Avtomatsko luščenje leksikalnih podatkov iz korpusa

Iztok Kosem*, Polona Gantar, Simon Krek*****

* Trojina, zavod za uporabno slovenistiko
Partizanska cesta 5, SI-4220 Škofja Loka
iztok.kosem@trojina.si
** Inštitut za slovenski jezik Frana Ramovša
Novi trg 4, SI-1000 Ljubljana
apolonija.gantar@guest.arnes.si
*** Amebis d.o.o.
Bakovnik 3, SI-1241 Kamnik
simon.krek@guest.arnes.si

Povzetek

Pri oblikovanju leksikalne baze za slovenščino v okviru projekta *Sporazumevanje v slovenskem jeziku* smo del gesel izdelali s postopkom avtomatskega luščenja leksikalnih podatkov iz korpusa Gigafida prek orodja Sketch Engine neposredno v program za izdelavo leksikalne baze iLex. V prispevku opisujemo posamezne korake pri pripravi avtomatizacijskega postopka, zlasti prilagoditev slovnice besednih skic, izdelavo konfiguracije za aplikacijo GDEX za izbor dobrih korpusnih zgledov in pripravo API skripte. Na kratko predstavimo prve rezultate izvedenega postopka in predlagamo izboljšave tako na ravni metodologije kot tudi vključitve dodatnih jezikovnotehnoloških funkcionalnosti pri avtomatičnem luščenju jezikovnih podatkov.

Automatic extraction of lexical information from a corpus

A selection of entries in the lexical database for Slovene, an activity within the Communication in Slovene project, was compilled using the automatic extraction of lexical information from the Gigafida corpus (via the Sketch Engine corpus tool) and importing the obtained information directly into the dictionary writing system iLex. The paper describes individual steps in the preparation of the automatic extraction procedure, especially the adjustment of sketch grammar, development of the GDEX configuration for the selection of good corpus examples, and the programming of the API script. We briefly present the initial results and suggest improvements in methodology of automatic extraction of lexical data, and inclusion of additional language technologies.

1. Uvod

Sodobne tehnologije so prinesle velike spremembe na področju leksikografije, tako za slovarske uporabnike kot za leksikografe. Nova generacija slovarskih uporabnikov, ki je vešča uporabe elektronskih medijev, zlasti spleta, zahteva zelo hiter in uporabniku prijazen dostop do podatkov o sodobnem jeziku, s tem pa posredno odloča tudi o vrsti, obliki in načinu njihove predstavitve. Če uporabniki česa ne najdejo v (spletnem) slovarju – do katerih največkrat dostopajo prav tako prek spletnih iskalnikov (Lorentzen in Theilgaard, 2012) –, obstaja velika verjetnost, da bodo podatek skušali poiskati na spletu.

Težava, s katero se soočajo leksikografi, je, da adekvaten opis jezika še vedno zahteva precej časa, tudi zaradi vse večjih besedilnih korpusov, ki jih je potrebno analizirati. Zato razvijalci leksikografskih orodij iščejo nove rešitve, ki bi leksikografom olajšale delo in jim pomagale leksikografske podatke čim hitreje spraviti do uporabnika. V zadnjih letih so bila tako razvita orodja, kot je Sketch Engine s funkcijami Besedna skica (ang. Word sketches; Kilgarriff in Rundell, 2002) in Kliksikografija (ang. Tickbox lexicography), od katerih prva nudi hiter pregled slovničnih in kolokacijskih vzorcev, v katerih se besede sopojavljajo, druga pa hiter in preprost prenos želenih podatkov v slovarsko orodje. Tovrstna orodja smo s pridom uporabili tudi pri oblikovanju Leksikalne baze za slovenščino (LBS), ki je potekalo v obdobju 2008–2012 pri projektu *Sporazumevanje v slovenskem jeziku* (http://www.slovenscina.eu/; SSJ[1]).

