjenem primeru Sinclairja in Renoufa (1988), določamo po relativni frekvenci. Prislove, razporejene po relativni frekvenci, razdelimo glede na stopnjo znanja učencev, in sicer jih poskusno ločimo na tri stopnje. Prislove z visoko relativno frekvenco poučujemo na osnovni stopnji, prislove z relativno frekvenco v srednjem območju poučujemo na srednji stopnji in tiste z nizko relativno frekvenco na višji stopnji. Če območje relativne frekvence razdelimo na tri enake dele, so rezultati precej prepričljivi, kot lahko vidimo v Tabeli 31: pod 5 %, od 5 %–10 % ter več kot 10 %. Prislovov z visoko relativno frekvenco je malo. Ker se uporabljajo v širokem in splošnem kontekstu, so primerni za osnovno stopnjo. Prislovi z relativno frekvenco od 5 %–10 % so nekoliko posebni in jih lahko uvrstimo v srednjo stopnjo. V primerjavi s seznamom besedišča JLPT (seznam besedišča za osrednji izpit iz znanja japonščine kot tujega jezika) (JLPT, 2002), ki velja kot standard pri poučevanju japonščine, je veliko podobnosti, zato lahko potrdimo, da so omenjeni prislovi primerni za srednjo stopnjo. Podobno velja za prislove za višjo stopnjo, pri katerih je relativna frekvenca manjša od 5 %.

Standard za besedišče ni absoluten, ampak se spreminja z izobraževalnimi cilji. Z metodološkega vidika je torej pomembna pridobitev objektivno zasnovanega predloga uvajanja prislovov glede na njihovo relativno frekvenco, vsekakor pa je možno občasno prehajanje predlaganih prislovov med območji osnovne, srednje in višje stopnje glede na izobraževalne cilje.

5.4.1.4 Predlogi za izdelavo učnega načrta za besedišče prislovov

V tem poglavju razpravljamo o možnosti vključitve vrste korpusov in korpusne frekvence v izdelavo učnega načrta za besedišče na konkrentem primeru rezultatov analize prislovov (Tabela 12). Kot smo že omenili, se med izdelavo učnega načrta

pri vključevanju posamezne učne enote upošteva celotna struktura. Pod predpostavko, da gre za učenje splošne japonščine, namenjeno odraslim, pri izdelavi učnega načrta iz korpusov izberemo temu primerna gesla. Za to so primerni korpusi z visoko stopnjo splošnosti v podatkih. Med obstoječimi pisnimi korpusi so primerni tisti, pri katerih ni odstopanj pri podatkih, npr. BBCWJ-jev korpus publikacij (KokkenBK) in spletni korpus (JpWaC), do določene mere tudi korpus časopisov (Mai2002). Med obstoječimi govornimi korpusi sta za ta namen primerna Oikawa in KokkenOC, čeprav bi bilo seveda še bolje imeti na voljo večji uravnotežen govorni korpus japonščine. Ker je NUJCC neformalen, je treba biti pri uvajanju novih prislovov previden. Tabela 31 prikazuje zaporedje učinkovitega uvajanja prislovov. Seznam temelji na frekvenci na osnovi prej omenjenih korpusov.

Tabela 31: *Zaporedje učinkovitega uvajanja prislovov glede na stopnje pri splošnem poučevanju japonščine.*

Stopnja učenja	Prislov	Relativna frekvenca
osnovna	*tabun, kanarazu, kitto, osoraku, (zettai), (zettai ni)*	> 10 %
srednja	*kanarazushimo, doumo, douyara, taitei, moshika shitara, taigai*	5 %–10 %
višja	*yohodo (yoppodo), angai, hyotto shitara (hyotto suru to), koto ni yoru to (koto ni yoreba), ookata, sazo, moshika sureba (moshika suru to)*	< 5 %

Tabun, *kanarazu*, *kitto*, *osoraku* se najbolj pogosto pojavljajo v uravnoteženih korpusih JpWaC in KokkenBK. Če je vsaj v enem korpusu relativna frekvenca višja od 10 % (gl. Tabelo 12), je umestitev v osnovno stopnjo smiselna. Prislova *zettai* in *zettai ni* se zelo pogosto pojavljata v korpusu časopisov (11, 12 %), govornem korpusu NUJCC (skupno 52 %) in v belih knjigah (KokkenOC) (14,11 %). Ker sta zelo pogosta v govornem korpusu neformalnega sloga, ju lahko predlagamo kot izbirne enote na osnovni stopnji. *Taitei* in *sazo* imata visoko frekvenco v učbeniških korpusih, medtem ko se v korpusu JpWaC in korpusu publikacij ne pojavljata pogosto. Prislov *taitei* se med vsemi preiskovanimi učbeniki pojavi v dveh na osnovni stopnji, a glede na rezultate korpusne analize ni potrebno, da se ta prislov poučuje na osnovni stopnji. Nasprotno pa prislova *osoraku* v učbenikih skorajda ne najdemo, čeprav se v korpusih pogosto pojavlja in je zato treba razmisliti o njegovi vključitvi v učni načrt za osnovno stopnjo. Za gesla, primerna za srednjo stopnjo, so bila izbrana tista, ki so se vsaj v enem korpusu s splošno vsebino pojavila med 5 % in 10 %. Prislova *kanarazushimo* in *moshika shitara* nista obravnavana v učbenikih, a spadata v drugo najvišjo stopnjo pri seznamu besedišča JLPT. Če upoštevamo

njuno frekvenco v korpusih, je smiselno ju učiti na srednji stopnji. Na višji stopnji poučujemo gesla z nizko frekvenco, pod 5 %, kamor npr. spada prislov *sazo*.

Pri poučevanju, namenjenem specifičnim ciljem, je pomembno, da najprej vključimo gesla, ki se pogosto pojavljajo v specializiranih korpusih, katerih tematika je usklajena z izobraževalnimi cilji, pri čemer je seveda nujno upoštevati tudi ostale dejavnike. Na primer, glede na podatke pričujoče raziskave je pri specializiranem poučevanju na znanstvenem in političnoekonomskem področju smiselno prislov *kanarazushimo* predstaviti že v zgodnjih fazah učenja.

5.4.1.5 Predlogi za izdelavo učnega načrta za besedišče prislovov glede na kolokabilnost med prislovi in modalno obliko na koncu stavka

Iz analize korpusov je postalo jasno, da je skupaj s poučevanjem prislovov smiselno poučevati tudi reprezentativno modalno obliko, ki z njim tvori kolokacijo. Zaporedje in način poučevanja teh kombinacij se ne osredotoča le na modalne oblike, ki najpogosteje tvorijo kolokacije s prislovi, ampak je pomembno razmišljati tudi o pomenski funkciji in težavnostni stopnji te modalne oblike v stavčnem vzorcu. Tabela 27 v razdelku 5.3.2 prikazuje tendenco frekvence in izpostavljenosti kolokacij med prislovom in modalno obliko oz. tipom modalnosti. Iz nje lahko razberemo, da se prislovi najpogosteje vežejo s tipi modalnosti za pričakovanje in nujnost. Tudi statistično izračunan indeks izpostavljenosti, ki izraža moč kolokacij, izkazuje visoko vrednost za ta dva tipa modalnosti. Vse to pomeni, da imata pri komunikaciji v japonskem jeziku modalni obliki za pričakovanje in nujnost večjo funkcionalno prioriteto kot obliki za možnosti in domnevo. Pomembno je upoštevati tudi pomensko funkcijo tovrstne komunikacije pri izdelavi učnega načrta. Z vidika pomenske funkcije modalne oblike, ki tvori kolokacijo s prislovom, obstajajo štiri vrste modalnosti. Za učinkovitost velja, da so modalne oblike vsake vrste obravnavane v kompletu, vendar pa je določene oblike bolje vpeljati na nekoliko višji stopnji znanja, na primer tiste z višjo težavnostjo ter sestavljene oblike.

Vrsti modalnosti, ki najpogosteje tvorita kolokacijo s prislovi, sta pričakovanje in nujnost. K prvi spadajo modalne oblike *darou, darou to omou, no darou, no da to omou, koto darou,* k drugi pa *no da, hazu da, ni chigai nai* in *beki da* (za podrobnosti gl. Tabelo 27). Ker se tudi prislovi, ki tvorijo kolokacije s temi modalnimi oblikami, pojavljajo pogosto, je najbolje, da se kolokacije poučujejo že na osnovni stopnji. Nadalje se razmišlja tudi o zgradbi modalne oblike s strani težavnostne stopnje.

V trenutnem učnem načrtu za osnovno stopnjo so vključene modalne oblike *darou, to omou* in *darou to omou. No darou, no da, koto da, hazu da* in podobne oblike so obravnavane v učbenikih do konca srednje stopnje. Pri uradnem testu za znanje japonskega jezika so umeščene v tretjo stopnjo, tako da se lahko poučujejo v okviru druge polovice osnovne stopnje ali pa po zaključeni osnovni stopnji. Če povzamemo, smiselno je upoštevati funkcionalno pomembnost vrst modalnosti in obravnavanje kompletov tistih modalnih oblik ter poučevati pogoste kolokacije med določenim prislovom in določeno modalno obliko v primeru prostejših vzorcev modalnih oblik nekje med prvo in drugo polovico osnovne stopnje.

