Gradnja in analiza korpusov za prevodoslovne raziskave

Špela Vintar in Darja Fišer

This paper presents guidelines for the construction and analysis of corpora in translation studies. The first part introduces some basic concepts in corpus linguistics and discusses the goals and types of corpus-based translation studies, gives theoretical and practical guidelines regarding the representativeness of specialized corpora and outlines the corpus annotation process at several levels and for different languages. The second part presents methods of corpus analysis with an emphasis on the tools that support the Slovene language and are appropriate for analysing both publicly available and custom-built corpora. The most important functions, such as the use of concordancers, frequency lists, collocations and keywords, are described and illustrated with practical examples.

Ključne besede: gradnja korpusov, korpusno prevodoslovje, reprezentativnost, korpusna orodja, označevanje

1 UVOD

Številne prevodoslovne raziskave temeljijo na besedilnem gradivu. Gradivo je lahko najrazličnejših oblik in vsebin; lahko vsebuje prevedena besedila in njihove izvirnike, samo prevedena besedila, izvirnik in več prevodov v isti jezik ali v več jezikov; vsebuje lahko pisna in govorjena besedila; besedila v celoti ali zgolj izpisane fragmente. Kadar nabor gradiva izpolnjuje določene zahteve, lahko govorimo o korpusu, pri čemer zgolj obstoj korpusa ne pomeni nujno, da je tudi na njem izvedena raziskava korpusna. Pričujoče poglavje se ukvarja s korpusi v prevodoslovju, in sicer tako z njihovo izgradnjo kot z metodološkimi pristopi h korpusni analizi.

> *Corpus: A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.* (EAGLES 1996)

Po definiciji EAGLES[12] je korpus zbirka besedil ali njihovih delov, izbranih in urejenih v skladu z določenimi jezikoslovnimi kriteriji, ki naj bi se uporabljala kot vzorec jezika. K tej definiciji moramo dodati še, da so korpusi danes izključno na računalniških medijih in da tiskanih zbirk gradiva ne imenujemo več korpus. Pomemben element pri opredelitvi korpusa, ki ga nekatere druge definicije tudi eksplicitno omenjajo (Atkins et al. 1992, Gorjanc 2005), je namembnost za pridobivanje jezikoslovnih spoznanj, kar nam daje vedeti, da se korpusi razlikujejo glede na to, kaj želimo z njimi početi, nadaljnji pomemben poudarek pa je na besedi vzorec, saj nam ta sporoča, da je od korpusa upravičeno mogoče pričakovati, da daje karseda verno sliko jezikovne zvrsti, ki jo predstavlja, z drugimi besedami, da je reprezentativen.

2 GRADNJA KORPUSOV ZA PREVODOSLOVNE RAZISKAVE

2.1 Vrste korpusnih raziskav

S pomočjo korpusov je mogoče izvajati tako jezikoslovne kot kulturološke študije prevodov. Z združevanjem kvantitativne in kvalitativne korpusne analize lahko prevode raziskujemo na leksikalni, skladenjski, pa tudi diskurzivni ravni. Zanimajo nas lahko pogosti/tipični oz. redki/nenavadni pojavi v prevajanem jeziku na splošno, kar sodi v skupino raziskav o t. i. prevodoslovnih univerzalijah (Baker 1993). V drugi sklop pa sodi proučevanje sloga posameznih prevodov glede na določene kriterije, kjer se posvečamo predvsem prevajalčevemu/avtorjevemu

[12] http://www.ilc.cnr.it/EAGLES/typology/node4.html

slogu, razlikam med žanri, primerjavi prevodov v različnih časovnih obdobjih oz. družbenih okoliščinah in podobno (Tymoczko 2000).

Glede na status prevodov v raziskavah in glede na vrste korpusov, ki jih pri tem uporabljamo, je korpusne študije v prevodoslovju mogoče razdeliti na tri skupine (Olohan 2004):

a) V kontrastivnih študijah je prevod obravnavan v odvisnosti od izvirnika. Za tovrstne raziskave uporabljamo vzporedne korpuse, s pomočjo katerih proučujemo, kako prevajalci rešujejo konkretne (jezikovne ali kulturološke) prevajalske probleme (Kenny 2001).
b) V študijah, ki prevode obravnavajo kot samostojna besedila brez neposrednega vpogleda v izvirnike oziroma brez upoštevanja odnosa med izvirnikom in prevodom, uporabljamo enojezične oz. primerljive korpuse prevodov. V njih iščemo pojave, značilne za prevodni jezik, in proučujemo vplive spremenljivk, kot so jezik izvirnika, žanr, način prevajanja ipd. (Laviosa 2002).
c) Za proučevanje značilnosti prevedenih besedil v razmerju do neprevedenih besedil v ciljnem jeziku pa poleg korpusa prevodov uporabljamo še referenčni korpus v ciljnem jeziku. Na podlagi izsledkov je mogoče pridobiti dragocene informacije o slovnici, semantičnih odnosih, sprejemljivosti določene rabe, neologizmih oz. zastarelih izrazih ter pragmatičnih vidikih ciljnega jezika (McEnery in Wilson 2001).

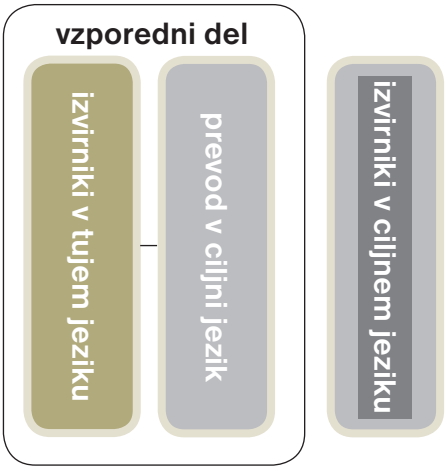
Kadar načrtujemo jezikoslovno ali prevodoslovno raziskavo in bi zanjo radi uporabili korpus, je prvi korak navadno poizvedba, ali že obstaja primeren korpus, ki predstavlja reprezentativen vzorec za tisto jezikovno zvrst, ki jo nameravamo raziskovati. Za raziskovanje prevodov in s prevajanjem povezanih pojavov se največkrat uporabljajo vzporedni korpusi, se pravi korpusi, ki vsebujejo izvirnike in njihove prevode v enem ali več jezikih in ki so stavčno poravnani, kar pomeni, da so izvirni in prevedeni segmenti med seboj povezani. Za prevodoslovje pa so zanimive še druge vrste korpusov, denimo:

- enojezični, kadar raziskujemo zgolj prevode v določeni jezik in nas izvirniki ne zanimajo,
- večjezični primerljivi, kadar raziskujemo prevode sorodnih besedil v več jezikov,
- prevodoslovni, ki vsebuje prevode v določeni jezik ter primerljiva izvirna besedila v tem jeziku.

Baker (1995) sicer predlaga prevodoslovno korpusno tipologijo, ki razlikuje med vzporednimi, večjezičnimi in primerljivimi korpusi. V vzporednih korpusih so izvirniki stavčno poravnani s prevodi v ciljni jezik, kar prav tako velja za večjezične korpuse s to razliko, da je v večjezičnih korpusih vključenih več jezikovnih parov.

Primerljivi korpusi pa namesto izvirnikov in prevodov vsebujejo besedila v dveh ali več jezikih, ki niso izvirniki oz. prevodi, temveč so si med seboj podobna glede na žanr, področje ipd. Vendar Baker kasneje uvede še koncept prevodoslovnega korpusa, ki je pravzaprav enojezični primerljivi korpus. Namen prevodoslovnega korpusa je, da omogoča raziskave značilnosti prevedenega jezika v primerjavi z neprevedenim jezikom, te značilnosti so znane tudi kot prevodne univerzalije (prim. Toury 1995, Baker 1996). Če k prevodoslovnemu korpusu dodamo še izvirnike prevedenih besedil, dobimo vzporedno-primerljivi korpus z značilno tripartitno strukturo (Slika 1). Prvi prevodoslovni korpus, na katerem je bilo opravljenih že precej raziskav, je Translational English Corpus (TEC)[13] z Univerze v Manchestru. Sorodni projekti po svetu so še CroCo,[14] finski prevodoslovni korpus, zgrajen na Univerzi v Joensuu (Eskola 2004) in drugi. Za slovenščino nastaja obsežen prevodoslovni korpus v sklopu projekta Slovensko prevodoslovje – viri in raziskave,[15] ki poteka od maja 2009. Nastal naj bi petjezični vzporedno-primerljivi korpus prevodov v slovenščino, njihovih izvirnikov ter primerljivih neprevedenih besedil v slovenščini.

S slovenščino kot enim od jezikov je bilo zgrajenih že kar nekaj vzporednih korpusov, v glavnem za prevajalske, pa tudi za jezikovnotehnološke namene, npr. IJS-Elan,[16] Evrokorpus,[17] Trans, JRC-ACQUIS,[18] vendar so ti korpusi za prevodoslovne namene žal le pogojno uporabni, saj večinoma ne vsebujejo informacije o smeri prevoda (prim. Vintar 2008).



Slika 1: Struktura prevodoslovnega korpusa

[13] http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm
[14] http://fr46.uni-saarland.de/croco/index_en.html
[15] http://lojze.lugos.si/spook/
[16] http://nl2.ijs.si/index-bi.html
[17] http://evrokorpus.gov.si
[18] http://wt.jrc.it/lt/Acquis/

Za veliko večino raziskovalnih vprašanj, ki se porajajo prevodoslovcem v slovenskem prostoru, je torej potrebno primeren korpus šele zgraditi. To samo po sebi ni nič strašnega, je pa ob tem dobro opraviti nekaj temeljnih razmislekov, saj je – kot pravi Biber v spodnjem citatu – uporabnost korpusa neločljivo povezana z uporabnikovimi raziskovalnimi nameni na eni strani in reprezentativnostjo korpusa na drugi strani.

A corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what it is meant to represent. The representativeness of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalizability of the results of the research.
(Biber et al., 1998: 246)

Pri zbiranju besedil za korpus pravzaprav zbiramo primerke iz populacije z namenom ustvariti tak vzorec, ki bo kar najbolje predstavljal lastnosti preiskovane populacije. Statistika in teorija vzorčenja poznata več načinov izbiranja vzorca, denimo slučajnostnega, sistematičnega, namenskega itd. (Kožuh 2008: 130–135). Neslučajnostno vzorčenje se uporablja predvsem takrat, kadar ima populacija lastnosti, ki niso normalno porazdeljene in jih želimo sistematično vključiti v vzorec. Pri raziskovanju jezika se moramo vseskozi zavedati, da besede v jeziku niso razporejene naključno, zato ne moremo privzeti normalne porazdelitve, obenem pa se vseskozi srečujemo s paradoksalno situacijo, da želimo na podlagi vzorca sklepati o lastnostih določene jezikovne zvrsti, vendar te jezikovne zvrsti ne moremo sistematično vzorčiti, dokler ne vemo, katere so njene bistvene lastnosti, ki jih želimo zajeti.