Vendar pa mora leksikograf kljub uporabi omenjenih polavtomatskih funkcionalnosti še vedno analizirati velike količine podatkov, na podlagi katerih se odloča o njegovi relevantnosti in primernosti za vključitev v podatkovno bazo ali slovar. Zamuden pa je tudi postopek vnašanja podatkov v slovarsko orodje, saj ne obstaja neposreden prehod med Besednimi skicami in različnimi strukturami slovarskih baz ali priročnikov.

Pri gradnji LBS smo predvideli, da bi se proces izdelave gesla precej skrajšal, če bi imel leksikograf možnost avtomatsko izluščiti relevantne podatke o besedi neposredno iz korpusa v program za izdelavo slovarja, jih pregledati, selekcionirati in po potrebi dopolniti. Dodatna analiza korpusa ne bi bila več potrebna oz. bi se zreducirala na preverjanje aktualne rabe.

Zamisel smo začeli uresničevati ob zaključku aktivnosti, saj je bilo pomembno, da sta bila vrsta leksikalno-gramatičnih podatkov in način njihove organizacije v LBS dokončno opredeljena z DTD strukturo in da je bila zadostna količina gesel ročno izdelana Na ta način je bilo mogoče predvideti optimalne rešitve, ki bi jih prinesel postopek avtomatizacije.

V tem prispevku opisujemo posamezne korake pri vzpostavljanju postopka avtomatskega luščenja podatkov za izdelavo slovarskih gesel ter prva opažanja.

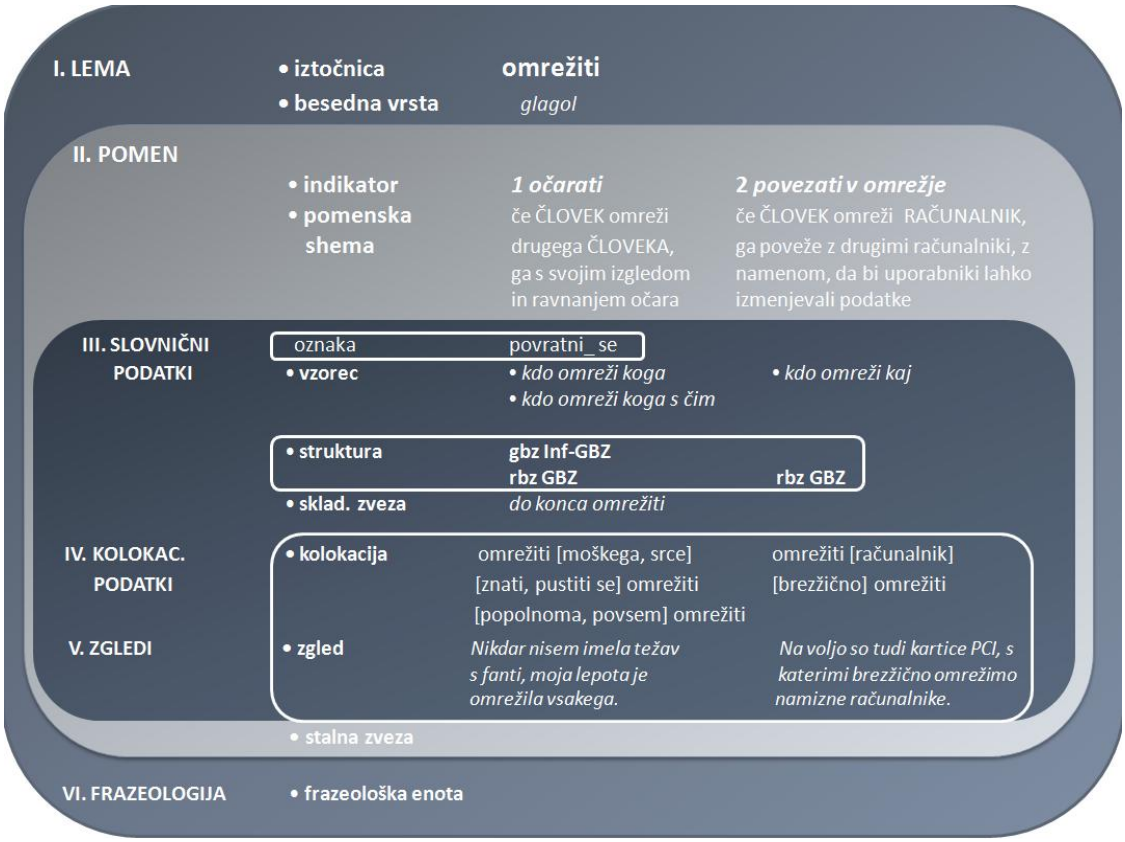
[1] Operacijo, v okviru katere je nastala raziskava, delno financira Evropska unija iz Evropskega socialnega sklada ter Ministrstvo za izobraževanje, znanost, kulturo in šport Republike Slovenije.