Kolokacije, ki se poučujejo na srednji stopnji, so večinoma sestavljene iz modalnih oblik in prislovov, ki se pojavljajo srednje pogosto. Z vidika pomenske funkcije vrste modalnosti je za to stopnjo primerna predstavitev kolokacij prislovov in modalnih oblik, ki izražajo domnevo in možnost. Med modalne oblike za domnevo spadajo *you da, rashii, ki ga suru, no you da, mitai da, sou da* in *you ni omou*. Med modalne oblike za možnost spadajo *kamo shirenai, no kamo shirenai, no kana, no de wa nai ka (to), no ka to omou, no ka, darou to omou* in *ka to omou to*. Med temi modalnimi oblikami so nekatera v trenutnem učnem načrtu kot gesla modalnih oblik tretje stopnje umeščena v osnovno stopnjo (*you da, rashii, kamo shirenai*), vendar prislovov, ki tvorijo kolokacije s temi modalnimi oblikami, tako samostojno kot v kolokacijah, v obstoječem učnem načrtu ne najdemo pogosto. Glede na pomenske funkcije in frekvence modalnih oblik bi bilo treba razmisliti o njihovem poučevanju v kombinaciji s prislovi na srednji stopnji. Poleg tega bi lahko na srednji stopnji vpeljali srednje pogoste kolokacije prislovov in modalnih oblik, ki izražajo nujnost in pričakovanje. Med njimi je *kanarazu*, ki tvori kolokacijo z modalnimi oblikami za nujnost, kot so *to wa kagiranai, wake de wa nai, mono de wa nai, koto de wa nai* in *to ienai*. Ker jih spremlja *nai*, postanejo delno zanikane povedne vsebine, kar lahko obravnavamo kot nenujnost.

Na višji stopnji se poučujejo kolokacije med modalnimi oblikami in prislovi z nizko pojavnostjo in modalne oblike, ki niso vključene v osnovno in srednjo stopnjo, oblike z nizko stopnjo pojavnosti ter sestavljene oblike, kot so *to kangaeru kamo shirenai, mono darou* in *kamo shirenai no da*.

Prislov v določenih primerih spremeni pomen stavka glede na modalnost, s katero tvori kolokacijo. Na primer, prislov *kitto* se lahko veže tako z nujnostno kot tudi s pričakovalno vrsto modalnosti. Zato je pomembno, da za vsako vrsto pokažemo primeren zgled in da učence opozorimo na take primere.

Z uporabo uravnoteženih korpusov, kot sta JpWaC in BCCWJ, lahko predlagamo na več virih zasnovan uravnotežen učni načrt. Pomemben je tudi premislek o variantah modalnih oblik in o njihovih frekvencah, kot prikazuje Tabela 26 v razdelku 5.3.2. Ne le kolokacije z modalnimi oblikami na koncu stavka, raziskati je treba tudi kolokacije z drugimi elementi v stavku, kot so glagol, pridevnik itd.

5.4.2 Predlogi za urejanje slovarja

V tem razdelku rezultate kolokacijskih odnosov med prislovi in modalnimi oblikami na koncu stavka primerjamo z opisom teh podatkov v treh slovarjih japonskega jezika. Omenjeni kolokacijski odnosi so pridobljeni iz uravnoteženega korpusa publikacij in iz korpusa JpWaC. Dva sta japonsko-angleška slovarja, široko uporabljena s strani učencev (*Genius Japanese-English Dictionary, New Century Japanese-English Dictionary*). Tretji je *Slovar slovničih vzorcev* (jp. *Nihongo bunkei jiten,* ang. *Dictionary of Japanese Grammatical Expressions, BJ*) (Group Jamashii, 2005), v katerem so izčrpno popisani stavčni vzorci, ki učencem na višjih stopnjah povzročajo največ težav (Tabela 32).

S stališča razpršenosti prislovov se v dveh japonsko-angleških slovarjih pojavijo vsi prislovi, ki so predmet preučevanja v pričujoči raziskavi. Nasprotno v slovarju stavčnih vzorcev *BJ* nekaterih prislovov ne najdemo. Poleg prislovov *angai, ookata, koto ni yoru*, ki se že tako ne pojavljajo pogosto, pa ni najti niti *zettai* in *zettai ni*, ki sta v korpusih pogosta. V Tabeli 12 lahko denimo vidimo, da sta ta prislova v neformalnih govornih korpusih zelo pogosta. Pojavljata se tudi v korpusu publikacij, korpusu JpWaC in formalnih govornih korpusih, zato je smiselno, da ju vključimo v slovar japonskega jezika za srednjo stopnjo in višje. Nasprotno sta prislova *sazo* in *taigai* zelo pogosta v učbeniških korpusih, v korpusu JpWaC in korpusih publikacij pa se skoraj ne pojavljata. Še posebej *sazo* velja za nekoliko starejši izraz, tako da niti ni nujno, da bi bil vključen v slovar za učence.

Modalne oblike, ki tvorijo kolokacije s prislovi, pogosto zasledimo v stavčnih primerih v obeh japonsko-angleških slovarjih, vendar se modalne oblike določenega prislova v stavčnih primerih razlikujejo v vsakem slovarju, prav tako pa ni dovolj primerov kolokacij prislovov z reprezentativnimi modalnimi oblikami. Na primer, k prislovu *tabun* je namesto *kamoshirani* in *mono da* bolje dodati *no darou* ali *to omou*, k prislovu *osoraku* pa je dobro dodati *to omou* ali *ni chigai nai*.

Tabela 32: *Kolokacije med prislovi in modalnimi oblikami na koncu stavka v japonskih slovarjih.*

PRISLOV	GENIOUS	NEW CENTURY	SLOVAR STAVČNIH VZORCEV
aruiwa	▲kamo shirenai	▲kamo shirenai	▲kamo shirenai ●no darou, to omowareru
angai	△	▲no kamo shirenai	/
osoraku	▲deshou, no darou, darou	△	▲darou, mono to omowareru, ni chigai nai
ookata	△	△	/
kanarazu	▲deshou, nasai, darou	▲deshou, te kudasai, hitsuyou ga aru, koto ni shite iru	▲te kudasai, shinakereba naranai, you ni shiyou
kanarazushimo	▲to wa kagiranai, to iu wake de wa nai, mono de wa nai	▲to wa kagiranai, to wa ienai	▲to wa kagiranai, de wa nai to watashi wa omotte iru, shinakereba naranu mono da to omotte iru wake de wa nai ●wake de wa nai, to wa kagiranai
kitto	▲no desu, ni chigai nai, n'da, darou	▲darou, ni chigai nai, deshou	▲deshou, darou, ni chigai nai, te kudasai
koto ni yoru to	△	▲kamo shirenai	/
sazo (kashi)	▲deshou	▲deshou, koto deshou	▲koto darou, koto de gozaimashou
taigai	▲darou	△	koto ni shite iru, koto ni natte iru
taitei	▲darou	▲de arou	Enaki stavčni primeri in razlaga kot za *taigai*.
tabun	▲darou, kamo shirenai, deshou	▲deshou, darou, kamo shirenai, mono da	▲deshou, te okou, deshou ka, to omou
doumo	▲you da, you ni omou	▲sou da, you da, n'da, ki ga shinai	sou da, you da, rashii
moshika shitara	▲kamo shirenai	▲kamo shirenai	▲kamo shirenai, no de wa nai darou ka
hyotto shitara	▲darou, kamo shirenai, deshou ka	▲no kamo shirenai, ka to omou	/
yohodo	△	▲you da	▲ni chigai nai, rashii, n'darou, n'da, n'darou to omou
zettai	△	△	/
zettai ni	▲shite wa ikenai, suruna	fukanou da, koto de wa nai	/

Legenda: △ prislov je v slovarju, modalna oblika, s katero je v kolokaciji, se ne pojavi ▲ modalna oblika je v stavčnem primeru ● modalna oblika je omenjena v drugem vzorcu

V slovarju stavčnih vzorcev BJ se pri vseh geslih obravnavanih prislovov pojavljajo modalne oblike, vendar lahko tudi za potrebe tega slovarja uporabimo podatke modalnih oblik z visoko frekvenco ali izpostavljenostjo v korpusu. Kar se tiče prislova *osoraku*, je k njemu namesto *mono to omowareru* bolje postaviti modalne oblike *no darou, koto darou* in *to omou*. Ker se pri geslu *kanarazu* ne pojavi domnevna modalna oblika, je bolje dodati *hazu da, no da, darou* ali *ni chigai nai*. In ker se prislova *taigai* in *taitei* razlikujeta v rabi in frekvenci, ni primerno, da v slovarju za ti dve gesli uporabimo enaki stavčni primeri in razlage.

Omenjeni predlogi temeljijo na podatkih, ki so v precejšni meri uravnoteženi oz. ne kažejo večjih odstopanj. Iz tega razloga jih lahko uporabimo za izdelavo slovarja, namenjenega opisu ali učenju splošne japonščine. Ker se zavedamo, da se reprezentativne modalne oblike razlikujejo glede na vrsto korpusa, je tudi pri slovarju za učenje japonščine pomembno, da odraža razporejene podatke glede na zvrst besedila.