Pri prevodoslovnih raziskavah je odločilnega pomena hipoteza, iz katere izhaja korpusna metoda, sledita pa vprašanji populacije in primernega vzorca, ki predstavljata ključni dejavnik pri gradnji korpusa. S korpusno metodo mislimo na kvantitativni način obdelave korpusnega gradiva, kjer s pomočjo delno ali v celoti avtomatiziranih postopkov poizvedovanja po korpusu in statistične obdelave korpusnih podatkov pridobivamo dokaze in spoznanja. Podrobneje bodo ti postopki predstavljeni v drugem delu poglavja, na splošno pa se korpusna metoda uporablja za dva tipa raziskav (Tognini-Bonelli 2001: 65):

1. Korpusno podprte (corpus-based) raziskave uporabljajo korpus za preskušanje, dokazovanje ali dopolnjevanje teorij in jezikovnih opisov, ki so nastali neodvisno od korpusnih podatkov.
2. Nasprotno pa korpusno usmerjene (corpus-driven) raziskave v korpusu vidijo mnogo več kot le zbirko primerov, s katerimi bi bilo mogoče dokazati vnaprej oblikovano hipotezo. Teoretična izhodišča tu v celoti izhajajo iz korpusa in odsevajo jezikovno rabo, kot jo kaže korpus.

V tem smislu naj bi veljal napotek, da korpusno metodo izbirajmo le takrat, kadar smo korpusnim dokazom tudi res pripravljeni zaupati. Kljub temu da se zavedamo, da nam popolne objektivnosti tudi tak pristop ne nudi, je obenem metodološko nesprejemljivo selektivno navajanje korpusnih primerov in načrtno prikrojevanje rezultatov z namenom potrditve vnaprej oblikovane hipoteze.

Če smo še vedno prepričani, da naša prevodoslovna hipoteza zahteva korpusno metodo in primeren korpus še ne obstaja, se prične gradnja korpusa. Temeljna vprašanja pri gradnji korpusa, s katerimi se utegnemo srečati, so tale:

1. tip korpusa,
2. velikost/reprezentativnost,
3. avtorske pravice,
4. predobdelava in označevanje besedil,
5. poizvedovanje po korpusu in obdelava korpusnih podatkov.

V nadaljevanju razpravljamo o vsaki od navedenih tem.

2.2 Tip korpusa

Korpuse delimo na eno- in večjezične, pri čemer so slednji lahko vzporedni ali primerljivi; diahrone in sihnrone glede na to, ali želijo predstavljati jezikovno rabo v določenem trenutku ali skozi daljše obdobje; govorne in pisne glede na prenosnik; zaključene in spremljevalne glede na dotok novih besedil (prim. Erjavec 1997, Gorjanc 2005: 8–11). Kot smo razložili že v uvodu, imamo pri prevodoslovnih raziskavah skoraj neizogibno – vsaj posredno – opravka z najmanj dvema jezikoma, čeprav ni nujno, da je vsaka prevodoslovna raziskava večjezična ali da zahteva večjezični korpus.

Če raziskujemo zgolj prevode, bi morda zadoščal enojezični korpus prevodov, vendar je prevodoslovne pojave izjemno težko interpretirati brez vpogleda v izvirno besedilo. Če namreč v korpusu prevodov odkrijemo določeno jezikovno posebnost, se brez omenjenega vpogleda postavlja vprašanje, ali gre za sistemsko značilnost ciljnega jezika ali za s prevajanjem povezani pojav, saj je bi lahko šlo tudi za interferenco med izvirnim in ciljnim jezikom.

Iz zgornjega sledi, da moramo pri vsaki prevodoslovni hipotezi, ki jo želimo dokazati s pomočjo korpusa, poskrbeti za "kontrolni korpus". Če denimo raziskujemo jezikovne lastnosti prevodov v nematerni jezik, jih moramo primerjati s prevodi v materni jezik in po možnosti še z izvirnimi besedili, da bodo pridobljene trditve predstavljene v ustreznem kontekstu.

Kadar se odločamo za vzporedni korpus, je dobro že vnaprej predvideti tehnične rešitve za poravnavo besedil ter načine iskanja po končanem korpusu. Za vzporedne korpuse je namreč manj kakovostnih programskih rešitev kot za enojezične, pogosto pa je treba posegati po programskih orodjih lastne izdelave.

2.3 Velikost in reprezentativnost korpusa

So when you design a corpus it is probably best to write down what you would ideally like to have, in terms of the amount and the type of language, and then see what you can get; adjust your parameters as you go along, keeping a careful record of what is in the corpus, so that you can add and amend later, and if others use the corpus they know what is in it. It is important to avoid perfectionism in corpus building. (Sinclair 2005)

Zagotoviti, da bo korpus dobro predstavljal izbrano jezikovno zvrst, je verjetno eno najkompleksnejših vprašanj korpusnega jezikoslovja, saj je reprezentativnost težko meriti. Kadar gradimo velike referenčne korpuse jezikov, zajemamo besedila v čim širšem žanrskem, geografskem in medijskem razponu, velikosti takšnih korpusov pa danes dosegajo že prek pol milijarde pojavnic (npr. Cosmas II,[19] Collins WordBanks,[20] Fidaplus[21]). A še vedno se lahko zgodi, da iskane besede ni v korpusu – mar to pomeni, da je ni v jeziku? Ker so pojavi, ki jih raziskujemo v prevodoslovju, pogosto kompleksni in zaobjemajo več jezikoslovnih ravni, je toliko teže zbrati res reprezentativen vzorec za njihovo opazovanje, ker jih je tudi toliko teže prešteti. Kako lahko na primer samodejno preštejemo izpuste v prevodu? Z današnjimi korpusnimi metodami zgolj tako, da jih najprej ročno označimo.

Zares reprezentativen je lahko le t. i. kumulativni korpus, se pravi korpus, ki zbere vsa besedila raziskovane jezikovne zvrsti. Tak korpus je bil denimo ustvarjen v okviru raziskovalnega projekta Slovenski prevodi nemških besedil 1848–1919 pod vodstvom dr. Ericha Prunča in je prek bibliografije TraDok[22] dostopen prevodoslovni javnosti. V tem projektu so bila zabeležena vsa objavljena besedila ciljnega obdobja, ki so bila prevedena iz nemščine v slovenščino, večina besedil pa je bila nato digitalizirana in oblikoskladenjsko označena. Pri takšnem korpusu ne vzorčimo, zato predstavlja takšna zbirka najboljši približek reprezentativnemu korpusu.

Ker je kumulativni korpus pogosto nemogoče zgraditi, moramo biti pri izboru vzorca nadvse pazljivi, da bo korpusne dokaze res mogoče posploševati na celot-

[19] http://www.ids-mannheim.de/cosmas2/
[20] http://www.collinslanguage.com/wordbanks/Default.aspx
[21] http://www.fidaplus.net
[22] http://itat2.uni-graz.at/pub/tradok/

no raziskovano jezikovno zvrst. Prvi korak pri tem je, da ciljno populacijo čim bolje zamejimo. Če denimo raziskujemo prevode sodobnih književnih avtorjev in nameravamo zajemati prevode del, ki so nastala v zadnjih desetih letih, smo populacijo sicer navidez zamejili, vendar se bodo v tem izboru znašli tako prvenci mladih avtorjev kot dela zrelih književnikov, ki se morda slogovno in jezikovno med seboj bolj razlikujejo, kot če bi izbrali daljši časovni razpon in obdobje raje zamejili z avtorjevo letnico rojstva.

Nato moramo določiti osnovno korpusno enoto, ki bo predstavljala posamezne vzorčene jezikovne zvrsti. Korpusna enota je lahko članek, novela, pesem, roman, poglavje, spletna stran ali kaj drugega. Biber in dr. (1998) pri napotkih za sestavo korpusa poudarja raznolikost, ki je za zagotavljanje reprezentativnosti pomembnejša od velikosti korpusa. Korpus, ki bo sestavljen iz krajših izsekov iz čimveč različnih besedil, dolgih denimo 1.000 besed, bo bolj reprezentativen v smislu predstavljanja jezikovne raznolikosti kot korpus, sestavljen iz manjšega števila daljših besedil. Pomemben dejavnik pri izbiri vzorčnih korpusnih enot pa je seveda tudi predmet preučevanja. Nekatere slovnične strukture dosegajo reprezentativne pogostosti že pri kratkih, denimo 1.000-besednih izsekih, po drugi strani pa za raziskovanje kompleksnejših pojavov potrebujemo neprimerno več gradiva. Danes diskovne kapacitete niso več težava in ni posebnih razlogov, da bi besedila pri vključevanju v korpus krajšali.

Obstaja več načinov, kako ovrednotiti reprezentativnost korpusa. Dickinson (2009) v navezavi na Biberja (1993) reprezentativnost korpusa definira kot "mero, do katere vzorec vsebuje variabilnost celotne populacije" in ki nam omogoča, da pridobljene rezultate posplošujemo na celotno vzorčeno jezikovno zvrst. Če želimo opazovati določeni jezikovni pojav X v korpusu velikosti N in nas pri tem zanima, ali je korpus dovolj velik za načrtovano opazovanje, je informativna statistika standardne napake:

$$s_{\bar{x}} = \frac{s}{N}$$

V formuli s pomeni standardna deviacija izbrane spremenljivke, N pa velikost korpusa. Intuitivno jasno je, da za čim manjšo standardno napako potrebujemo čim večji N.

Podobno razmišljanje nas lahko vodi tudi pri vrednotenju reprezentativnosti, in sicer denimo pri merjenju stabilnosti pogostosti izbranega jezikovnega pojava v različno velikih korpusih. Zelo pogosti jezikovni pojavi, denimo besedna vrsta samostalnik, so razmeroma stabilni že pri majhnih korpusnih vzorcih, redkejši

pojavi pa zahtevajo daljše vzorce. Tako bo denimo pogostost samostalnika v več 200-besednih vzorcih približno enaka, kar zagotovo ne bo veljalo za pogostost priredij z *niti-niti*.

Medtem ko za nekatere prevodoslovne raziskave potrebujemo korpus, ki bo čim bolje predstavljal jezikovno raznolikost opazovane zvrsti, pa za druge, denimo terminološke raziskave, potrebujemo čim bolj homogene specializirane korpuse. Kilgarriff (2001) predlaga metodo za merjenje homogenosti korpusa, ki v grobem korpus primerja s samim seboj, in sicer tako, da korpus razdeli na več enako velikih delov in nato izračunava podobnost vsakega dela z vsemi ostalimi, za testiranje različnih cenilk pa vzpostavi kontrolno populacijo, za katero je podobnost znana. Svojo metodo zato poimenuje metoda korpusov znanih podobnosti oziroma *Known-Similarity Corpora*.

Kako simuliramo znano podobnost? Postopek je precej preprost. Vzamemo dva korpusa precej različnih jezikovnih zvrsti, A in B. Nato zgradimo skupino kontrolnih korpusov KK tako, da KK1 vsebuje 100 % A, KK2 90 % A in 10 % B, KK3 80 % A in 20 % B, KK4 70 % A in 30 % B in tako naprej. Zdaj lahko trdimo, da je KK2 bolj podoben korpusu A kot KK3, KK4 je bolj podoben korpusu B kot KK1 in podobno. S temi kontrolnimi skupinami lahko testiramo različne metode merjenja podobnosti med korpusi.