2. Leksikalna baza za slovenščino glede na avtomatsko luščenje leksikalnih podatkov iz korpusa

Leksikalna baza za slovenščino je zasnovana z dvema temeljnima ciljema: (a) zagotoviti relevantne leksikografske podatke, zlasti pomenske, skladenjsko-pomenske, kolokabilne, besednozvezne, frazeološke ipd. za izdelavo sodobnih jezikovnih priročnikov za slovenščino in (b) izdelati jezikovni vir, ki bo zagotavljal računalniško procesljive jezikovne podatke, ki so interpretirani in označeni v skladu z jezikoslovnimi kategorijami (Gantar in Krek, 2011). Celotna leksikalna baza je na voljo v formatu XML, posamezni njeni deli pa so določeni z DTD strukturo, tako da jih je mogoče avtomatsko identificirati. S tem naj bi LBS služila izdelavi in izboljšanju orodij za avtomatsko analizo jezika, kot je skladenjsko razčlenjevanje, večnivojsko označevanje, pomensko razdvoumljanje besed itd. Tako strukturirane podatke pa smo uspešno uporabili tudi pri postopku avtomatizacije.

Jezikovni podatki so v LBS strukturirani na šestih medsebojno povezanih nivojih: najvišji nivo vključuje lemo, podatke o besedni vrsti in korpusni frekvenci. Sledi semantični nivo, ki predvideva pomenske podatke v treh različnih oblikah: kot *indikatorje* za oblikovanje pomenskega menija, kot t. i. *pomenske sheme*, izdelane po vzoru baze FrameNet s formaliziranimi semantičnimi tipi na posameznih mestih prototipičnega vezljivostnega vzorca in kot *stavčne definicije*. Vsaka leksikalna enota, kamor štejemo (pod)pomene, stalne zveze in frazeološke enote, je na skladenjskem nivoju opredeljena glede na to, ali nastopa v določenih zanjo tipičnih *skladenjskih strukturah*, pri glagolih pa je ta podatek razviden iz *stavčnih vzorcev*. Na kolokacijskem nivoju sledijo skladenjskim strukturam *kolokacije*, ki so tipične leksikalne zapolnitve skladenjskih struktur, in t. i. *skladenjske zveze*, ki so pomensko in strukturno ustaljeni delčki jezika, kot na primer: *[deževati] od jutra do večera; [angažirati se] v smeri česa; [vožnja, voziti] pod vplivom alkohola*. *Stalne zveze* in *frazeološke enote* so od pomenov neodvisne in imajo lastno notranjo zgradbo, ki predvideva vse elemente na enak način kot posamezni pomeni besede. Vsi našteti podatki so potrjeni s *korpusnimi zgledi*.

Kot je razvidno iz spodnje slike, smo s postopkom avtomatizacije pridobili podatke na ravni skladenjskih struktur, pripadajočih kolokacij in relevantnih korpusnih zgledov (v okvirčkih). Poleg tega še podatke o tipičnem skladenjskem ali besedilnem obnašanju leme v korpusu, kot je denimo sopojavljanje z lastnimi imeni ali količinskimi izrazi, možnost tretjeosebne rabe glagola ali nastopanje v *se*-glagolskih ali citatnih konstrukcijah.