5.5 Povzetek

V tem poglavju smo najprej raziskali rezultate predhodnih raziskav o prislovih in modalnih oblikah na koncu stavka in pokazali, da primanjkuje empiričnih podatkov o reprezentativnih modalnih oblikah in da nimamo izčrpnega seznama modalnih oblik. Japonski različici sistema za luščenje kolokacij smo dodali funkcijo za iskanje oddaljenih kolokacijskih odnosov med prislovi in modalnimi oblikami na koncu stavka. Proces je potekal v naslednjih štirih korakih.

1. Izdelava seznama modalnih oblik in njihovih variant.
2. Določanje morfemske analize modalnih oblik v orodju ChaSen ter izdelava nove oznake za modalne oblike.
3. Ponovno označevanje korpusa z morfološko analizo orodja ChaSen in dodelitev oznak modalnim oblikam.
4. Dodajanje pravil, ki določajo kolokacijske odnose med prislovi in modalnimi oblikami, v slovnično datoteko besednih skic.

V nadaljevanju smo ovrednotili kolokacije med prislovi in modalnimi oblikami na koncu stavka, ki so bile pridobljene s pomočjo besednih skic iz spletnega korpusa in pokazali visoko natančnost izluščenih podatkov. Rezultate smo primerjali s Kudōjevo raziskavo o prislovih in razkrili razlike ter podobnosti. Nadalje smo opravili primerjavo z uravnoteženim korpusom publikacij ter tudi v tem segmentu prikazali podobnosti in razlike. V obeh korpusih je bilo veliko skupnih točk v

rezultatih preiskovanih kolokacijskih odnosov, zato smo lahko opazovali tendenco kolokacij s podobnimi modalnimi oblikami. Poleg tega smo rezultate pridobljenih kolokacij med prislovi in modalnimi oblikami analizirali in primerjali v različnih tipih korpusov.

V sklepnem delu smo razpravljali o tem, kako bi lahko izluščene rezultate aplicirali na poučevanje japonskega jezika. Najprej smo z uporabo različnih korpusov iskali način, kako izpopolniti učni načrt, in predlagali postopek za njegovo izdelavo. Ugotovili smo, da se razpršenost prislovov in tendenca kolokacijskih odnosov med prislovi in modalnimi oblikami glede na vrsto/tip korpusa razlikujejo. Zaradi tega je tudi pri izdelavi učnega načrta za besedišče zaželena gradivna osnova, ki temelji na tipu korpusa, ki se sklada s cilji učnega načrta. Glede na vrednost izpostavljenosti in frekvenco v primernih korpusih ter upoštevajoč druge faktorje smo predlagali postopek uvajanja prislovov v učni načrt po prioriteti. Da bi empirično podprli primernost učnega načrta, smo z vrednotenjem učbenikov potrdili predvsem učinkovitost rezultatov o pogostosti kolokacij v uravnoteženih korpusih.

Izluščene rezultate smo primerjali s slovarji japonskega jezika ter razpravljali, v kolikšni meri so v njih obravnavane informacije o kolokacijskih odnosih med prislovi in modalnimi oblikami. Obravnavani slovarji so v številnih primerih zajemali kolokacijski odnos med prislovom in modalno obliko, a pokazalo se je tudi nekaj pomanjkljivosti in nedoslednosti v leksikalnem opisu. Na temeljih korpusnih rezultatov smo podali predloge za izpopolnitev japonskih slovarjev.

Izzivi za prihodnost zajemajo objavo seznama modalnih oblik in variant v XML-obliki ter nadaljno nadgradnjo seznama modalnih oblikah, zasnovano na različnih tipih korpusov, ne samo na podatkih spletnega korpusa. Poleg tega stremimo k dodajanju različnih korpusov v orodje SkE, primerjavi podatkov kolokacij med prislovi in modalno obliko na koncu stavka glede na zvrst ter širitev rabe pri poučevanju japonskega jezika.

6 Zaključek in nadaljnja vprašanja

6.1 Povzetek rezultatov in splošne opazke

V tem poglavju omenimo rezultate pričujoče raziskave in povzamemo vsebino posameznih poglavij v monografiji.

Prvo poglavje (Empirični in aplikativni pristop v analizi kolokacij) oriše razvoj korpusnega jezikoslovja na Japonskem in širše ter opiše uporabo korpusov pri poučevanju tujih jezikov. Poseben poudarek je namenjen pomembnosti učenja kolokacij za učence tujih jezikov. V tem poglavju so opisane raziskave tudi o luščenju kolokacij iz obsežnega korpusa ter o podpori pri učenju kolokacij in obstoječa učna gradiva. Posvetimo se pomanjkljivosti učnega gradiva, ki obravnava kolokacije, za učence japonskega jezika. Na koncu je orisano ozadje raziskav o oddaljenih kolokacijskih odnosih med prislovi in modalnimi oblikami na koncu stavka v japonskem jeziku.

Drugo poglavje (Razvoj japonske različice orodja za luščenje kolokacij) opiše pripravo japonske različice sistema SkE za iskanje kolokacij. Za ta namen je bil uporabljen spletni korpus JpWaC, izdelana in uporabljena je bila tudi japonska slovnična datoteka, temelječa na besednih vrstah orodja ChaSen in regularnih izrazih. Rezultat je omogočil luščenje izčrpnega in sistematičnega sežetka jezikoslovnih informacij, kot so različni kolokacijski in slovnični odnosi med besedami in morfemi. Tovrstne informacije lahko pridobimo iz obsežnega korpusa japonščine z veliko hitrostjo in natančnostjo. V slovnični datoteki je zajetih 22 vzorčnih pravil kolokacijskih odnosov, ki zajemajo več kot 50 različnih kolokacijskih odnosov z glagoli, samostalniki, pridevniki na *-i* in *-na* ter prislovi. Pri gradnji obsežnega spletnega korpusa smo težili k čim bolj uravnoteženim podatkom, pri primerjavi s časopisnimi podatki pa se je pokazalo, da je pri korpusu JpWaC res manj odstopanj. Nadalje se je pokazalo, da je iz korpusa izluščene pogoste jezikovne podatke smiselno uporabiti pri poučevanju japonskega jezika.

Tretje poglavje (Aplikacija in vrednotenje japonske različice sistema za luščenje kolokacij) se ukvarja s tem, kako japonsko različico pripomočka SkE, razvito za luščenje kolokacij v japonskem jeziku, uporabiti pri raziskavah, poučevanju in učnih gradivih v japonskem jeziku. Posebna pozornost je namenjena aplikaciji za izdelavo slovarjev za učence japonščine. Vrednotenje je bilo izvedeno z dveh vidikov. Prvi zajema primerjanje rezultatov, pridobljenih z japonsko verzijo besednih skic v orodju SkE, z obstoječim slovarjem kolokacij japonskega jezika. Da bi odkrili, ali je uporaba rezultatov iskanja kolokacij, pridobljenih z japonsko verzijo sistema,

primerna za slovar za učence japonskega jezika ali ne, smo rezultate primerjali z naključno izbranimi primeri kolokacij iz *Slovarja uporabe japonskih izrazov v praksi* (Himeno, 2004). Odkrili smo, da bi bilo pri izdelavi prihodnjega slovarja kolokacij, za čim boljšo pokritost podatkov, pomembno vključiti tudi sistem za luščenje kolokacij. Drugo stališče je bilo vrednotenje posameznih gesel s strani strokovnjakov za poučevanje japonščine. Strokovnjaki so – z mislijo na razvoj obsežnega slovarja japonskih kolokacij – ocenjevali, ali so z japonsko verzijo sistema dobljeni rezultati kolokacij primerni za vključitev v slovar ali ne. Ker delež kolokacij, ki so ovrednotena z »dobro« in »morda«, zajema 68 % in 18 %, delež kolokacij, ovrednoten z »slabo« pa 11 %, lahko sklenemo, da je japonska različica SkE učinkovit leksikografski pripomoček.

Četrto poglavje (Kolokacijski odnos na daljavo med prislovi in modalno obliko na koncu stavka) opiše pojav oddaljene kolokacije, ko se dve besedi ali beseda in stavčni vzorec istočasno nahajata na določeni razdalji. Osredotoča se na pojav, ki se tradicionalno obravnava kot strukturna odvisnost in ujemanje, in ga analizira v številnih korpusih japonskega jezika kot kolokacije med prislovi in modalnimi oblikami na koncu stavka. Kot rezultat uporabe različnih pripomočkov in metod smo za vsak prislov razjasnili, s katero modalnostjo na koncu stavka tipično tvori kolokacijo. Pokazali smo tudi, da se tendenca spreminja glede na vrsto korpusa. Razpršenost prislovov odraža vrsto besedil v korpusih, na podlagi tega podatka pa smo lahko klasificirali korpuse in njihove registre. Ročno pridobljeni rezultati korpusne analize so pomagali pri izvajanju raziskave, opisane v naslednjem poglavju.