Rezultati pokažejo, da ta statistika dobro predstavi razlike v pogostosti besed med dvema korpusoma, vrednost cenilke pa postopoma narašča z naraščajočo pogostostjo besede. To ustreza intuitivni predpostavki, da so bolj pogoste besede boljše merilo različnosti oziroma podobnosti korpusov kot manj pogoste. Za preskus homogenosti strokovnega korpusa bi opisano metodo uporabili tako, da bi test enkrat izvedli s korpusoma različnih področij, drugič pa z istim korpusom, naključno razdeljenim na dva dela (Vintar 2008).

2.4 Avtorske pravice

Pri gradnji korpusov se prej ali slej srečamo tudi z vprašanjem avtorskih pravic. Četudi zbiranje besedil in njihovo vključevanje v korpus navadno predpostavljata raziskovalno rabo in izključujeta reproduciranje in distribucijo teh besedil ter materialno okoriščanje z njimi, pri tem vseeno nemalokrat prihaja do težav in spornih praks, ki v eni skrajnosti pretirano ščitijo avtorje oziroma predvsem založbe ter s tem onemogočajo jezikoslovno raziskovanje, v drugi pa povsem razvrednotijo avtorske vsebine in jih razširjajo kot javno dobro, kar niso.

V Sloveniji vprašanje avtorskih pravic ureja *Zakon o avtorski in sorodnih pravicah* (Ur.l. RS, št. 16/2007). Po tem zakonu je avtorska pravica skupno poimenovanje za več različnih pravic, ki pripadajo avtorju avtorskega dela in ki se delijo v tri sklope:

– moralne avtorske pravice (npr. pravica priznanja avtorstva, pravica spoštovanja avtorskega dela);
– materialne avtorske pravice (npr. pravica reproduciranja, pravica javnega izvajanja, pravica javnega prenašanja, pravica dajanja na voljo javnosti);
– druge pravice avtorja (npr. pravica dostopa k izvirniku ali primerku dela, pravica do nadomestila za tonsko ali vizualno snemanje in za fotokopiranje).

Kot v drugih državah po svetu tudi slovenska zakonodaja določa, da avtorska pravica nastane s samo stvaritvijo avtorskega dela in zanjo ni potrebna nobena registracija ali drugo uradno dejanje. To pomeni, da je načeloma za vsako uporabo avtorskega dela potrebno pridobiti dovoljenje avtorja oziroma nosilca avtorskih pravic.

Čeprav o spletnih vsebinah zakon ne govori izrecno in je pogosto slišati mnenja, da so avtorska dela na spletu javna dobrina, ki jo lahko vsak prosto uporablja, je dejstvo, da ta avtorska dela uživajo povsem enako varstvo kot dela, objavljena na klasični način.[23]

Uporaba avtorskih del pomeni njihovo objavo, distribucijo, dajanje na voljo javnosti (objava na spletni strani), predvajanje in podobno. Kot v "fizičnem" tudi v digitalnem svetu velja pravilo, da moramo za uporabo avtorskega dela pridobiti dovoljenje avtorja ali drugega nosilca avtorskih pravic (založbe, upravitelja spletnega portala ipd.) Vsaka uporaba brez dovoljenja je prepovedana in ima lahko za posledico odškodninsko ali celo kazensko odgovornost tistega, ki krši avtorske pravice.

Korpus je v mnogih vidikih specifična oblika reproduciranja in distribuiranja. Večina spletnih korpusnih iskalnikov ne omogoča vpogleda v celotno delo, temveč uporabniku v obliki konkordance ponuja zgolj kratke izseke iz besedil, bibliografski podatki o vsakem delu posebej pa so navedeni na koncu vsake konkordančne vrstice. Načeloma bi lahko trdili, da gre pri taki vrsti dostopa do korpusa za posebno obliko citiranja, ki – še posebej pri prosto dostopnih korpusih brez komercialne izrabe – služi javnemu dobru. Žal večina avtorjev in založnikov ni tega mnenja, zato za spletno dostopne korpuse velja napotek, da je za vsako avtorskopravno zaščiteno besedilo potrebno pridobiti dovoljenje oziroma z besedilodajalcem skleniti pogodbo; takšna praksa je ustaljena tudi pri vseh resnejših korpusnih projektih v Sloveniji.

[23] Več o tem v zanimivi razpravi na portalu pravozatelebane.com, http://www.pravozatelebane.com/index.php?option=com_content&task=view&id=11&Itemid=107, objavljeno 27.8.2007.

Korpus, ki ga gradi posameznik ali institucija za lastne raziskovalne namene in ga ne namerava javno objaviti prek iskalnika, se lahko sklicuje na 50. člen ZASP, ki pravi:

(2) Fizična oseba lahko prosto reproducira delo:
1. na papirju ali podobnem nosilcu z uporabo fotokopiranja ali druge fotografske tehnike s podobnimi učinki,
2. na katerem koli drugem nosilcu, če to stori za privatno uporabo, če primerki niso izročeni ali priobčeni v javnosti in če pri tem nima namena dosegati neposredne ali posredne gospodarske koristi.

(3) Javni arhivi, javne knjižnice, muzeji ter izobraževalne in znanstvene ustanove lahko za lastne potrebe prosto reproducirajo delo na kateremkoli nosilcu, če to storijo iz lastnega primerka in če pri tem nimajo namena dosegati neposredne ali posredne gospodarske koristi.

2.5 Predobdelava in označevanje besedil

Ko smo se odločili za nabor besedil in razjasnili vprašanja v zvezi z avtorskimi pravicami, se začne fizično zbiranje. Pri besedilih, ki niso na razplago v elektronski obliki, je prva faza skeniranje in optično razpoznavanje znakov (OCR), nato pa po potrebi še ročno pregledovanje elektronske različice. Že v tej fazi se moramo tudi odločiti, katere podatke o fizičnem nosilcu bomo ohranjali; za nekatere raziskave je denimo pomembno ohraniti elektronski faksimile izvirnika, številko strani v izvirni izdaji, sklice na slike in druge nejezikovne prvine.

Ko so besedila na razpolago v elektronski obliki, se jih pretvori v enotno obliko. Ta je navadno golo besedilo (txt), določiti in zagotoviti moramo še enoten kodni nabor (navadno UTF-8). Če iz oblikovanih in strukturiranih dokumentov, kot so formati Word, QuarkExpress, HTML itd., izluščimo zgolj besedilo, se precej metajezikovnih informacij izgubi. Če so za našo raziskavo podatki o funkciji besedila (naslov, citat, celica v tabeli) in njegovi obliki (krepko, ležeče, pisava, velikost pisave) pomembni, jih moramo ohraniti v obliki korpusnih oznak. Iz spletnih dokumentov navadno odstranimo tudi navigacijske elemente (Domov, Nazaj, Naprej).

Tako poenoteni in prečiščeni dokumenti so primerni za nadaljnje jezikoslovne analize, denimo za poravnavo ali oblikoskladenjsko označevanje. Najprimernejše sosledje teh korakov je odvisno od specifik korpusnega projekta, ne glede na to pa se je treba še prej odločiti za format zapisa korpusa in za strukturo ter nabor metabesedilnih oznak. Ker so napotki za standardizacijo korpusnih zapisov in strukturiranje glave podrobno podani drugod (Erjavec 1997; Erjavec 2003; Gorjanc 2005: 56–63), se v nadaljevanju ukvarjamo predvsem z vprašanji, ki se porajajo pri gradnji večjezičnih korpusov.

2.5.1 Stavčna poravnava

Kadar imamo na razpolago izvirnike in njihove prevode v enem ali več jezikih, govorimo o vzporednem korpusu. Polna funkcionalnost vzporednega korpusa je omogočena šele, ko za vsak segment v izvirniku poznamo njegov pripadajoči segment v vsakem od prevodov. Vzporedni korpus, ki vsebuje več kot dva jezika, je lahko poravnan za vsak jezikovni par posebej (za štirijezični vzporedni korpus je to 6 jezikovnih parov), ali pa je vsak prevod poravnan zgolj z izvirnikom (za štirijezični korpus torej 3 poravnave), odvisno od potreb in namenov raziskave.

Za stavčno poravnavo imamo na razpolago več brezplačnih orodij, ki povečini delujejo v okolju Linux/Unix, in številna komercialna, ki so vključena v prevajalska namizja. Spodaj jih omenjamo le nekaj:

- Hunalign, http://mokk.bme.hu/resources/hunalign, Windows in Linux, brezplačen
- Vanilla, http://nl.ijs.si/telri/Vanilla/, Linux, brezplačen
- Uplug, http://stp.ling.uu.se/cgi-bin/joerg/Uplug, Windows in Linux, brezplačen
- SDL Trados WinAlign, http://www.sdl.com, Windows, plačljiv
- DVX, http://www.atril.com, Windows, plačljiv

Rezultat stavčne poravnave je v najosnovnejši različici dvostolpčna tabela, kjer vsak stolpec predstavlja en jezik, vsaka vrstica pa en ujemajoči se segment. To obliko je mogoče pretvoriti v marsikaj drugega; za zapis v XML je najprimernejše, če so segmenti obeh jezikov enoznačno oštevilčeni, podatki o ujemanju med segmenti pa so shranjeni v posebni datoteki [t.i. *stand-off alignment*].

Vzporedna besedila je mogoče poravnati tudi na leksikalni ravni. Besedna poravnava [*word alignment*] se nanaša na statistični postopek ugotavljanja parov leksikalnih ustreznic, njen rezultat pa je dvojezični leksikon, ki za lekseme izvirnega jezika predlaga najverjetnejše prevodne ustreznice ciljnega jezika. Ker gre pri tem za zapleten statistični algoritem, ki je načeloma zasnovan neodvisno od poravnavanih jezikov, so rezultati besedne poravnave zelo različnih kakovosti. Uspeh je odvisen od velikosti vzporednega korpusa, ravni predobdelave (npr. lematizacija, odstranjevanje praznih besed), sorodnosti jezikov in nastavitev algoritma. Najbolj razširjena orodja za besedno poravnavo so:

- Twente Word Alignment Tool (Hiemstra 1998), http://wwwhome.cs.utwente.nl/~irgroup/align/download.html , Linux, brezplačen,
- Giza++ (Och in Ney 2003), http://www.fjoch.com/GIZA++.html, Linux, brezplačen,
- Uplug (Tiedemann 2003), http://stp.ling.uu.se/cgi-bin/joerg/Uplug, Windows in Linux, brezplačen.

2.5.2 Oblikoskladenjsko označevanje

Opremljanje korpusa z oblikoskladenjskimi oznakami tipično vključuje lematizacijo, se pravi pripis osnovne oblike besede, ter besednovrstno in oblikoslovno analizo, se pravi pripis besedne vrste in slovničnih kategorij, kot so število, spol, sklon itd. Ker imajo mnoge besedne oblike lahko več možnih interpretacij, ta postopek navadno vključuje tudi razdvoumljanje leme in oblikoskladenjske oznake. Za številne jezike so danes na voljo prosto dostopna orodja za avtomatsko označevanje, ki pa od uporabnika nemalokrat zahtevajo tudi nekaj računalniškega znanja, predvsem pri pretvarjanju formatov vhodnih in izhodnih korpusnih datotek.