Slika 1: Podatki, pridobljeni s pomočjo avtomatskih postopkov, v strukturi LBS

3. Avtomatsko luščenje leksikalnih podatkov

Odločitev za izvedbo postopka avtomatičnega luščenja leksikalnih podatkov iz korpusa (ALLP) izhaja iz potrebe po skrajšanju časa in zmanjšanju stroškov pri izdelavi slovarskih priročnikov ter po drugi strani iz novih možnosti, ki jih ponujajo sodobna orodja za analizo velikih besedilnih korpusov.

3.1. Metodologija in zaporedje postopkov

Ideja ALLP predvideva prenos relevantnih leksikalnih podatkov iz korpusa Gigafida prek orodja Sketch Engine (Kilgarriff et al., 2004; SkE) oz. aplikacije Besedne skice in v leksikalni bazi registriranih skladenjskih struktur z uporabo API skripte (ang. Application Programming Interface) v program iLex (Erlandsen, 2004), v katerem se izdeluje LBS. Kot relevantne leksikalne podatke upoštevamo kolokabilno okolje besede ter dobre, tj. berljive ter pomensko in skladenjsko relevantne korpusne zglede, ki kolokacije potrjujejo v realnem besedilnem okolju. Te podatke zagotavlja aplikacija Besedne skice v orodju SkE, vendar pa je bilo za njihovo avtomatsko luščenje potrebno izdelati oz. slovenščini prilagoditi slovnico besednih skic (Krek, 2012) in pripraviti konfiguracijo orodja GDEX za pridobivanje dobrih korpusnih zgledov. Pred začetkom ALLP smo s sodelavci ekipe SkE v Brnu pripravili še API skripto, prilagodili DTD leksikalne baze novim elementom in atributom, specifičnim za ALLP, namestili korpus Gigafida v orodje SkE ter pripravili izbor relativno frekventnih in po možnosti enopomenskih lem za prvo fazo avtomatskega luščenja.

3.2. Izbor lem

Zaradi obvladljivosti količine podatkov pri evalvaciji izluščenih gesel in posledično zaradi možnosti postopnega izboljševanja nastavitev orodja GDEX in API skripte smo pri naboru lem upoštevali tri parametre: (a) relevantno frekvenco za posamezno besedno vrsto, kar pomeni predvsem dovolj obsežno besedno skico; (b) potencialno enopomenskost glede na slovenski Wordnet (Fišer, 2009; sloWNet) in SSKJ; (c) vključenost v sloWNet zaradi možnosti nadaljnjih povezav in nevključenost v SSKJ zaradi potencialnih novih izrazov in pomenov. Pri prvih poskusih ALLP smo se omejili na manj pogoste leme (pribl. 600 pojavitev v korpusu Gigafida), vendar pa so izdelane besedne skice pokazale premalo relevantnih podatkov, zato smo v nadaljevanju pri posamezni besedni vrsti določili do pet frekvenčnih skupin, znotraj njih pa smo se osredotočili na tiste frekvenčne razpone, ki so zagotavljali obvladljivo število lem ter hkrati ponudili optimalno besedno skico za pomensko relativno nerazvejano lemo.

Pri določanju enopomenskih ali pomensko manj zahtevnih lem smo se oprli na stanje, kot je izpričano v sloWNetu in SSKJ: izbrali smo leme z enim ali dvema sopomenskima nizoma (sinsetoma) v sloWNetu in/ali iztočnice z enim ali dvema pomenoma v SSKJ. Znotraj posamezne frekvenčne skupine – na končnem seznamu prevladujejo leme s pogostostjo v razponu med 1000 in 10.000 pojavitev, nekaj pa je tudi redkejših oziroma pogostejših lem, ki smo jih vključili za namene dodatnega testiranja ALLP, pogostejša gesla pa tudi zato, da bi lahko preverili delovanje API skripte za vse relacije v slovnici besednih skic – smo s prekrižanjem preostalih parametrov izdelali seznam s 515 samostalnškimi, 260 glagolskimi, 275 pridevniškimi in 117 prislovnimi lemami.