Peto poglavje (Luščenje kolokacij na daljavo med prislovi in modalnimi oblikami na koncu stavka ter predlog za uporabo v učnem gradivu za japonščino) opiše pripravo velikega števila modalnih oblik in njihovih variant. Ker modalne oblike in njihove variante niso bile prepoznane v orodjih za obdelavo naravnih jezikov, kakršno je ChaSen, in so bile razdeljene na več posameznih morfemov, je bila med raziskavo na novo dodeljena oznaka za modalne oblike, čemur je sledilo ponovno označevanje korpusa in manjša sprememba slovnične datoteke. S tem je postalo mogoče iskanje (oddaljenih) kolokacij med prislovi in modalnostjo na koncu stavka v vzorčnem korpusu s pomočjo orodja SkE in japonskih besednih skic. Prikazano je bilo visoko vrednotenje (93 %–96 %) natančnosti izluščenih kolokacij med prislovi in modalnimi oblikami na koncu stavka. Nadalje smo dobljene rezultate primerjali z uravnoteženim, obsežnim japonskim korpusom. Porazdelitve prislovov in kolokacije med prislovi in modalnimi oblikami na koncu stavka so imele enako tendenco kot spletni korpusi, tako da je bila znova potrjena verodostojnost spletnih podatkov.

Na primerih kolokacij med prislovi in modalnimi oblikami na koncu stavka smo predlagali izdelavo učnega načrta za besedišče japonščine, osnovanega na različnih korpusih. Rezultati raziskovanja pojavnosti kolokacij med prislovi modalnimi oblikami na koncu stavka v japonskih učbenikih so pokazali, da te kolokacije niso obravnavane sistematično in da je prilagoditev sedanjega učnega načrta japonščine nujna. Poleg tega smo na podlagi rezultatov primerjave izluščenih kolokacij med prislovi in modalnimi oblikami na koncu stavka z japonskimi slovarji pokazali, da se obravnava kolokacij med slovarji razlikuje. Obstajajo številni primeri, v katerih je reprezentativna modalna oblika, ki tvori kolokacijo z določenim prislovom, odsotna. Zato smo prišli do zaključka, da je izpopolnitev japonskih slovarjev nujna, seveda s pomočjo obsežnih korpusov ter sistemov za luščenje kolokacij.

6.2 Nadaljnja vprašanja

V pričujoči raziskavi razvita japonska verzija sistema za luščenje kolokacij pokriva 50 tipov različnih kolokacijskih. Nadaljnji izziv je postopno povečevanje kolokacijskih odnosov in njihova izpopolnitev. Za premostitev šibkih točk orodja za morfemsko analizo ChaSen je nujno idiomom ter prepodrobno razčlenjenim morfemom ponovno dodati oznako za eno enoto.

Eden od nadaljnjih izzivov je prav tako analiza drugih orodij za morfološko analizo. Zelo pomemben napredek je že narejen z uporabo daljših in krajših enot v slovarju UniDic, ki je bil razvit v okviru projekta Korpus japonščine za označevanje japonskega korpusa BCCWJ (Maekawa et al., 2013). Elekronski slovar UniDic je z morfološkim analizatorjem MeCab že bil uporabljen pri označevanju novega in večjega japonskega korpusa JpTenTen (za podrobnosti glej Srdanović et al. 2013). Vendarle pa ta nova orodja še vedno ne ponujajo potrebnih oznak za modalne oblike in bi bilo treba raziskavo, opisano v 5. poglavju, ponoviti in jo aplicirati na nove oznake novih orodij za morfološko analizo. Da bi sistem luščenja kolokacij uspešno uporabljali pri poučevanju japonščine, je treba dodati oznake kolokacij in stavčnih primerov glede na stopnjo učnih sposobnosti ter omogočiti uporabo furigane za pridobljene podatke v sistemu.

V pričujoči raziskavi je bila z uporabo obsežnih korpusov in rezultatov analize kolokacijskih odnosov v japonščini predstavljena izdelava učnega načrta za besedišče, kot tudi predlogi za izpopolnitev japonskih slovarjev. Tu je bila posebna pozornost namenjena kolokacijam med prislovi in modalnimi oblikami na koncu stavka. Za učence tujih jezikov je obvladovanje kolokacijskih odnosov zelo pomembno, zaradi česar je v prihodnje treba razmisliti, katere kolokacije in na katerih stopnjah se

bodo poučevale v okviru japonskega jezika. Sestavljanje slovarja kolokacij za učence japonskega jezika ter izdelava učnih gradiv za podporo pri usvajanju kolokacij na osnovi takih analiz je pomembna naloga.

Predlogi za poučevanje japonščine v pričujoči raziskavi temeljijo na obsežnem spletnem korpusu in uravnoteženem korpusu BCCWJ. Postalo pa je jasno, da se razpršenost prislovov in tendenca kolokacij med prislovi in modalnimi oblikami razlikujeta glede na vrsto korpusa. Zato je pomembno, da se v okviru poučevanja japonščine in izdelave slovarja ta ugotovitev upošteva in se prilagodi učnim ciljem.

V pričujoči raziskavi smo kot primer kolokacijskega odnosa med prislovi in modalno obliko na koncu stavka opazovali tudi pojav oddaljene kolokacije. Razdalja med enotama kolokacije se sicer od primera do primera razlikuje, med nekaterimi enotami pa je lahko razmak precejšen. Izhajali smo iz predpostavke, da se ta tendenca spreminja glede na prislov. Pojav kolokacij na daljavo je tudi v drugih jezikih v veliki meri neraziskan, zato moramo preučiti, v kolikšni meri in na kakšen način se pojavlja z vidika splošnih lastnosti jezika.

V raziskavi smo uporabili različne korpuse in sistem za luščenje kolokacij ter prikazali možnosti aplikacije pri poučevanju japonščine in v leksikografiji. Vendar pa v svetu korpusnega jezikoslovja obstajajo tudi jezikovna sredstva, ki se ne uporabljajo pogosto. Z anketo pridobljena podatkovna zbirka besednih zvez oz. asociacij (Joyce, 2005) je dragocen vir podatkov v kognitivnem jezikoslovju, ki ga lahko apliciramo na izdelavo slovarja ali poučevanje jezika. V poskusni raziskavi smo primerjali kolokacije, izluščene iz obsežnih zbirk besedil v orodju SkE, s predlogi asociacij, kombinacij dveh besed, ki so jih podali ljudje glede na svojo intuicijo in so nato bili zbrani v bazo podatkov in kvantitativno ovrednodeni, ter ugotavljali skupne točke in razlike v rezultatih teh dveh pristopov (Joyce in Srdanović 2008). Podatki v korpusih so neprimerno obsežnejši kot v podatkovnih zbirkah asociacij, a v nekaj primerih kombinacije, ki se je pojavljala kot asociacija, v izluščenih rezultatih iz korpusa ni bilo mogoče najti. V prihodnje bi veljalo primerjati kolokacije, pridobljene s tema pristopoma, in še podrobneje primerjati ter raziskati razlike. Tovrstna raziskava bi lahko s teoretičnega in praktičnega vidika osvetlila razlike in podobnosti med pojavoma asociacij in kolokacij, uporabna pa bi bila tudi pri izboljšavi sodobnih orodjih za luščenje kolokacij s predlogi o novih, korpusno še neobdelanih tipih odnosov med besedami ter izdelavi učnih pripomočkov za učence japonskega jezika.

Viri in literatura

Akama, H., Miyake, M. & Jung, J. (2008). A New Evaluation Method for Graph Clustering of Semantic Networks Built from Lexical Co-occurrence Information, 18th International Congress of Linguists (CIL 18). *The 18th International Congress of Linguists. Abstracts, Volume 1*, 206–207.

Akano, I. (2006). Angleško korpusno jezikoslovje in poučevanje angleščine. [Eigo koopasu gengogaku to eigo kyouiku.] *Nihongo kyouiku (tokushuu) koopasu to nihongo kyouiku – genjou to kadai*, 130, 11–21.

Atkins, S., Charles J. F. & Christopher R. J.(2003). Lexicographic Relevance: Selecting Information from Corpus Evidence. *International Journal of Lexicography,* 16(3). Oxford, UK: Oxford University Press. 251–280.

Backhouse, A. E. (2004). *Collocational aspects of near-synonyms: Illustrations from a small corpus (Nihongo gakushuu jishohensan ni muketa denshi-ka koopasu riyou ni yoru korokeeshon kenkyuu)*. 89–102.

Baroni, M. & Bernardini, S. (2004). BootCat: Bootstrapping corpora and terms from the web. *Proceedings of the Fourth Language Resources and Evaluation Conference, LREC2004*. Lisbon, Portugal.

Baroni, M. & Bernardini, S., eds. (2006). *Wacky! Working papers on the Web as Corpus*. Bologna, Italy: GEDIT.

Baroni, M. & Kilgarriff, A. (2006). Large linguistically-processed Web corpora for multiple languages. *Proceedings EACL*. Trento, Italy.

Baroni, M., Kilgarriff, A., Pomikalek, J. & Rychly, P. (2006). WebBootCaT: a web tool for instant corpora. *Proceedings of the EuraLex Conference 2006*, 123–132.