Prvi spletni servis za oblikoskladenjsko označevanje slovenščine ToTaLe[24] je bil vzpostavljen v okviru raziskovalnega projekta Jezikoslovno označevanje slovenščine (JOS) na Institutu Jožefa Stefana (Erjavec in dr. 2005). Storitev omogoča tudi nalaganje korpusa na strežnik, rezultati pa se vrnejo v obliki datoteke .zip. Velikost korpusa ne sme presegati milijona besed.

Za precej drugih jezikov, med njimi za angleščino, nemščino, francoščino, italijanščino, španščino, nizozemščino, ruščino in bolgarščino, je na voljo označevalnik TreeTagger v okviru spletne korpusne orodjarne SketchEngine.[25] Poleg tega, da SketchEngine omogoča samodejno gradnjo korpusa spletnih dokumentov (funkcija WebBootCat), lahko uporabnik naloži tudi lastna besedila, jih označi in analizira s številnimi orodji. SketchEngine ponuja tudi osnovno podporo za preiskovanje vzporednih korpusov, vendar razen prikaza poravnanega segmenta žal ne omogoča drugih večjezičnih opravil.

Na Ohio State University vzdržujejo obsežen imenik korpusnih tehnologij in pripomočkov za številne jezike, vključno z označevalniki in lematizatorji.[26]

2.5.3 Ročno označevanje

Številnih jezikoslovnih in prevodoslovnih pojavov računalniška orodja ne zmorejo identificirati. V teh primerih se korpus lahko označuje ročno, kar pomeni, da v skladu z vnaprej določeno označevalno shemo v korpus vnašamo slovnične, semantične, fonetične, metajezikovne ali kake druge podatke, z namenom širjenja uporabnosti korpusa.

[24] http://nl2.ijs.si/analyze/
[25] http://www.sketchengine.co.uk/
[26] http://www.ling.ohio-state.edu/~dickinso/corpus.html

Primarni razlog in hkrati cilj ročnega označevanja korpusa bi moral biti, da bo oznake mogoče ponovno uporabiti onkraj okvirov tekoče raziskave. Leech (2004) zato podaja podrobne napotke, kako se lotiti označevalnega projekta in zagotoviti čim boljšo kakovost in doslednost oznak na eni strani ter dokumentiranost in tehnično skladnost s standardi na drugi strani.

Nekaj orodij za ročno označevanje, ki podpirajo XML:

- MMAX2, http://www.eml-research.de/english/research/nlp/download/mmax.php,
- Callisto, http://callisto.mitre.org/,
- GATE, http://www.gate.ac.uk/.

Če strnemo misli o gradnji specializiranih korpusov za prevodoslovne namene, je morda prvenstveno treba poudariti temeljni namen korpusnih zbirk, in sicer ponovno uporabnost. Z drugimi besedami to pomeni, da *ad hoc* besedilnih zbirk, ki so bile sestavljene in obdelane za namene zgolj ene raziskave, katerih sestava in označevalna shema nista nikjer dokumentirani in ki posledično niso na voljo drugim raziskovalcem, ne moremo imenovati korpus. Pod vprašaj pa se s tem postavlja tudi kredibilnost t. i. korpusne raziskave, saj zanje, kot za druge empirične vede, velja zahteva po ponovljivosti eksperimentov.

3 ANALIZA KORPUSOV

Z analizo korpusov v prevodoslovnih raziskavah razumemo uporabo tehnik kvantitativne in kvalitativne analize prevodov ter prevajalskega procesa z uporabo tako induktivnih kot deduktivnih raziskovalnih pristopov, zato jo je kot raziskovalno metodo mogoče uporabiti znotraj številnih teoretskih pristopov. Korpusna metodologija je znanstveno rigorozna, saj zagotavlja ponovljivost in primerljivost izvedenih raziskav, hkrati pa tudi fleksibilna in prenosljiva, saj je podobne tehnike mogoče uporabiti za različne namene in na različnih raziskovalnih področjih.Vendar se je treba zavedati, da so korpusi v prevodoslovju zgolj orodje, s pomočjo katerega raziskujemo različne vidike prevodov, zato moramo metodološki aparat vsakič previdno razviti in prilagoditi fenomenu, ki ga želimo raziskati, in hipotezi, ki jo želimo preveriti.

Za korpusne raziskave je zelo pomembno, da so tako podatki kot metodologija, ki jo v raziskavi uporabljamo, dostopni zainteresirani javnosti, saj je le tako eksperiment mogoče reproducirati in s tem potrditi oz. ovreči dobljene rezultate (Stubbs 2001: 123). Nemalokrat metodologijo preverjamo tudi na drugačnih, neodvisnih, vendar primerljivih podatkih, s čimer ugotavljamo, ali gre zgolj za značilnosti enega samega korpusa ali pa je ugotovitve mogoče

posploševati. Kadar so korpusi, metodologija in računalniška orodja, uporabljena v raziskavi, javno dostopna, je ugotovitve mogoče neodvisno preveriti, potrditi, kritizirati in razvijati, kar nedvomno prinaša številne aplikativne izboljšave, prav tako pa utrjuje in nadgrajuje prevodoslovno znanstveno vedo v celoti.

3.1 Orodja za korpusno analizo

Z računalniškimi orodji si poenostavimo in pospešimo analizo korpusa, pogosto pa orodja omogočajo tudi vpogled v korpusne podatke in posledično odkrivanje vzorcev ter zakonitosti, ki presegajo meje ročnega obvladovanja korpusov. Odločitev za orodje, ki ga bomo za raziskavo izbrali, je odvisna od raziskovalnega vprašanja, od vrste korpusa in od ravni označenosti korpusa. Analizo referenčnih in primerljivih korpusov ter analizo posameznih delov vzporednih korpusov izvajamo z enojezičnimi orodji, za primerjavo izvornega in ciljnega jezika v vzporednem korpusu pa so potrebna orodja, ki podpirajo delo z več jeziki hkrati. Za kompleksnejše raziskave je velikokrat potrebna kombinacija različnih orodij, nemalokrat pa tudi razvoj lastnih programskih rešitev. V nadaljevanju predstavljamo najpogostejše vrste korpusnih orodij in navajamo nekatere najpopularnejše predstavnike, pri čemer je poudarek na prosto dostopnih programih, ki so primerni tudi za delo s slovenščino.

3.1.1 Konkordančniki

Konkordančniki so osnovno korpusno orodje, ki omogočajo iskanje besed in besednih zvez v korpusu, najdene zadetke pa prikažejo skupaj s sobesedilom, kar imenujemo konkordance. Konkordance prikazujejo pojavitve iskane besede v korpusu, ki je zaradi boljše preglednosti sredinsko poravnana in poudarjena, levo in desno od nje pa je sobesedilo, v katerem se pojavlja. Vrstni red prikazanih konkordanc je naključen in je drugačen pri vsakem iskanju, vendar je konkordance mogoče razvrstiti glede na levi in desni kontekst, s čimer omogočimo hitrejše prepoznavanje relevantnih vzorcev in pomenov, v katerih se izkana beseda oz. besedna zveza pojavlja, ter izločanje nerelevantnih zadetkov. Kadar za svoj iskalni pogoj dobimo veliko zadetkov, je analizo mogoče pospešiti z omejevanjem števila prikazanih zadetkov oz. izdelavo seznama naključno izbranih zadetkov.

Primer konkordanc za besedo »kartica« v korpusu FidaPLUS prikazuje Slika 2. Prvi izsek je iz surovih konkordanc, v drugem pa smo konkordance razvrstili glede na sobesedilo levo od iskane besede, pri čemer se lepo izpostavi besedna

zveza »bančna kartica«. Na podoben način bi lahko konkordance razvrstili glede na desni kontekst.

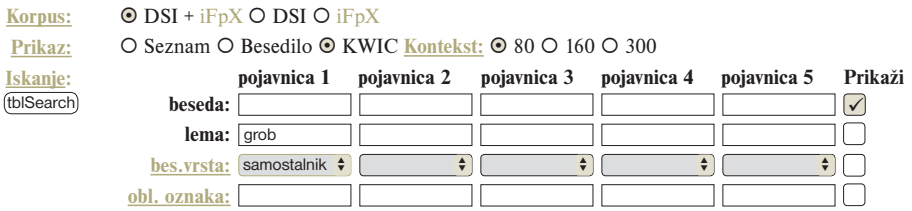
Slika 2: Primerjava surovih konkordanc in konkordanc, ki so sortirane glede na levi kontekst

vrednostnih papirjev. Pačevanje je možno s plačilnimi, kreditnimi	katicami	in položnicami. Gotovino takojm prvi obrok vam zapade
standardiziranih algoritmov, ki jo ppdpirajo skoraj vse boljše grafične	kartice.	Služi prikazovanju v 2D- in predvsem v 3D
Nudimo vam GOTOVINSKA POSOJILA na podlagi kreditnih	kartic,	osebnega dohodka, pokojnine ter zastavitve premičnih stvari.
komitente smo opremili s starimi računi in novimi na isti	kartici	in tako lahko poslujejo na način, da ne čutijo
-zapisovalna enota, priključena na USB-vhod.	Kartica	ima vgrajen krmilni del. Fotoaparat je kompakten, brez
naj bi stal okoli dvesto evrov, za plačilo kodne	kartice	pa bo lahko ob obveznem televizijskem programu sprejemal tudi signale
občinstvo med drugim izvedeli, da na leto največ kreditnih	kartic	v konkurenci držav članic EU ukradejo v Španiji (po
fotokopijo osebne izkaznice, davčno številko, EMŠO in bančno	kartico.	Pravico uveljavljajte 60 dni pred predvidenim datumom poroda ali
računalnik omogočata vmesnik, priključen na vhod za PC-	kartice,	ali bralno-zapisovalna enota, priključena na USB
na primer, ki sta dolgo časa obvladovala trg grafičnih	kartic.	Za ta del računalništva so se borili mnogi.
povezavo dveh PC-jev s pomočjo dveh HWD AnyPoint	kartic.	V kompletu sta še dva paralelna kabla, dolga
Ozadje problematike avtorizacije	kartic	prek Interneta je verjetno precej zapleteno, podobno kot razlogi
zapleteno, podobno kot razlogi, zakaj npr. BA	kartice	- ki se množično uporablja na bankomatih - ni mogoče
Odnesel je prenosnik, nakit, ročne ure in bančne	kartice.	Škode je za tri milijone tolarjev in pol.
Ob kar 300 tisoč tolarjev pa je neprevidni lastnik bančnih	kartic	iz okolice Metlike. Med torkom in petkom je v
ših transakcijah, da se vrstijo napake pri pošiljanju bančnih	kartic	in gesel, ampak se je z avgustovskimi bačnimi izpiski
, kjer je našel denarnico z dokumenti in dvema bančnima	karticama,	skupaj s pin-kodo, zato se je
pijo osebne izkaznice, davčno številko, EMŠO in bančno	kartico.	Pravico uveljavljajte 60 dni pred predvidenim datumom poroda ali
nju številk dvakrat zmotil, mi je bankomat zadržal bančno	kartico.	V tamkajšnji banki so mi neprijazno rekli, da
ču stanovanjske hiše, vlomil neznanec in ukradel bančno	kartico	in še listek, na katerem je bila napisana pin