3.3. Slovnica besednih skic

Za potrebe ALLP iz korpusa Gigafida je bila izdelana nova slovnica besednih skic (ang. sketch grammar), ki izkorišča tudi nekatere elemente, ki so bili v orodje Sketch Engine dodani v novejših različicah. Med njimi so predvsem t. i. direktive[2] (directives) *CONSTRUCTION, *COLLOC in *SEPARATEPAGE. Prva omogoča prepoznavanje skladenjskih struktur brez kolokacij, kar je primerno predvsem za luščenje glagolskih vezljivostnih vzorcev. Druga je namenjena izločanju elementov, ki v LBS spadajo v kategorijo skladenjskih zvez, denimo zveza predlog-samostalnik-predlog (primer: *v primerjavi z, v odnosu do*), tretja pa je namenjena odpiranju relacij s tremi elementi (direktiva *TRINARY) na novi spletni strani, kar omogoča uvedbo natančnejših relacij s predlogi (npr. samostalnik-predlog-samostalnik, glagol-predlog-samostalnik, pridevnik-predlog-samostalnik itd.), kjer po novem lahko upoštevamo tudi sklon predloga, ki v prejšnji slovnici besednih skic ni bil upoštevan zaradi prevelikega števila tako pridobljenih relacij/stolpcev v besedni skici. Nova slovnica je bila izdelana z upoštevanjem vseh struktur, registiriranih v leksikalni bazi v času izdelave, in ima tako bistveno več slovničnih relacij kot slovnica besednih skic, ki je bila uporabljena pri ročni izdelavi LBS. Skupaj je slovničnih relacij 103, število po relacijah je navedeno v tabeli 1.

direktiva	število
SEPARATEPAGE + TRINARY	36
DUAL	23
UNARY	2
CONSTRUCTION	13
CONSTRUCTION +UNARY	6
COLLOC	3
SYMMETRIC	2
brez	18
skupaj	103

Tabela 1: slovnične relacije po direktivah

Kot je razvidno iz tabele, so vse direktive s tremi elementi (*TRINARY) uporabljene v kombinaciji z izpisom na novi strani. Kombinacija direktiv CONSTRUCTION+UNARY je uporabljena v primeru, ko želimo, da nas sistem z izpisom v posebnem stolpcu "Constructions" opozori, da se neka kombinacija pogojev v korpusu pojavlja nadpovprečno pogosto (kar je sicer osnovna funkcija direktive UNARY). S pomočjo te direktive je pri luščenju mogoče tudi avtomatsko generirati opozorila, ki bi jih v klasičnih slovarjih pričakovali v t. i. slovničnih kvalifikatorjih, npr. pogosto zanikano, pogosto v 3. os. ednine itd. V postopku ALLP smo te podatke vključili v LBS v element <oznaka>, ki ima podobno vlogo kot kvalifikatorji. Pri vsaki od slovničnih relacij je naveden tudi podatek, kako se posamezna relacija prevaja v strukture, beležene v leksikalni bazi, s pomočjo katerih je mogoče identificirati neposredno povezavo med relacijo in elementom v leksikalni bazi. Primer:

*DUAL

=S_v_rodil-s/S_s-koga-česa

[2] Direktive določajo, kako program obravnava zapise v vrsticah, ki jim v slovnici besednih skic sledijo.

Struktura, s pomočjo katere luščimo kombinacije samostalnika v kateremkoli sklonu in samostalnika v rodilniku (npr. *delovanje motorja, valovanje morja*) se v leksikalni bazi pojavlja v strukturi SBZ0 sbz2, če je iztočnica jedrni samostalnik, ali v strukturi sbz0 SBZ2, če je iztočnica samostalnik v rodilniku. Ustrezen podatek o povezavi je dodan vsaki relaciji:

```
# LBS-XX ##########
# /1/ <struktura>SBZ0 sbz2</struktura>
# /2/ <struktura>sbz0 SBZ2</struktura>
####################
```

Opisana slovnica besednih skic je namenjena zgolj avtomatskemu luščenju podatkov iz korpusa, saj je za človeškega uporabnika razmeroma težko berljiva zaradi velikega števila relacij in kompleksih poimenovanj relacij.