Baroni, M., Kilgarriff, A., Pomikalek, J. & Rychly, P. (2006). WebBootCaT: instant domain specific corpora to support human translators. *Proceedings of EAMT 2006,* 47–252. Oslo, Norway.

Baugh, S., A. Harley, S. Jellis (1996). The Role of Corpora in Compiling the Cambridge International Dictionary of English. *International Journal of Corpus Linguistics,* 1(1), 39–59.

Bekeš, A. (2006). Japanese suppositional adverbs in speaker-hearer interaction. *Proc. of the 3rd conference on Japanese language and Japanese language teaching,* 34–48. Venezia, Italy: Cafoscarina.

Bekeš, A. (2008a). Japanese suppositional adverbs: probability and structure in speaker-hearer interaction. *Linguistica*, 48, 277–292. Ljubljana, Slovenia.

Bekeš, A. (2008b). Text and boundary: a sideways glance at textual phenomena in Japanese. *Zbirka Razprave FF 1. izd*. Ljubljana, Slovenia: Znanstvena založba Filozofske fakultete.

Biber, D. (1988). *Variation across speech and writing*. Cambridge, UK: Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation. A Cross-Linguistic Comparison*. Cambridge, UK: Cambridge University Press.

Cao, H. & Nishina, K. (2006). Priporočila za poučevanje kolokacij z vidika esejističnih napak pri tujih učencih. [Gaikokujin gakushuusha no sakubun goyou rei kara miru kyouki hyougen no shuutoku oyobi kyouiku he no taigen – Meishi to keiyoushi oyobi keiyoudoushi no kyouki hyougen ni tsuite.] *Nihongo kyouiku (tokushuu) koopasu to nihongo kyouiku*, 130, 70–79.

Chantree, F., de Roeck, A., Kilgarriff, A. & Willis, A. (2005). *Disambiguating Coordinations Using Word Distribution Information*. In Proceedings RANLP Bulgaria.

Chen, A., Rychly, P., Huang, C.-R., Kilgarriff, A. & Smith, S. (2007). A corpus query tool for SLA: learning Mandarin with the help of Sketch Engine. *Practical Applications of Language Corpora (PALC)*. Lodz, Poland.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. *COMPLEX« 94*. Budapest.

Church, K. W. & Hanks, P. (1989). Word association norms, mutual information, and lexicography. *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, 76–83.

Church, K. W. & Hanks, P. (1990). Word Association Norms, Mutual Information, and. Lexicography. *Computational Linguistics,* 16(1), 22–29.

Cook, G. (1998). The uses of reality: A Reply to Ronald Carter. *ELT Journal, 52(1)*, 57–63.

Crystal, D. (2006). *Language and the Internet*. Cambridge, UK: Cambridge University Press.

De Cock, S., Granger, S., Leech, G. & McEnery, T. (1998). *An automated approach to the phrasicon of EFL learners*. Granger, S. (ed.) Learner English on Computer. London and New York: Addison Wesley Longman, 67–79.

Den, Y., Kogiso T., Ogura, H., Yamada, A., Minematsu N., Uchimoto K. & Koiso, H. (2007). Jezikovni viri za japonsko korpusno jezikoslovje: Razvoj in aplikacije morfoloških analiz za elektronske slovarje. [Koopasu nihongogaku no tame no gengo shigen: Keitaiso kaisekiyou denshika jisho no kaihatsu to sono ouyou.]. *Nihongo kagaku*, 22, 101–123.

EDR (1994). *EDR Collocation dictionary Technical Report TR-043*. Japan Electronic Dictionary Research Institute, 1994.

Ellis, N. C. (2001). *Memory for language.* P. Robinson (Ed.) Cognition and second language instruction. Cambridge, UK: Cambridge University Press, 33–68.

Erjavec, T., Hmeljak, S. K. & Srdanović, E. I. (2006). JaSlo, A Japanese-Slovene Learners« Dictionary: Methods for Dictionary Enhancement. *Proceedings of the 12th EURALEX International Congress.*

Erjavec, T., Kilgarriff, A. & Srdanović, E. I. (2007). *A large public-access Japanese corpus and its query tool, CoJaS 2007*. The Inaugural Workshop on Computational Japanese Studies.

Erman, B. & Warren B. (2000). *»The idiom principle and the open choice principle« and »the idiom principle«.* Text: An interdisciplinary journal for the study of discourse.

Fillmore, C. J., Johnson, C. R. & Petruck, M. R. L. (2003). Background to FrameNet. *International Journal of Lexicography,* 16(3), 235–250. Oxford, UK: Oxford University Press.

Firth, J. R. (1951). Papers in Linguistics. Oxford University Press.

Fujii, S. & Ohara, K. (2003). Semantični okvir in FrameNet. [Fureemu imiron to fureemunetto.]. *Eigo seinen*, 149(6), 373–376.

Fujiike, Y. (2008). *Pregled dolžine enot korpusa BCCWJ. [Gendai nihongo kakikotoba kinkou koopasu ni okeru choutani no gaiyou.]*. Tokutei ryouiki kenkyuu „Nihongo koopasu" heisei 19 nendo koukai waakushoppu, 51–58.

Fukada, A. (2008). Aplikacija na področje raziskovanja in poučevanja japonščine v korpusnem jezikoslovju. [Koopasu gengogaku no nihongo kenkyuu to nihongo kyouiku he no ouyou.]. *Fifteenth Princeton Japanese Pedagogy Forum Proceedings*, 1–18.

Fukada, A. & Oso, M. (2007). Vsakodnevni pogovori z vidika sistema *chakoshi*. [Chakoshi de miru nichijou kaiwa.]. *Proceedings from the Fourth International Ceonference on Computer Assisted Systems for teaching and learning/Japanese (CASTEL/J)*, 125–128.

Fukada, A. (2007). Sistem *chakoshi* za luščenje kolokacijskih podatkov in primerov v japonščini. [Nihongo yourei to korokeeshon jouhou chuushutsu shisutemu „chakoshi".]. *Nihongo kagaku,* 22, 161–172.

Gahl, S. (1998). Automatic Extraction of subcategorization frames for corpus-based dictionary-building. *Proc. EURALEX 1998*, 445–452.

Gatt, A. & van Deemter, K. (2006). Conceptual coherence in the generation of referring expressions. *Proceedings of the COLING-ACL 2006 Main Conference Poster Session*.

Genius Japanese-English Dictionary [*Genius Waei Jiten*] (1994), 2. izd. Tokyo: Taishukan.

Ghani, R., Jones, R. & Mladenic, D. (2001). Using the Web to Create Minority Language Corpora. *Proceedings of the 2001 ACM CIKM: Tenth International Conference on Information and Knowledge Management*, 279–286.

Gillard, P. & Gadsby, A. (1998). Using a learners« corpus in compiling ELT dictionaries. S. Granger (ed.) Learner English on Computer, 159–171. London/New York: Longman.

Gotou, H. (2003). Jezikovna teorija in jezikovni materiali – korpusni in nekorpusni podatki. [Gengo riron to gengo shiryou – Koopasu to koopasu igai no deeta.]. *Nihongogaku, 22*, 6–15.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: collocations and formulae. *Cowie A. (ed.) Phraseology: theory, analysis and applications*, 145–160. Oxford, UK: OUP.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer.

Group Jamashii. (2005). *Slovar slovničnih vzorcev. [Nihongo bunkei jiten./ Dictionary of Japanese Grammatical Expressions.]*. Tokyo: Kuroshio shuppan.

Halliday, M.A.K. (1966). Lexis as a Linguistic Level. In Bazell et al. (eds.) *In memory of J. R. Firth.* Longman. 148–162.

Halliday, M. A. K. (1978). *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. London, UK: Edward Arnold.

Halliday, M. A. K. (1991). *Corpus studies and probabilistic grammar*. Aijmer, Karin & Bengt Altenberg (eds.), 30–43.

Hashimoto, N. (2008). Poskus izdelave seznama japonskih besed z uporabo korpusa. [Koopasu wo riyoushita nihongo kyouiku no tame no goi risuto sakusei no kokoromi.]. *Daihyousei wo riyousuru kakikotoba koopasu wo riyoushita nihongo kyouiku kenkyuu*, 137–139.

Hashimoto, N. (2008). Temeljne raziskave za izdelavo seznama japonskih besed za z uporabo korpusa. [Koopasu wo riyoushita nihongo kyouiku goi risuto sakusei no tame no kiso kenkyuu.]. *Daihyousei wo yuusuru kakikotoba koopasu wo riyoushita nihongo kyouiku kenkyuu*.

Heid, U., Evert, S., Docherty, V., Worsch, W. & Wermke, M. (2000). Computational tools for semi-automatic corpus-based updating of dictionaries. *EURALEX 2000 Proceedings*, 183–196.

Heylighen, F. (2002). Variation in the Contextuality of Language: An Empirical Measure. *Foundations of Science,* 7(3), 293–340.

Himeno, M. (2004). Slovar uporabe japonskih izrazov v praksi. [Nihongo hyougen katsuyou jiten.]. *Kenkyusha*.