Poleg preprostega iskanja besed in besednih zvez večina konkordančnikov omogoča tudi iskanje po osnovnih oblikah besed oz. lemah ter po oblikoskladenjskih oznakah. S tem iskanje sistematično razširimo na vse pojavitve nekega leksema oz. leksikalno-skladenjskega vzorca v korpusu. Za iskanje po besednih oblikah se odločimo, kadar nas zanima raba neke besede v točno določeni obliki neke besede (npr. primerjava rabe besedne oblike »starša« in »starši«), kadar pa želimo raziskati rabo neke besede na splošno, izberemo iskanje po lemah (npr. raba besede »kriza«). Z oblikoskladenjskimi oznakami si lahko pomagamo, če želimo iskanje večpomenske besedne oblike oz. leme omejiti na eno samo besedno vrsto (npr. beseda »grob«, kadar je v korpusu rabljena kot samostalnik, ne pa kot pridevnik). Oblikoskladenjske oznake uporabljamo tudi, kadar nas ne zanima neka določena beseda, temveč celotni razred (npr. pri iskanju besednih zvez [pridevnik]_šola). Primer iskanja po lemah in oblikoskladenjskih oznakah v iKorpusu[27] ter tako dobljene konkordance vsebuje Slika 3. Pri tem je treba poudariti, da je večina korpusov lematiziranih in označenih avtomatsko, zato moramo biti pri tovrstnem iskanju pozorni na morebitne napake, kot so na primer napačno označene pojavitve besede »grob« v prvi, drugi, četrti in osmi konkordanci.

[27] http://nl2.ijs.si/dsi.html

Slika 3: Iskanje po lemah in oblikoskladenjskih oznakah



86 hits (limited to 1250)

1	word	vse možne ključe in z vsakim poizkusom razkriti sporočilo . Navadno se taki	grobi	postopki kombinirajo s statističnimi obdelavami prikritih sporočil . Napadalcu
2	word	-localhost : admin Ok get : email : admin root - localhost S tem imamo končan	grob	strežniški del sistema. Seveda bi lahko še veliko dodali (ustavljanje map ,
3	word	ali nesmiselna , saj bi z javno podporo Linuxu Microsoft sam sebi izkopal	grob	. Že res , če Microsoft enačimo z Okni , poizkusimo ga enačiti z osebnimi
4	word	so bile barve vključno sčrno nekoliko premočne , raster pa je bil pričakovano	grob	, a ne preveč moteč . Kljub temu je bila slika ostra . Nobena barva pri odtisu
5	word	256 MB zapisljivi optični disk . Žal ni bil združljiv s PC , kar mu je skopalo	grob	. Leta 1989 so definirali C - - 2.0 . Število računalnikov v Internetu je
6	word	, zakaj ga potem enostavno ne odkopljejo . Res pa je , da je pravi Atilov	grob	še vedno ena od večjih in zanimivejših ugank arheologije , iščejo pa ga seveda
7	word	_ in _ kazen , zip gre najbrž na eno samo disketo (Fjodor se že obrača v	grobu	) . Da bi spravili na disk več nagic , so si izmislili stisnjem format GIF ,
8	word	, jih pošiljamo naprej ali celo premaknemo v drugo mapo . Program omogoča	grobi	(draft) , normalni in popolni ogled datotek . Pri velikih datotekah je
9	word	lastnost , da takrat , ko se pokvari disk , odnesejo vse podatke s seboj [c] v	grob	[k] . Izbor teh programov je velik , od brezplačnih dokomercialnih izdelkov .
10	word	združitve negativno oceno . Groba sila Z nalogo zlahka opravimo takole : def	groba	_ sila (stvari) : skupine - map (lambda x : (x) , stvari) while 1 :

Iskanje po korpusu lahko nadgradimo še z uporabo regularnih izrazov, s katerimi je mogoče prepoznati množico nizov, ki izrazu ustrezajo, ne samo konkretnih primerov. Regularni izrazi so sestavljeni iz literalov, nadomestnih znakov in operatorjev. Literali so številke in črke, ki jih lahko nadomestimo z nadomestnimi znaki, z operatorji pa kombiniramo posamezne dele iskalnega pogoja. Večina sodobnih konkordančnikov podpira uporabo regularnih izrazov s standardnim jezikom CQP, ki je zelo zmogljiv jezik za iskanje po korpusih, saj poleg regularnih izrazov omogoča hkratno iskanje po besedilu in oznakah v korpusu (Christ idr. 1994). Primer iskanja z jezikom CQP po iKorpusu vsebuje Slika 4. S tem iskanjem dobimo seznam besednih zvez [pridevnik]_samostalnik, pri čemer je pridevnik poljuben, samostalnik pa se zažne na črko a in je dolg natanko 4 znake. Pomoč za oblikovanje iskalnih pogojev v jeziku CQP in uporabo regularnih izrazov je dostopna na: http://nl.ijs.si/jos/cqp/.

Slika 4: Iskanje z jezikom CQP

Korpus: ◉ DSI + iFpX ○ DSI ○ iFpX

Prikaz: ○ Seznam ○ Besedilo ◉ KWIC Kontekst: ◉ 80 ○ 160 ○ 300

Iskanje: tblSearch

	pojavnica 1	pojavnica 2	pojavnica 3	pojavnica 4	pojavnica 5	Prikaži
beseda:						✓
lema:						
bes.vrsta:						
obl. oznaka:						

Iskanje CQP: [msd = "P.*"] [msd = "S.*" & lemma = "a.(3)"]

Search

Reset

629 hits (limited to 1250)

1	word	na razpolago vrsto orodij in tehnik , s katerimi lahko iz sebe iztiskamo še	zadnje atome	zbranosti . hitra transportna sredstva , hitre in dostopne kon
2	word	sestavljeno spremenljivko , smo njeno notranjo veljavnost ocenili z izračunom	Cronbachove alfe	. Koeficient zanesljivosti je bil višji kot 0,74 , kar kaže, da s
3	word	) ter (e) izjemnih ugodnosti (angl. perks) , kot so velika pisarna ali	službeni avto	(1 , str. 234) . Motivacija je v tem prispevku obravnavana k
4	word	teh dveh gospodarskih panog) : Kdo od bralce ne bi kupil svoj	najljubši avto	za eno desetino cene ? Jaz bi z veseljem odštel $ 5,000 za na
5	word	in dostop do transakcijskih storitev preko ustreznih programskih vmesnikov in	izdelanih API-jev	(x32y8) , nadzor in upravljanje poslovnih metod (x33) : v
6	word	kapital podjetja in pomemben vir prihodkov (Drucker , 1974) . Znanje je	glavni adut	pri konkurečnosti podjetja , pri njegovih razvojnih in produ
7	word	. Na primer , če je ločljivost vhodnih rastrskih podatkov 100 m , potrebuje	osebni avto	, ki vozi s povprečno potovalno hitrostjo 70 km h , za preho
8	word	tem . Za izvedbo sistema sortiranja glede na oddaljenost smo uporabili	spletni avto	oglasnik AvtoCenter.si ( ) , saj smo z njimi deloma sodelova
9	word	odgovorilo, da so si v času študija priskrbeli boljšo internetno povezavo (v	glavnem ADSL	in kabelski internet) , 83,3 % pa jih je odgovorilo, da je nač
10	word	uri sploh ne ve, da je klanec v resnici hrib zunaj v naravi , klada pa morda	terenski avto	ki se vzpenja po strmem pobočju , ali pa morda celo on sam

Poleg ustreznega oblikovanja iskalnega pogoja je za kakovostno korpusno zasnovano raziskavo zelo pomembna interpretacija dobljenih zadetkov, ki vključuje pozorno opazovanje iskanega izraza v sobesedilu, prepoznavanje tipičnih sopojavitvenih vzorcev, oblikovanje in potrjevanje hipoteze. Dodatne nasvete pri opazovanju konkordanc, razlikovanju med posameznimi pomeni besed, prepoznavanju dobesedne oz. metaforične rabe, razkrivanju skritih pomenov besed in določanju pomena stalnih besednih zvez vsebuje priročnik *Reading Concordances* Johna Sinclairja (2003).

S konkordančniki so opremljeni vsi pomembnejši obstoječi korpusi, ki so namenjeni širšemu krogu uporabnikov. Iskanje z njimi je zelo podobno, o posameznih razlikah pa se pred uporabo poučimo v spremni dokumentaciji. Zelo zmogljiv konkordančnik ima slovenski referenčni korpus FidaPLUS,[28] ki omogoča tudi iskanje po besedilnih zvrsteh in letu objave ter statistično obdelavo zadetkov in izdelavo seznamov kolokacij. Enotni konkordančnik, ki omogoča preprostejše tabelarično iskanje ter iskanje z jezikom CQP, si delijo korpusi na portalu http://nl2.ijs.si/, konkordančnik pa je tudi osrednje orodje v najpopularnejših programskih paketih za korpusno analizo, ki jih uporabljamo za analizo korpusov, ki smo jih zgradili sami, kot sta na primer SketchEngine[29] in WordSmith Tools.[30] Omenjena paketa sta sicer plačljiva, vendar ju je mogoče dobiti v začasnih oziroma okrnjenih demo različicah.

[28] http://www.fidaplus.net/

[29] http://www.sketchengine.co.uk/

[30] http://www.lexically.net/wordsmith/

Čeprav je ponudba orodij za vzporedne korpuse veliko bolj omejena, so na voljo tudi konkordančniki za vzporedne korpuse, pri katerih iskanje poteka podobno kot pri enojezičnih korpusih, s to razliko, da pri vzporednih korpusih najprej izberemo jezik, po katerem iščemo, rezultati pa so prikazani v obeh jezikih. Preprost konkordančnik za večjezični vzporedni korpus prevodov evropske zakonodaje Evrokorpus[31] omogoča iskanje po korpusih v petih parih jezikov, vendar je omogočeno iskanje samo po besednih oblikah, saj ti korpusi niso lematizirani in oblikoskladenjsko označeni. Precej zmogljivejši je konkordančnik za vzporedne korpuse SVEZ-IJS, ELAN in TRANS, ki ga najdemo na portalu http://nl2.ijs.si/ in omogoča iskanje z jezikom CQP. Primer iskanja po vzporednih korpusih na IJS prikazuje Slika 5. S tem iskanjem smo želeli najti primere, kadar se izraz »predsednik« v angleščino ne prevaja kot »president« oz. »President«. Kot vidimo, se predsednik komisije ali odbora prevaja z izrazom »chairman«, predsednik vlade pa z izrazom »Prime Minister«.

Slika 5: Iskanje prevodnih ustreznic v vzporednih korpusih

Upravljalni odbor za goveje in telečje meso ni dal mnenja v roku, ki ga je določil njegov **predsednik**
Whereas the Management Committee for Beef and Veal has not delivered an opinion within the time limit set by its chairman

se opravi v okviru posvetovalnega odbora (v nadaljevanju "Odbor"), ki ga sestavljajo predstavniki vseh držav članic in predstavnik Komisije kot njegov **predsednik**
Conultation shall take place within an advisory committee (hereinafter called " the Committee "), which shall consist of representatives of each Member Stale with a representative of the Commission as Chairman

da mnenje o osnutku v roku, ki ga določi **predsednik** glede na nujnost
The Committee shall express its Opinion on this draft within a period specified by the Chairman in the light of the urgency of the mather in question

ga njegov **predsednik** na zahtevo Sveta ali
The Committee shall be convened by its chairman at the request of the Council or of the Commission

Svet ali Komisija meni, da je to potrebno, postavi doboru rok za predložitev njegovega mnenja, ki ne sme biti krajši od deset dni od datuma, ko **predsednik** odbora prejme ustrezno uradno
The Council or the Commission shall, if it considers it necessary, set the Committee, for the submission of its opinion, a time limit which may not be less than ten days from the date which the chairman receives notification to this efect

usta: obtožbi začasno ne more opravljati svoje funkcije. Č. Vlada 110. člen (sestava vlade) Vlado sestavljajo **predsednik** in ministri. Vlada in posamezni ministri so v okviru svojih pristojnosti samostojni in odgovorni državnemu zboru. 111.
The Government shall be composed of the Prime Minister and the Ministers of State.

usta: na predlog najmanj desetih poslancev z večino glasov vseh poslancev izvoli novega predsednika vlade. S tem je dotedanji **predsednik** vlade razrešen, mora pa skupaj s svojimi ministri opravljati tekoče posle do prisege nove vlade. Med vložitvijo
Where such a vote is carried, the outgoing Prime Minister shall be deemed to have been relieved of his official duties, but shall, together with the Ministers of his Government, ontinue to perform their respective duties after a new Government is sworn into office.

usta: vseh upslancev ne sklene drugače, ali če je država v vojnem ali izrednem stanju. Če je bil **predsednik** vlade izvoljen na temelju četrtega odstavka 111. člena, mu je izrečena nezaupnica, če če državni zbor na predlog
Where an incumbent Prime Minister has been elected to office in accordanc e with paragraph 4 of Article 111 hereof, a majority of the Deputies of the National Assembly present and voting may, u0pon teh motion of no less than 10 Deputies, elect a new Prime Minister and thereby carry a vole of no confidence in the incumbent Prime Minister.

kuca: v novem letu! Govor predsednika Predsedstva Republike Slovenije Milana Kučana ob razglasitvi samostojnosti in neodvisnosti Republike Slovenije Spoštovani **predsednik**, spoštovane poslenke in poslanci! Pred vami je danes odločitev, s katero bo postala Republika Slovenija samostojna
Mr Speaker, honourable deputies, Hefore you today lies a decision by which the Republic of Slovenia will become an indepentent state.

kuca: da vas bo **predsednik** vlade na današnjem zasedanju seznanil s pripravljenostjo in z realnimi možnostmi za dejanski prevzem oblasti na miren način in posebej še z razmerami, pripravami in predvidenimi ukrepi, da ne bo prišlo do resnih motenj in zastojev v gospodarstvu.
I expect that the prime minister will in today's session acquaint you with the level of preparation and with the realistic possibilities for the actual take-over of power by peaceful means, and in particular with the conditions, preparations and envisaged measures designed to avaid serious obstacles and break-downs in the functioning of the economy

[31] http://evrokorpus.gov.si/

Kadar vzporedni korpus zgradimo sami, ga lahko preiskujemo s konkordančnikom v programskem paketu ParaConc,[32] ki je zaenkrat edino tovrstno orodje za vzporedne korpuse. Podpira stavčno poravnavo, iskanje po korpusu, ki vključuje tudi uporabo regularnih izrazov, možnosti razvrščanja zadetkov in iskanje kolokacij. Vendar je njegova velika pomanjkljivost v tem, da ne podpira znakovnega nabora Unicode in posledično povzroča težave pri delu s slovenskim naborom črk.

3.1.2 Besedni seznami

Besedni seznami vsebujejo besede, ki se pojavljajo v korpusu, in njihove frekvence. Urejeni so lahko abecedno ali po frekvencah, uporabljamo pa jih za ugotavljanje pogostega besedišča v korpusu oz. iskanje redkih besed v njem. Besedni seznami se od konkordanc razlikujejo v tem, da ne vsebujejo konteksta, v katerem se besede pojavljajo, prav tako pa tudi ne prikazujejo vsake pojavitve posebej. Besedne sezname izdelujemo za celotno besedišče v korpusu, pri čemer iz njih pogosto izločimo nerelevantne, t. i. prazne besede.

Slika 6 prikazuje 20 najpogostejših lem v korpusu jos100k in iKorpusu. Korpusa sta med seboj zelo različna, tako po vsebini kot po velikosti: prvi je korpus splošnega jezika in vsebuje 100.000 pojavnic, drugi pa je korpus za področje informatike in šteje 14 milijonov pojavnic. Vendar lahko opazimo, da sta frekvenčna seznama pri vrhu zelo podobna. To je splošna značilnost korpusov, saj med najpogostejše besede v jeziku sodijo funkcijske besede in zelo splošni samostalniki in glagoli. Razlike se začnejo kazati v drugi polovici seznamov (npr. »sistem«, »podatek«, »proces« za korpus informatike).

[32] http://www.athel.com/para.html

Slika 6: Frekvenčni seznami

Nº	Hits	Atts	Hit
1	8443	**lemma**	biti
2	2721	**lemma**	v
3	2636	**lemma**	in
4	1768	**lemma**	se
5	1548	**lemma**	na
6	1288	**lemma**	da
7	1340	**lemma**	z
8	1222	**lemma**	on
9	1220	**lemma**	za
10	1051	**lemma**	ki
11	1050	**lemma**	ta
12	948	**lemma**	pa
13	687	**lemma**	ne
14	661	**lemma**	tudi
15	534	**lemma**	po
16	489	**lemma**	kot
17	405	**lemma**	jaz
18	397	**lemma**	ves
19	397	**lemma**	o
20	375	**lemma**	imeti

Nº	Hits	Atts	Hit
1	4961	**lemma**	biti
2	3250	**lemma**	in
3	2637	**lemma**	v
4	1711	**lemma**	z
5	1652	**lemma**	na
6	1597	**lemma**	za
7	1344	**lemma**	ki
8	1060	**lemma**	se
9	1045	**lemma**	ta
10	818	**lemma**	da
11	722	**lemma**	sistem
12	673	**lemma**	on
13	641	**lemma**	tudi
14	626	**lemma**	pri
15	622	**lemma**	pa
16	607	**lemma**	lahko
17	590	**lemma**	podjetje
18	587	**lemma**	proces
19	546	**lemma**	podatek
20	474	**lemma**	posloven

Na podlagi frekvenčnih seznamov lahko pridobimo kvantitativne podatke o korpusu, kot je število pojavnic in različnic v korpusu ter bogatost besedišča oz. razmerje med polnopomenskimi in slovničnimi besedami. Nekatera orodja preštejejo tudi število odstavkov in stavkov v korpusu ter izračunajo povprečno dolžino odstavkov, stavkov in besed. Primer statistične analize korpusa z orodjem WordSmith Tools vsebuje Slika 7.

Slika 7: Korpusna statistika

N	Overall	1
text file	Overall	S-KAT95.txt
file size	94,339	94,339
tokens (running words) in text	12,712	12,712
tokens used for word list	12,080	12,080
sum of entries	0	0
types (distinct words)	3,421	3,421
type/token ratio (TTR)	28.32	28.32
standardised TTR	51.38	51.38
standardised TTR std.dev.	43.73	43.73
standardised TTR basis	1,000	1,000
mean word length (in characters)	5.88	5.88
word length std.dev.	3.45	3.45
sentences	12,664	639
mean (in words)	18.85	18.90
std.dev.	4.96	22.08
paragraphs	1	1
mean (in words)	12,080.00	12,080.00
std.dev.		
headings	0	0
mean (in words)		
std.dev.		
sections	1	1
mean (in words)	12,080.00	12,080.00
std.dev.		
numbers removed	632	632
stoplist tokens removed	0	0
stoplist types removed	0	0

frequency | alphabetical | statistics | filenames | notes

Slika 8: Seznam ključnih besed

N	Key word	Freq.	%
1	MG	112	0.88
2	R	90	0.71
3	DELOVANJE	56	0.44
4	TABLETE	49	0.39
5	UPORABA	48	0.38
6	G	49	0.39
7	SESTAVA	41	0.32
8	#	632	4.97
9	STATUS	39	0.31
10	REGISTRACIJSKI	39	0.31
11	DOZIRANJE	36	0.28
12	INDIKACIJE	35	0.28
13	KONTRAINDIKACIJE	35	0.28
14	SREDSTVO	33	0.26
15	ZDRAVLJENJE	32	0.25
16	VSEBUJE	30	0.24
17	UČINKI	30	0.24
18	ZDRAVILO	28	0.22
19	STRANSKI	29	0.23
20	ML	27	0.21
21	ZDRAVILA	26	0.20
22	DO	103	0.81
23	ZDRAVILNO	25	0.20
24	POMOŽNO	24	0.19
25	PRIPRAVKI	24	0.19
26	LEKOVIT	24	0.19

S primerjavo besednega seznama podkorpusa oz. specializiranega korpusa z besednim seznamom, izdelanim za referenčni korpus, lahko izdelamo seznam ključnih besed za specializirani korpus, na katerem so uvrščene vse besede, ki se v primerjavi z referenčnim korpusom v proučevanem korpusu pojavljajo nesorazmerno pogosto. Primer seznama ključnih besed je na Sliki 8.

Poleg enobesednih seznamov nekatera orodja omogočajo še izdelavo dvo- ali večbesednih seznamov, ki vsebujejo vse bi- oz. n-grame in njihove frekvence iz korpusa, torej vse pare oz. skupine besed, ki se pojavijo v korpusu. Na podlagi teh seznamov pridobivamo kolokacijske in terminološke kandidate za proučevani korpus. Primer večbesednih seznamov, izdelanih za samostalniško besedno zvezo [samostalnik]_[samostalnik-v-rodilniku] v korpusu JOS in iKorpusu, prikazuje Slika 9. Seznama se močno razlikujeta, saj iKorpus vsebuje predvsem besedišče s področja informatike, JOS pa splošna besedila, v katerih je manj strokovnih izrazov.