Himeno, M. (ed.) (2012). *Kenkyusha Japanese Collocation Dictionary (Kenkyuusha Nihongo Korokeeshon Jiten)*. Tokyo, Japan: Kenkyusha.

Hirata, M. (2001). Pomen besede *kamo shirenai* – Raziskovanje modalnosti in pragmatičnega stika. [Kamo shirenai no imi – modariti to goyouron no setten wo saguru.]. *Nihongo kyouiku*, 108.

Hirotani, M. (2007). Verifikacija učinka uporabe korpusa CMC. [CMCkoopasu riyou no kouka kenshou.]. *Proceedings from the Fourth International Conference on computer Assisted Systems for teaching and learning/Japanese (CASTEL/J)*, 177–180.

Hodošček, B. & Nishina, K. (2012). Japanese Learning Support Systems: Hinoki Project Report. *Acta Linguistica Asiatica*, 2(3), 95–124. Ljubljana, Slovenia: Ljubljana University Press.

Horby, A. S. (1954). *A Guide to Patterns and Usage in English*. London, UK: Oxford University Press.

Hornby, A. S., ed. (1942). *Idiomatic and Syntactic English Dictionary*. Kaitakusha. (http://www.lang.nagoya-u.ac.jp/nichigen/0-kyouiku/seminar/2007xian/takizawa.pdf)

Howatt, A. P. R. (1984). *A History of English Language Teaching*. Oxford: Oxford University Press.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge, UK: Cambridge University Press.

Ikehara, S., Shirai, S. & Uchino, H. (1996). A Statistical Method for Extracting Uninterrupted and Interrupted Collocations from Very Large Corpora. *Proceedings of 16th International Conference on Computational Linguistics (COLING-96)*, 574–579.

Inoue, E. & Akano I. (eds.) (2003). *Modrosten angleško-japonski slovar. [Uizudamu Eiwa jiten./ The Wisdom English-Japanese Dictionary.]*. Sanseidou.

Inoue, E. (2003). *Leksikografija, ki temelji na korpusih. Angleško korpusno jezikoslovje: Osnova in praksa. [Koopasu ni motozuku jisho henshuu. Eigo koopasu gengogaku: Kiso to jissen.]*, 10. poglavje.

Ishikawa, S. (2008). *Korpus angleščine in jezikovno izobraževanje. [Eigo koopasu to gengo kyouiku.]*. Taishuukanshoten.

Ishikawa, S. (2009). Zmožnost nekontroliranja, kontroliranja in posredovanja v osnovnem raziskovanju japonščine – Web As Corpus. [Nihongo kiso kenkyuu ni okeru hitouseigata, touseigata, baikaigata Web as Corpus no kanousei – kengo koopasu ni okeru kihongo hindo no anteisei ni tsuite.]. *Tokutei ryouiki kenkyuu „Nihongo koopasu" Heisei 20 nendo koukai waakushoppu sateraitosesshon yokoushuu*, 29–38.

Itou, M. (2002). *Uvod v računsko jezikoslovje. [Keiryou gengogaku nyuumon.]*. Taishuukanshoten.

Itou, M. (2003). Korpusi in statistika. [Koopasu to toukei.]. *Nihongogaku, April, 22*, Rinji zookango Koopasu gengogaku, 26–35.

JACET Materials Research Commitee (ur.) (1993). JACET 4000 osnovnih besed. [JACET kihongo 4000.]. Daigaku Eigo kyouiku gakkai.

JACET Materials Research Commitee (ur.) (2003). JACET 8000 osnovnih besed. [JACET list of 8000 Basic Words.]. Daigaku Eigo kyouiku gakkai.

James, C. (1998). *Errors in language learning and use*. London, UK: Longman.

Johns, T. F. (1991). Should you be Persuaded: Two Samples of Data-Driven Learning Materials. *Johns, T. F. and King, P. (eds.) 'Classroom Concordancing' Birmingham University English Language Research Journal,* 4, 1–13.

Joyce, T. (2005). Constructing a large-scale database of Japanese word associations. *Katsuo Tamaoka (ed.) Corpus Studies on Japanese Kanji (Glottometrics 10)*, 82–98. Tokyo, Japan: Hituzi Syobo & Germany: RAM-Verlag: Ludenschied.

Joyce, T. & Srdanović, E. I. (2008). Comparing Lexical Relationships Observed within Japanese Collocation Data and Japanese Word Association Norms. Coling 2008, 22nd International Conference on Computational Linguistics. *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*. 24 August 2008, Manchester, UK, 1–8.

Kameya, Y. & Sato, T. (2005). Computation of probabilistic relationship between concepts and their attributes using a statistical analysis of Japanese corpora. *Proceedings of Symposium on Large-scale Knowledge Resources (LKR2005).*

Kawahara, D. and Kurohashi, S. (2006). Case Frame Compilation from the Web using High-Performance Computing. *Proceedings LREC*. Genoa, Italy.

Kawahara, D. and Kurohashi S. (2006). *Izdelava obsežnega sklonskega okvirja prek spleta z uporabo učinkovitega izračuna. [Kouseino keisan kankyou wo mochiita Web kara no oukibo kaku fureemu kouchiku], Jouhoushorigakkai shizen gengo shori kenkyuukai,* 171(12), 67–73.

Kawamura, Y. (1993). Analiza besedila za bralno razumevanje z uporabo jezikovnega pregleda. [Goi chekkaa wo mochiita dokkai tekisuto no bunseki.]. *Kooza nihongo kyouiku*, 34, 1–22.

Keller, F. & Lapata, M. (2003). Using the Web to Obtain Frequencies for Unseen Bigrams. *Computational Linguistics,* 29(3), 459–484.

Kida, A., Takanashi K., Inui Y. & Isahara, H. (2002). Analiza lastnosti japonščine, ki omogočajo predvidevanje stavčne strukture. [Bunkouzou no zenshinteki yosoko wo kanounisuru nihongo no shotokuchou no bunseki.]. *Gengo shori gakkai dai 8 kai nenji taikai happyou ronbunshuu*, 65–68.

Kudō, H. (2000). *Prislovi in tipi stavkov: Modalnost v japonski slovnici. [Fukushi to bun no chinjutsu no taipu. Nihongo no bunpou 3 modariti.]*. (Moriyama, Takurou, Nitta Yoshio, Kudō, Hiroshi). Iwanamishoten, 161–234.

Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the Special Issue on Web as Corpus. *Computational Linguistics,* 29(3).

Kilgarriff, A. & Rundell, M. (2002). Lexical Profiling Software and its Lexicographic Applications - a Case Study. *EURALEX 2002 Proceedings*, 807–818.

Kilgarriff, A. & Tugwell, D. (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. *Proc. workshop „COLLOCATION: Computational Extraction, Analysis and Exploitation. 39th ACL & 10th EACL,* 32–38.

Kilgarriff, A. (2001). Comparing Corpora. *International Journal of Corpus Linguistics,* 6(1), 1–37.

Kilgarriff, A., Rychly, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Proc. Euralex*, 105–116.

Kilgarriff, A., Kovář, V., Krek, S., Srdanović, I. & Tiberius, C. (2010). A Quantitative Evaluation of Word Sketches. *The 14th EURALEX International Congress*.

Kindaichi, H. (2006). Japonski slovar kolokacij, ki ga želimo poznati. [Shitte okitai nihongo korokeeshon jiten]. *Gakushuu kenkyuusha*.

Kintou. (2003). Korpus klasičnega jezika. [Kotengo no koopasu]. Nihongogaku, 22 (special issue: *Koopasu gengogkau),* 62–81.

Kjellmer, G. (1991). A mint of phrases. *K. Ajimer and B. Altenberg (eds.) English Corpus Linguistics,* 111–127. London, UK: Longman.

Kobayashi Y., Tokunaga T. & Tanaka H. (1996). *Analiza sestavljenih samostalnikov z uporabo pomenskih kolokacijskih podatkov med samostalniki. [Meishikan no imiteki kyouki jouhou wo mochiita fukugou meishi no kaiseki.] Shizen gengo shori,* 3(1), 29–43.

Kokuritsu kokugo kenkyuujo. (1962–64). 90 vrst terminov v sodobnih revijah. [Gendai zasshi 90 shu yougo youjiji daiichi ~san bunsatsu]. *Houkoku* 21, 22, 25. Shuuei shuppan.

Kokuritsu kokugo kenkyuujo. (1970). Raziskovanje časopisnega besedišča z računalnikom. [Denshi keisanki niyoru shimbun no goichousa.]. *Houkoku*, 37, 36. Shuuei shuppan.

Kokusai kouryuu kikin Nihon kokusai kyouiku kyoukai. (2002). *Naloge za test JLPT: Popravljena izdaja. [Nihon gonouryokushiten shutsudai kijun kaiteiban.].*

Komori, S. (2004). *A study of L2 lexical collocations of English-speaking learners of Japanese. [Nihongo gakushuu jishohensan ni muketa denshi-ka koopasu riyou ni yoru korokeeshon kenkyuu].* 117–129.

Krek, S. & Kilgarriff, A. (2006). Slovene Word Sketches. *Proceedings of the Fifth Slovenian / First International Languages Technology Conference.* Ljubljana, Slovenia.