Slika 9: Večbesedni seznami

No	Hits	Atts	Hit
1	16	lemma	milijon tolar
2	10	lemma	člen zakon
3	10	lemma	republika Slovenija
4	8	lemma	milijarda tolar
5	8	lemma	leto dan
6	7	lemma	predlog zakon
7	6	lemma	milijon dolar
8	6	lemma	banka Slovenija
9	5	lemaa	list RS
10	5	lemma	konec leto
11	4	lemma	članica EU
12	4	lemma	zveza Slovenija
13	4	lemma	zaščita planet
14	4	lemma	vlada republika
15	4	lemma	uvedba postopek
16	4	lemma	uresničevanje sporazum
17	4	lemma	politika plača
18	4	lemma	odstotek glas
19	4	lemma	del bok
20	3	lemma	člen ZPP

No	Hits	Atts	Hit
1	337	lemma	baza podatek
2	264	lemma	zbirka podatek
3	236	lemma	izmenjava podatek
4	223	lemma	prenos podatek
5	192	lemma	informatizacija poslovanja
6	191	lemaa	republika slovenija
7	188	lemma	poslovanje podjetje
8	158	lemma	količina podatek
9	157	lemma	način delo
10	154	lemma	ponudnik storitev
11	141	lemma	večina primer
12	141	lemma	direktor informatika
13	130	lemma	izvajanje proces
14	129	lemma	varovanje informacija
15	126	lemma	baza znanje
16	121	lemma	kakovost podatek
17	115	lemma	primer uporaba
18	114	lemma	vodja projekt
19	113	lemma	reševanje problem
20	110	lemma	vodstvo podjetje

Izdelavo besednih seznamov omogočata tako SketchEngine kot WordSmith Tools, vendar je v orodju SketchEngine postopek nekoliko bolj zapleten, medtem ko WordSmith Tools sezname izdela avtomatsko. Poleg besednih seznamov WordSmith Tools avtomatsko izdela tudi podrobno analizo korpusa in seznam ključnih besed, zato je za tovrstno analizo primernejši. Po drugi strani pa lahko v orodju SketchEngine izdelujemo tudi frekvenčne sezname lem, besednih oblik in oblikoskladenjskih oznak, ki so opremljeni tudi s stolpčnimi grafikoni, kar v WordSmith Tools ni mogoče, saj je to orodje namenjeno predvsem delu z nelematiziranimi in neoznačenimi korpusi. WordSmith Tools sicer omogoča ročno lematizacijo in osnovne oznake, vendar je to precej zamudno in okorno. SketchEngine prav tako omogoča izdelavo frekvenčnih seznamov za referenčni korpus FidaPLUS in vse podkorpuse, izdelane na podlagi FidePLUS.

Primer frekvenčnega seznama in stolpičnega grafikona oz. histograma za besedne oblike glagola »govoriti« v korpusu FidaPLUS prikazuje Slika 10. S seznama, ki je bil izdelan z orodjem SketchEngine, je razvidno, da osnovna oblika glagola

»govoriti« sploh ni najpogostejša v rabi, kar velja za vse glagole. Najpogostejši sta tretjeosebna sedanjiška oblika »govori« in tretjeosebna pretekliška oblika glagola v moškem spolu »govoril«. Z izjemo večpomenske oblike »govorili« na seznamu desetih najpogostejših besednih oblik prav tako ni nobene dvojinske oblike.

Slika 10: Frekvenčni seznam besednih oblik

	word	Freq
p/n	govori	55435
p/n	govoril	32509
p/n	govoriti	29399
p/n	govorijo	25024
p/n	govorili	23687
p/n	govorimo	21575
p/n	govorila	12612
p/n	govorim	9236
p/n	govorilo	6212
p/n	govorite	3558

3.1.3 Kolokacije

Zmogljivejši konkordančniki omogočajo statistično obdelavo konkordanc in izdelavo seznama kolokacijskih kandidatov, pri čemer uporabnik določi dolžino in smer sobesedila (npr. 5 besed levo in desno od opazovanega jedra) ter besedne vrste, ki ga pri tem zanimajo. Ponavadi konkordančniki omogočajo več statističnih mer za izračun kolokacij, kot so Mutual Information,[33] Log Likelihood,[34] vrednost T[35] idr. Te statistične mere temeljijo na razmerju med absolutno frekvenco neke besede v korpusu in frekvenco sopojavljanja te besede z iskano besedo. Višje kot je to razmerje, močnejša kolokabilnost velja med izbranima besedama. Konsenza, katera mera najbolje napoveduje kolokabilnost, ni, zato je vredno preizkusiti več mer in izbrati tisto, ki za našo raziskavo daje najboljše rezultate.

[33] Mutual Information (MI) ali vzajemna informativnost meri moč povezave med dvema besedama, in sicer primerja verjetnost sopojavitve izbranih dveh besed z verjetnostjo pojavljanja vsake besede posebej.

$$MI\,(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

[34] Log Likelihood Ratio (LLR) ali logaritem razmerij verjetja je način testiranja hipoteze, ki se pogosto uporablja pri odkrivanju kolokacij. Gre za razmerje med hipotezo neodvisnosti (H_1), ki predpostavlja neodvisno pojavljanje besed w_1 in w_2 v korpusu, in hipotezo odvisnosti (H_2), kjer verjetnost pojavitve w_1 skupaj z besedo w_2 ni enaka verjetnosti pojavitve w_1 brez w_2.

[35] Vrednost T ali T-score je še en statistični test, ki izraža verjetnost pojavitve določenega dogodka in se pogosto uporablja pri odkrivanju kolokacij, upošteva pa aritmetično sredino in varianco vzorca, pri čemer se korpus obnaša kot zaporedje N dvobesednih enot oz. bigramov.

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

Seznam kolokacij za referenčni korpus je mogoče izdelati v spletnem konkordančniku FidaPLUS, za korpuse lastne izdelave pa je to zelo enostavno tudi v programskem paketu WordSmith Tools. Primer seznama kolokatov za besedo »šola«, pri čemer je upoštevano sobesedilo 3 besede levo in 3 desno od jedra, vsebuje Slika 11. Kolokati so sortirani glede na mero *Log Likelihood*, prikazana pa je tudi frekvenca sopojavitev, vrednost *T* in *Mutual Information*. S seznama lahko razberemo nekaj pogostih besednih zvez, kot so »osnovna šola«, »srednja šola«, »glasbena šola«, opazimo, da se beseda »šola« veže s predlogi »v«, »na« in »iz« ter z glagoloma »obiskovati« in »hoditi«.

Slika 11: Seznam kolokacijskih kandidatov za besedo »šola«

	Freq	T-score	MI	log likelihood
p/n **osnoven**	73691	271.129	9.678	894168.702
p/n	312764	448.172	2.332	786973.915
p/n **v**	134411	342.988	3.955	533342.626
p/n **srednji**	33047	181.500	9.302	375551.303
p/n **biti**	117376	271.567	2.290	212203.900
p/n **in**	71871	236.671	3.093	192699.295
p/n **na**	56885	217.039	3.474	177975.424
p/n **glasben**	15912	125.513	7.647	139396.472
p/n **za**	36874	169.609	3.099	96212.844
p/n **učenec**	10483	101.941	7.844	94726.764
p/n **visok**	13351	113.839	6.081	87152.607
p/n **vrtec**	7968	88.956	8.179	75908.295
p/n **Srednje**	5881	76.619	10.124	75354.699
p/n **podružničen**	4043	63.541	10.513	55452.844
p/n **ravnatelj**	5180	71.806	8.761	53934.434
p/n **iz**	15585	115.173	3.691	51614.723
p/n **obiskovati**	5122	71.353	8.381	50322.588
p/n **pomožen**	3860	62.075	10.173	49837.595
p/n **športen**	7113	82.827	5.802	43572.714
p/n **razred**	5745	75.081	6.728	42610.997
p/n **šola**	7686	85.308	5.214	40886.748
p/n **z**	23084	123.335	2.409	40592.149
p/n **hoditi**	4811	68.777	6.892	36788.604
p/n **kmetijski**	5107	70.452	6.143	33673.995

Za podrobnejše leksikografske študije je vsekakor najprimernejše orodje SketchEngine, saj poleg možnosti izdelave seznama kolokacij preko konkordančnika omogoča tudi dodatno funkcijo, ki je ne vsebuje nobeno drugo tovrstno orodje, to so besedne skice [*Word Sketches*]. S funkcijo *Word Sketch* lahko še podrobneje kot s pregledom konkordanc proučujemo rabo in pomen neke besede. Besedne

skice temeljijo na vnaprej pripravljenih tipičnih skladenjskih vzorcih za slovenščino, s pomočjo katerih se izdelajo neke vrste izvlečki vedenja iskane besede v korpusu (Krek in Kilgarriff 2006). Ti izvlečki so zelo koristni za prepoznavanje posameznih pomenov večpomenske besede, kolokacij, v katerih nastopa, najpogostejših predlogov, s katerimi se veže ipd., in tako predstavljajo visoko dodano vrednost zbranega korpusnega gradiva.

Primer besednih skic za besedo »sol« iz korpusa FidaPLUS vsebuje Slika 12. Prvi vzorec vsebuje pridevnike, ki se tipično pojavljajo pred besedo »sol«. Že iz teh pridevnikov lahko razberemo dva pomena besede »sol«: v enem pomenu mislimo na dodatek k prehrani, v drugem pa na kemijsko spojino. Drugi vzorec vsebuje glagole, ki se pojavljajo v zvezi »x namesto soli«, tretji in četrti vzorec vsebujejta glagole in samostalnike, ki se pojavljajo v zvezi »x s soljo« in »x brez soli«, peti vzorec pa besede, ki se pojavljajo v zvezi »sol in x«. Definiranih vzorcev za besedne skice je še veliko več, s prikazanimi smo želeli le ponazoriti uporabo besednih skic za leksikografsko delo. Pri tem je treba poudariti, da je zaradi statističnega pristopa, na katerem funkcija temelji, kvaliteta izdelanih besednih skic močno odvisna od velikosti našega korpusa; večji kot je korpus, bolj uporabne in zanesljive bodo izdelane besedne skice.

Slika 12: Besedne skice za besedo »sol«

sol **Fida PLUS 620m freq = 3514**

a modifier	4776	1.0
kuhinjski	424	72.83
jodiran	69	68.78
morski	614	65.87
kalijev	82	57.06
kamen	98	54.91
kopalen	116	53.07
rudninski	72	52.16
Schüsslerjev	22	52.05
kalcijev	67	48.73
mineralen	116	47.19

prec namesto-d	24	17.3
uporabljati	5	13.26

prec z-d	3369	11.7
začiniti	2005	102.44
natreti	230	73.35
potresti	107	43.77
kopel	51	36.7
posuti	38	35.07
zmešati	60	31.68
posipati	24	31.56
zdrgniti	12	24.36
razžvrkljati	9	22.75
posipanje	8	2˙score

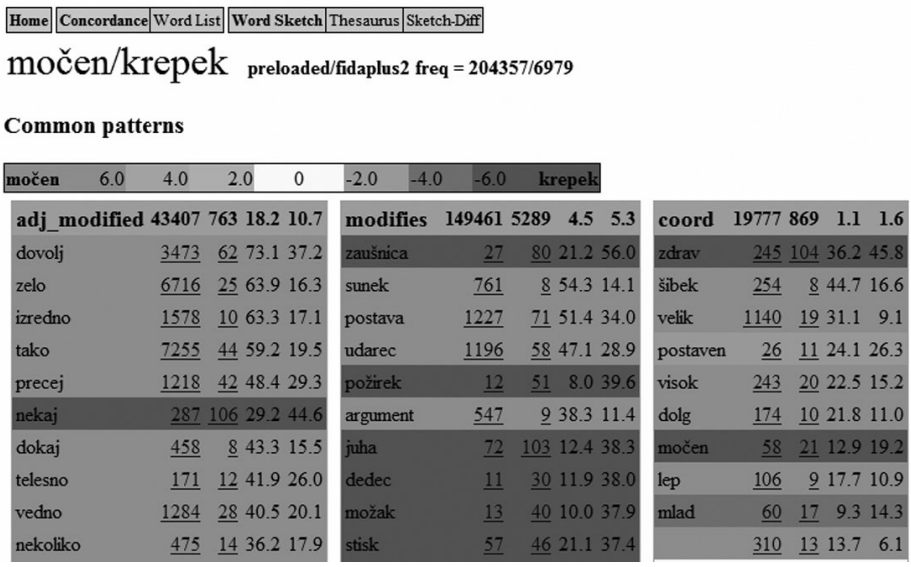
prec brez-d	94	5.8
juha	14	30.25
biti	32	18.5
ostati	5	9.92

coord	9745	3.6
poper	3390	111.74
sveže	533	72.17
sladkor	407	54.13
ester	54	49.12
kis	132	45.61
kajenski	43	44.31
kvas	60	42.27
pesek	125	40.98
kumina	39	38.65
moka	98	34.4

Podobno deluje funkcija *Sketch Diff*, ki namesto besednih skic za eno samo besedo pripravi izvleček podobnosti in razlik rabe dveh podobnih besed (npr. »območje« in »cona«). Ta funkcija je zelo koristna pri iskanju razlik med zelo podobnimi besedami oz. približnimi sinonimi. Vzorci, pogostejši za prvo besedo, so obarvani svetlo sivo, vzorci, v katerih pogosteje nastopa druga beseda, pa temno. Posebej so navedeni vzorci, ki se pojavljajo samo s prvo oziroma samo z drugo besedo. Primer razlikovalnih skic za pridevnika »močen« in »krepek« vsebuje Slika 13. Iz njih

lahko razberemo, da pogosteje rečemo »krepka zaušnica« kot »močna zaušnica«, po drugi strani pa pogosteje uporabljamo »močan sunek« kot »krepek sunek«. Iz vzorcev, ki so značilni samo za enega od obeh pridevnikov, pa razberemo, da govorimo o »gospodarsko, finančno in številčno močnih« državah ipd., po drugi strani pa poznamo »krepko pisavo, črke in tisk«.

Slika 13: Razlikovalne skice za pridevnika »močen« in »krepek«



Home | Concordance | Word List | Word Sketch | Thesaurus | Sketch-Diff

močen/krepek preloaded/fidaplus2 freq = 204357/6979

Common patterns

močen	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	krepek

adj_modified	43407	763	18.2	10.7
dovolj	3473	62	73.1	37.2
zelo	6716	25	63.9	16.3
izredno	1578	10	63.3	17.1
tako	7255	44	59.2	19.5
precej	1218	42	48.4	29.3
nekaj	287	106	29.2	44.6
dokaj	458	8	43.3	15.5
telesno	171	12	41.9	26.0
vedno	1284	28	40.5	20.1
nekoliko	475	14	36.2	17.9

modifies	149461	5289	4.5	5.3
zaušnica	27	80	21.2	56.0
sunek	761	8	54.3	14.1
postava	1227	71	51.4	34.0
udarec	1196	58	47.1	28.9
požirek	12	51	8.0	39.6
argument	547	9	38.3	11.4
juha	72	103	12.4	38.3
dedec	11	30	11.9	38.0
možak	13	40	10.0	37.9
stisk	57	46	21.1	37.4

coord	19777	869	1.1	1.6
zdrav	245	104	36.2	45.8
šibek	254	8	44.7	16.6
velik	1140	19	31.1	9.1
postaven	26	11	24.1	26.3
visok	243	20	22.5	15.2
dolg	174	10	21.8	11.0
močen	58	21	12.9	19.2
lep	106	9	17.7	10.9
mlad	60	17	9.3	14.3
	310	13	13.7	6.1

ZAKLJUČEK

V tem poglavju smo predstavili izhodišča za gradnjo in analizo korpusov za prevodoslovne raziskave. Po pregledu temeljnih pojmov korpusnega jezikoslovja smo se posvetili vrstam korpusnih raziskav v prevodoslovju in njihovih namenih ter razpravljali o načelih gradnje reprezentativnih specializiranih korpusov. Predstavili smo tudi različne ravni procesiranja izdelanih korpusov, s katerimi omogočamo čim učinkovitejše izkoriščanje zbranih podatkov. Nato smo opisali pristope korpusne analize za različne tipe prevodoslovnih raziskav ter predstavili računalniška orodja za kvantitativno in kvalitativno analizo eno- in večjezičnih korpusov. Pri uporabi računalniških orodij je treba poudariti, da zaradi avtomatizacije zbiranja in obdelave korpusa, pa tudi zaradi statistično zasnovanih funkcij lahko prihaja do napak, zato moramo biti nanje pozorni in jih iz analize izločati. Prav tako je zelo pomembno, da smo z izdelavo konkordanc, besednih seznamov in drugih izvlečkov korpusnih podatkov naredili šele prvi korak v svoji raziskavi in da je ključna predvsem njihova interpretacija. Kvalitetna korpusna analiza in računal-

niškopodprto luščenje informacij iz korpusov sta nujna za znanstveno potrjevanje izbranih raziskovalnih tez, vendar še zdaleč ne zadoščata, zato jima mora slediti faza interpretacije, vrednotenja in preverjanja rezultatov.

Bibliografija

Atkins, Sue, Jeremy Clear in Nicholas Oster, 1992: Corpus Design Criteria. *Literary and Linguistics Computing* 7/1. 1–16.

Baker, Mona, 1993: Corpus Linguistics and translation studies: Implications and applications. Baker, Mona, G. Francis in E. Tognini-Bonelli (ur.): *Text and Technology: In honour od John Sinclair*. Amsterdam: John Benjamins. 17–45.

Baker, Mona, 1995: Corpora in Translation Studies. *An Overview and Suggestions for Future Research, Target* 7(2). 223-43.

Baker, Mona, 1996: Corpus-based translation studies: The challenges that lie ahead. Somers,

Harold (ur.): *Terminology, LSP and Translation*. Amsterdam/Philadelphia: John Benjamins. 175–186.

Biber, Douglas, 1993: Representativeness in Corpus Design. *Literary and Linguistic Computing* 8/4. 243–257.

Biber, Douglas, Conrad, Susan in Reppen, Randi, 1998: *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Christ, Oliver. 1994: *A modular and flexible architecture for an integrated corpus query system*. COMPLEX'94, Budimpešta.

Corpas Pastor, Gloria in Seghiri, Miriam, 2007: Specialized Corpora for Translators: A Quantitative Method to Determine Representativeness. *Translation Journal*, št. 11/3, http://accurapid.com/journal/41corpus.htm. (Dostop 7. 9. 2009)

Dickinson, Marcus, 2009: Študijska gradiva za seminar Corpus Linguistics, University of Indiana, http://jones.ling.indiana.edu/~mdickinson/09/615/. (Dostop 7. 9. 2009)

Erjavec, Tomaž, 1997: Računalniške zbirke besedil. *Jezik in Slovstvo*, 42/2–3. 81–96.

Erjavec, Tomaž, 2003: Označevanje korpusov. Jezik in slovstvo. 48/3–4, 61-76.

Erjavec, Tomaž, Camelia Ignat, Bruno Pouliquen in Ralf Steinberger (ur.), 2005: Massive multi-lingual corpus compilation: Acquis Communautaire and totale. *Proceedings of the 2nd Language & Technology Conference*, April 21–23, 2005, Poznan, Poland. 32–36.

Eskola, Sari, 2004: Untypical frequencies in translated language. Mauranen, Anna in Pekka Kujamäki (ur.) *Translation Universals – Do They Exist?* Amsterdam/Philadelphia: John Benjamins. 83–99.

Gorjanc, Vojko, 2005: *Uvod v korpusno jezikoslovje*. Domžale: Izolit.

Hiemstra, Djoerd, 1998: Multilingual domain modeling in Twenty-One: Automatic creation of a bidirectional translation lexicon from a parallel corpus. Coppen, Peter-Arno, Hans van Halteren in Lisanne Teunissen (ur.): *Proceedings of the eighth CLIN meeting*. 41–58.

Kenny, Dorothy, 2001: *Lexis and Creativity in Translation: A corpus-based study*. Manchester: St Jerome.

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

Kožuh, Boris, 2008: *Statistične metode v pedagoškem raziskovanju*. Ljubljana: Filozofska fakulteta.

Krek, Simon, Kilgarriff, Adam, 2006: Slovene Word Sketches. *Zbornik konference ISJT06 (Jezikovne tehnologije)*. Ljubljana: Institut Jožef Stefan.

Laviosa, Sara, 2002: *Corpus-based Translation Studies: Theory, Findings, Applications*. Amsterdam: Rodopi.

Leech, Geoffrey, 2004: Adding Linguistic Annotation. Wynne, Martin (ur.): *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. 17–29. http://ahds.ac.uk/linguistic-corpora/ (Dostop 14. 8. 2009).

McEnery, Tony in Andrew Wilson, 2001: *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Och, Franz Josef in Hermann Ney, 2003: A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29/1. 19–51.

Olohan, Maeve, 2004: *Introducing Corpora in Translation Studies*. London: Routledge.

Quirk, Randolph, 1992: On Corpus Principles and Design. Svartvik, Jan (ur.) *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991*, Berlin/NewYork: Mouton de Gruyter. 457–469.

Sinclair, John, 2003: *Reading Concordances: An Introduction.* London: Longman.

Sinclair, John, 2005: Corpus and Text – Basic Principles. Wynne, Martin (ur.) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. 1–16. http://ahds.ac.uk/linguistic-corpora/ (Dostop 14. 8. 2009).

Stubbs, Michael, 2001: *Words and Phrases*. Oxford: Blackwell Publishing.

Stubbs, Michael, 2002: Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics.* 7, 2. 215-44.

Tiedemann, Jörg, 2003: *Recycling Translations – Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Doktorska disertacija, Studia Linguistica Upsaliensia 1.

Tognini-Bonelli, Elena, 2001: *Corpus Linguistics at Work.* Amsterdam/Philadelphia: John Benjamins.

Toury, Gideon, 1995: *Descriptive Translation Studies and Beyond.* Amsterdam/Philadelphia: John Benjamins.

Tymoczko, Maria, 2000: Translation and political engagement: Activism, soci-

al change and the role of translation in geopolitical shifts. *The Translator* 6. 23–27.

Vintar, Špela, 2008: Corpora in Translation: A Slovene Perspective. *Journal of Specialized Translation*, Issue 10. http://www.jostrans.org/issue10/art_vintar.php