Kucera, H. & Francis, W. N. (1967). *Computational Analysis of Present-Day American English.* Providence, Rhode Island: Brown University Press.

Leech, G. (1997). *Teaching and Language Corpora: A Convergence*. Teaching and Language Corpora, 1–23. London, UK: Longman.

Leech, G., Garside, R. & Bryand, M. (1994). CLAWS4: The tagging of the British National Corpus. *Proceedings of the 15th International Conference on Computational Linguistics (COLING »94),* 622–628.

Leech, G., Rayson, P. & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. Harlow, UK: Pearson Education Limited.

Lewis, M. (Ed.) (2000). *Teaching Collocation: Further Developments in the Lexical Approach*. Language Teaching Publications.

Lin, D. (1998). Automatic retrieval; and clustering of similar words. *Proceedings of COLING ACL,* 768–774.

Lüdeling, A. (Ed.) & Kytö, M. (Ed.) (2008). *Volume 1*. Berlin, Boston: De Gruyter Mouton. Retrieved 11 Apr. 2015 from http://www.degruyter.com/view/product/19320.

Maekawa, K. (2006a). Kotonoha. The Corpus Development Project of the National Institute for Japanese Language. *Proceedings of the 13th NIJL International Symposium: Language Corpora: Their Compilation and Application*, 55–62.

Maekawa, K. (2006b). Raziskava prioritetnega področja projekta Korpus japonščine. [Tokutei ryouiki kenkyuu „Nihongo koopasu" no mezasu mono.]. *Nihongo koopasu zentai kaigi soukatsu han houkoku.*

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M. & Den, Y. (2013). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation.* Netherlands: Springer.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

Masuoka, T. (1991). *Slovnica modalnosti. [Modariti no bunpou.]*. Kuroshio shuppan.

Matsumoto, Y., Asahara, M., Iwadate, M. & Morita T. (2009). Stanje in problemi orodja za dodajanje oznak *chaki*. [Tagutsuki koopasu kanri tsuuru „chaki" no genjou to kadai.]. *Tokutei ryouiki kenkyuu „Nihongo koopasu" heisei 20 nendo koukai waakushoppu*, 77–80.

Matsumoto, Y. (2003). Vrste korpusov za sodobni jezik in njihove značilnosti. [Gendaigo no koopasu no shurui to sorezore no tokuchou.]. *Nihongogaku,* 22, 54–60.

Matsuyoshi, T. & Sato M. (2007). Parafraziranje japonskih funkcionalnih izrazov, ki temeljijo Slovarju japonskih funkcionalnih izrazov. [Taikeiteki kinou hyougen jisho ni motoduku nihongo kinou hyougen no iikae.]. *Gengo shori gakkai dai 13 kai nenji taikai happyou ronbunshuu,* 899–902.

McEnery, T. & Hardie, A. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge, UK: Cambridge University Press.

Mel'čuk, I. (1998). Collocations and Lexical Functions. *A.P. Cowie (ed.): Phraseology. Theory, Analysis, and Applications*, 23–53. Oxford, UK: Clarendon Press.

Milkov, N. (2001). Logico-Linguistic Moleculism: Towards an Ontology of Collocations and other [Language] Patterns. *Proceedings of OntoLex«2000: Ontologies and Lexical Knowledge Bases*, 82–94. Sofia, Bulgaria: OntoText Lab.

Minami, F. (1974). *Struktura sodobne japonščine. [Gendai nihongo no kouzou.]*. Taishuukanshoten.

Minami, F. (1993). *Obris slovnice sodobne japonščine. [Gendai nihongo bunpou no rinkaku.].* Taishuukanshoten.

Moriyama T., Nitta, Y. & Kudō H. (2000). *Modalnost. [Modariti.]* Iwanamishoten.

Muraki, S. (2007). Korokeeshon to wa nani ka [What is a collocation?]. *Nihongo gaku (Japanese language linguistics)*, 26(12). Meiji shoin, 4–17.

Nation, P. (2001). Learning vocabulary in another language. Cambridge University Press.

New Century Japanese-English Dictionary [Nyuu senchurii waei jiten] (1999), 2. izd. Tokyo: Sanseido.

Nishina, K. & Yoshihashi, K. (2007). Japanese Composition Support System Displaying Occurrences and Example Sentences. *Symposium on Large-scale Knowledge Resources (LKR2007)*, 119–122.

Nishina, K., Yoshihashi K. & Liang F. (2007). *Razvoj sistema za podporo učenju japonščine in slovar EDR. [Nihongo gakushuu shien shisutemu kaihatsu to EDR jisho.]*. ICT-EDR shinpojiumu (demo).

Nishina, K. (2008). Struktura in vrednotenje sistema za podporo učencem japonščine pri pisanju spisov. [Nihongo gakushuusha sakubun shien no tame no kyouki hyougen to sakubun hyouji shisutemu kouchiku to hyouka.]. *Tokutei ryouiki kenkyuu „Nihongo koopasu" Heisei 19 nendo koukai waakushoppu*, 137–143.

Nitta, Y. (2002). *Vidik prislovnih izrazov. [Fukushiteki hyougen no shosou.].* Kuroshio shuppan.

Noda, H. (2007). Bunpoutekina korokeeshon to imitekina korokeeshon. *Nihongo gaku (Japanese language linguistics)*, 26(12). Meiji shoin. 18–27.

Ogino, T., Kobayashi, M. & Isahara H. (2003). *Vezljivost japonskega glagola. [Nihongo doushi no ketsugouka.].* Sanseidou.

Ogino, T., Kobayashi, M. & Isahara, H. (2007). Od slovarja kolokacij do slovarja vezljivosti. [Kyoukijisho kara ketsugouka jisho he.]. *ICT-EDR shinpojiumu kenkyuu happyou.*

Ogino, T. (ed.) (2008). Študija o metodi izdelave slovarja z uporabo korpusa. [Koopasu wo riyoushita kokugoshiten henshuuhou no kenkyuu.]. *Bunkashou kagaku kenkyuuhi tokutei ryouiki kenkyuu „Nihongo koopasu" jisho henshuuhan.*

Ogino, T., Kondou Y., Yazawa M. & Maruyama, N. (2006). Študija o metodi izdelave slovarja z uporabo korpusa. [Koopasu wo riyoushita kokugoshiten henshuuhou no kenkyuu.].*„Nihongo koopasu" zentai kaigi jisho henshuuhan houkoku*.

Oso, M. & Takizawa, N. (2003). Raziskovanje poučevanja japonščine s korpusom – Poudarek na kolokacijah in njihovih napačnih rabah. [Koopasu ni yoru nihongo kyouiku no kenkyuu – korokeeshon oyobi soono goyou wo chuushin ni.]. *Nihongogaku, April, Rinji zoukango*, Meiji shoin, 234–244.

Oso, M. (2006). Korpus japonščine in poučevanje japonščine. [Nihongo koopasu to nihongo kyouiku.]. *Nihongo kyouiku,* 130, 3–10.

Palmer, H. E. (1930). *Interim Report on Vocabulary Selection*. Tokyo, Japan: The Institute for Research in English Teaching.

Palmer, H. E. (1938). *A Grammar of English Words*. Harlow, UK: Longman.

Partington, A. (1998). *Patterns and Meanings: Using corpora for English language research and teaching*. Amsterdam, The Netherlands: John Benjamins.

Pardeshi, P. & Akasegawa, S. (2010). BCCWJ wo katsuyou shita kihon doushi handobukku sakusei: koopasu braujingu shisutemu NINJAL-LWP no tokuchou to kinou. *Tokutei ryouiki kenkyuu Nihongo koopasu Gendai Nihongo kakikotoba kinkou koopasu kansei kinen yokoushuu*. 205–216.

Rayson, P. & Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the ACL Workshop on Comparing Corpora*, 1–6. Hong Kong.

Rychlý, P. (2008). A Lexicographer-Friendly Association Score. *RASLAN 2008. 2. izd.*, 6–9. Brno, Czech Republic: Masarykova Univerzita.

Rundell, M, ed. (2002). *Macmillan English Dictionary for Advanced Learners*. London, UK: Macmillan.

Saitou, H. (1905). *Class-Books of English Idiomology*.

Sakakura, A. (1988). *Japonska pogovorna slovnica, druga izdaja. [Kaikou nihon bunpou no hanashi dainihan.]*. Kyouiku shuppan.

Sangawa, K. H., & Erjavec, T. (2012). JaSlo: Integration of a Japanese-Slovene Bilingual Dictionary with a Corpus Search System. *Acta Linguistica Asiatica*, 2(3), 125–140.

Sano, K. & Lee, J. (2007). Za kaj lahko uporabimo KH Coder – Implikacija na področje poučevanja in raziskovanja poučevanja japonščine. [KH Koder de naniga dekiruka ~nihongo shuutoku, nihongo kyouiku kenkyuu riyou he no shisa.]. *Gengobunka to Nihongo kyouiku*, 33, 47–48.

Sato, Y., Sakaue, T., Koizumi, T., Akkus, D. & Sugiura, M. (2007). Developing Ajax-Based Material for Learning Collocations [Ajax ni yoru jisaku kyouzai kaihatsu gihou.]. *The 4th International Conference on Computer Assisted Systems for Teaching & Learning Japanese (CASTEL-J)*. University of Hawaii, Kapiolany Community College.

Sharoff, S. (2006a). Creating general-purpose corpora using automated search engine queries. *WaCky! Working papers on the Web as Corpus*. Bologna, Italy: GEDIT.

Sharoff, S. (2006b). Open-source corpora: using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4), 435–462.

Siepmann, D. (2005). Collocation, colligation and encoding dictionaries. Part 1: Lexicological aspects. *International Journal of Lexicography,* 18(4), 409–443.

Siepmann, D. (2006). Collocation, colligation and encoding dictionaries. Part II: Lexicographical Aspects. *International Journal of Lexicography,* 19(1), 1–39.

Sinclair, J. M. (1987). *The nature of the evidence*. In J. McH. Sinclair (ed.) Looking Up: An Account of the COBUILD Project in Lexical Computing. London: Collins. 150–159.

Sinclair, J.M. & Renouf, A. (1988). A lexical syllabus for language learning. *R. Carter and M. Mccarthy, eds. Vocabulary and language teaching*. Harlow, UK: Longman.

Sinclair, J.M. (1991). *Corpus, concordance, collocation. Oxford, UK: Oxford University Press.*

Smith, S., Chen, A. & Kilgarriff, A. (2007). *A corpus query tool for SLA: learning Mandarin with the help of Sketch Engine.* Practical Applications in Language and Computers - PALC 2007.

Smith, S., Sommers, S. & Kilgarriff, A. (2008). *Learning words right with the Sketch Engine and WebBootCat: Automatic cloze generation from corpora and the web*. Proc. CCU. Taipei, Taiwan.

Smrž, P. (2004). *Integrating Natural Language Processing into E-learning — A Case of Czech, Proceedings of the Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning.* COLING 2004. 106–111.

Sparck, K. J. (1986). *Synonymy and Semantic Classification*. Edinburgh, UK: Edinburgh University Press.

Srdanović, I., Erjavec T. & Kilgarriff, A. (2008). A web corpus and word-sketches for Japanese. *Shizen gengo shori (Journal of Natural Language Processing),* 15(2), 137–159.

Srdanović, I. & Nishina, K. (2008). The Sketch Engine corpus query tool for Japanese and its possible applications [Koopasu kensaku tsuuru Sketch Engineno nihongo ban to sono riyou houhou]. *Nihongo kagaku (Japanese Linguistics),* 23, 59–80.

Srdanović, I., Bekeš, A. & Nishina, K. (2009a). Koopasu ni motozuita goi shirabasu sakusei ni mukete: suiryouteki fukushi to bunmatsu modariti no kyouki wo chuushin ni shite (Towards creation of lexical syllabus based on corpora - on suppositional adverbs and clause-final modality collocations. *Nihongo kyouiku (Journal of Japanese Language Education)*, 142, 69–79.

Srdanović, I., Hodošček B., Bekeš, A. & Nishina, K. (2009b). Extracting distant collocations of adverbs and modality forms using a web corpus and a query system. [Uebukoopasu to kensaku shisutemu o riyou shita suiryou fukushi to modariti keishiki no enkaku kyouki chuushutsu to nihongo kyouiku e no ouyou]. *Shizen gengo shori*, 16(4), 29–46.

Srdanović, I., Ida, N., Shigemori Bučar, C., Kilgarriff, A. & Kovar, V. (2011). Japanese word sketches: advantages and problems. *Acta Linguistica Asiatica*, 1(2), 63–82.

Srdanović, I. & Sakoda, K. (2013). Analysis of learner's production of adjectives using the Japanese language learner's corpus C-JAS: the case of takai. *Acta Linguistica Asiatica,* 3(2), 9–24.

Srdanović, I. (2013). Description of Adjective and Noun Collocations Based on Large-Scale Corpora: Towards Dictionary for Japanese Language Learners (in Japanese). *Kokuritsu kokugo kenkyuujo ronshuu (NINJAL Research Papers),* 6.

Srdanović, I., Suchomel, V., Ogiso, T. & Kilgarriff, A. (2013). Japanese Language Lexical and Grammatical Profiling Using the Web Corpus JpTenTen (in Japanese). *Proceeding of the 3rd Japanese corpus linguistics workshop,* 229–238. Tokyo, Japan: NINJAL, Department of Corpus Studies/ Center for Corpus Development.

Srdanović, I. (2014). Corpus-Based Collocation Research Targeted at Japanese Language Learners. *Acta Linguistica Asiatica*, 4(2), 25–36. (http://revije.ff.uni-lj.si/ala/) (DOI: 10.4312/ala.4.2.25-36).

Sunakawa, Y. (2007). Funkcijske besede kolokacij – O prislovu in kolokacijah. [Kinougo no korokeeshon (1) Fukushi to no kyouki ni tsuite.]. *Nihongo kyouiku han dainikai koukai kaigi.*

Sunakawa, Y. (2008). *Raziskava poučevanja japonščine z uporabo korpusa pisanega jezika. [Daihyousei wo yuusuru kakikotoba koopasu wo katsuyoushita nihongo kyouiku kenkyuu.].* Posebno področje raziskovanja »Korpusa japonščine«, odprta delavnica leta 2008.

Takizawa, N. (2004). Luščenje kolokacij iz japonskih elektronskih besedil. [Nihongo denshi tekisuto kara no korokeeshon chuushutsu]. *Nihongo gakushuu jisho ni muketa denshika koopasu riyou niyoru korokeeshon kenkyuu*, 27–40.

Takizawa, N. (2006). Računalniška pismenost za uporabo korpusov. [Koopasu riyou no tame no konpyuuta riterashii.]. *Nihongo kyouiku*, 130, 22–31.

Takizawa, N. (2007). *Raziskovanje japonščine in japonsko korpusno poučevanje. [Nihongo kenkyuu to nihongo kyouiku to koopasu.]*

Terashima, K. & Moriguchi, M. (2005). The first collocation dictionary for the Japanese language. *Words in Asian Cultural Contexts: Proceedings of the 4th Asialex Conference 2005*, 303–307.

Terashima, K. & Takizawa, N. (2004). Towards compiling a corpus-based dictionary of Japanese collocations. *Proceedings of the 3th Asialex Biennial International Conference 2003.*

Terashima, K. (2004). Sopomenke z vidika podobnosti in razlik v obnašanju. [Furumai no idou kara mita ruigigo – ureshii/tanoshii/omoshiroi wo rei ni shite]. *Nihongo gakushuu jisho ni muketa denshika koopasu riyou niyoru korokeeshon kenkyuu*, 253–258.

Tono, Y. (2007). Poskus implementacije sistema Sketch engine v korpus japonščine. [Nihongo koopasu de no Sketch Engine jissou no kokoromi.]. *Tokutei ryouiki kenkyuu „Nihongo koopasu" Heisei 18 nendo koukai waakushoppu, Monbukagakushou kagakukenkyuuhi tokutei ryouiki kenkyuu „Nihongo koopasu"*, 109–112.

Ueyama, M. & Baroni, M. (2005). Automated construction and evaluation of a Japanese web-based reference corpus. *Proceedings of Corpus Linguistics 2005*.

Weeds, J. & Weir, D. (2005). Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity. *Computational Linguistics*, 31(4), 439–475.

West, M. (1953). A general service list of English words with semantic frequencies and a supplementary word-list for the writing of popular science and technology. London, UK: Longman, Green and Company.

Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied Linguistics,* 21(1), 3–25.

Willis, D. (1990). *The lexical syllabus*. London, UK: Harper Collins.

Yamada, S. (2008). Kolokacijski opisi in pomensko razvrščanje samostalnikov. [Korokeeshon no kijutsu to meishi no imi bunrui.]. *Nihongogaku*, 26, 48–57.

Yamauchi, H.. (2004). Metode za raziskovanje usvajanja besedišča – Statistika N-gram in Chasen. [Goi shuutoku kenkyuu no houhou – Chasen to N-guramu toukei.]. *Dainigengo to shite no nihongo no shuutoku kenkyuu,* 7, 141–162.

Yamauchi, H. (2006). O pripravljanju učnega načrta za besedišče pri poučevanju japonščine. [Nihongo kyouiku ni okeru gou shirabasu no sakusei ni tsuite.]. *Tokutei ryouiki kenkyuu „Nihongo koopasu" heisei 18 nendo koukai waakushoppu*, 161–164.

Yamauchi, H. (ur.) (2008). *Standardni nabor besed za poučevanje japonščine. [Nihongo kyouiku sutandaado shian: goi.]*. Hitsuji shobou.

Yamauchi, H. (2015). Attempting to Maintain Objectivity in Syllabus Creation. *NINJAL Project Review,* 4(3), 197–204.

Yamazaki, M. (2006). Oblikovanje sodobnega japonskega pisanega korpusa z reprezentanti. [Daihyousei wo yuusuru gendai nihongo kakikotoba koopasu no sekkei.]. *Kokuritsu kokugo kenkyuusho*, 63–70.

Stvarno kazalo

L

P

R

S

Š

T

U

V

W

Y

Z