3.4. GDEX za selekcionirano luščenje korpusnih zgledov

Vključevanje korpusnih zgledov predstavlja v LBS pomemben del človeškemu uporabniku namenjene informacije, saj se z njimi potrjujejo pomenska členitev in definicije, kolokacijske lastnosti besed, njihovo obnašanje v stavčnih vzorcih, tipične besedilne in žanrske rabe, pragmatika ipd. Zato je izbira dobrega korpusnega zgleda, ki naj bi, kot pravita Atkins in Rundell (2008: 458), ustrezal vsaj trem merilom: pristnosti in tipičnosti, informativnosti in razumljivosti, še toliko bolj pomembna. Iskanje takih zgledov pa postaja zaradi čedalje večjih korpusov in posledično velike količine podatkov vse težje in vse bolj zamudno.

Pomoč leksikografom pri iskanju dobrih zgledov predstavlja orodje GDEX (Good Dictionary EXamples; Killgarriff et al., 2008), ki zglede razvršča glede na njihovo kakovost. Ker pa so tipičnost, informativnost in razumljivost težko merljive lastnosti, aplikacija GDEX pri oceni kakovosti zgleda meri predvsem značilnosti, ki so z omenjenimi merili posredno povezane. Sem sodijo zlasti dolžina zgleda, celostavčna oblika, preprosta ali manj kompleksna skladenjska zgradba povedi, prisotnost ali odsotnost redkih besed, spletnih in elektronskih naslovov ipd.

Prva različica orodja GDEX je bila razvita za angleški jezik in uporabljena pri izbiri dodatnih zgledov za kolokacije v spletni postavitvi slovarja Macmillan English Dictionary. Pri projektu SSJ je bila za namene LBS izdelana različica za slovenščino (Kosem et al., 2011), ki je izboljšala angleško in precej olajšala delo leksikografom. Cilj, ki je bil bolj ali manj dosežen, je bil izdelati konfiguracijo, ki bi ponudila vsaj tri dobre zglede med desetimi ponujenimi za vsak kolokator v Besedni skici, pri čemer naj bi bil vsak pomensko ali skladenjsko relevanten podatek v LBS potrjen z najmanj dvema korpusnima zgledoma.

Obstoječa verzija aplikacije GDEX za slovenščino ni bila ustrezna za potrebe ALLP zaradi razlik v konceptu računalniško-leksikografskega dela. Pri običajnem postopku leksikograf s pomočjo korpusnih orodij analizira jezikovne podatke, jih selekcionira in vnese v program za izdelavo slovarjev. Pri postopku ALLP pa se podatki avtomatsko izvozijo iz korpusa neposredno v program za izdelavo slovarjev, kjer jih leksikograf pregleda, selekcionira in uredi. Ker smo z ALLP želeli občutno skrajšati postopek ročnega polnjenja posameznih elementov geselske zgradbe, hkrati pa razbremeniti tudi postopek odstranjevanja nerelevantnih ali neustreznih podatkov, ki se v slovarsko orodje prenesejo zaradi korpusnega šuma, lematizacijskih in drugih bolj ali manj predvidljivih napak, je bil naš cilj izdelati GDEX konfiguracijo, pri kateri bi bili prvi trije ponujeni zgledi že dovolj dobri za pojasnitev predhodno registriranih kolokacij.

Iz izkušenj pri procesu izdelave prvotne konfiguracije GDEX je bilo jasno, da bodo rezultati pri izboru kakovostnih zgledov pri posameznih besednih vrstah različni. Zato smo za vsako besedno vrsto, ki je zastopana v LBS, tj. za samostalnik, glagol, pridevnik in prislov, izdelali samostojno konfiguracijo, pri čemer se konfiguracije niso razlikovale v merilih, naštetih v tabeli 2, temveč v posameznih nastavitvah. Pri določanju nastavitev za posamezno besedno vrsto smo analizirali zglede, ki so bili v LBS že ročno izbrani na podlagi meril dobrih korpusnih zgledov. Na ta način smo dobili izhodiščne statistične vrednosti za klasifikatorje, na podlagi katerih smo izdelali konfiguracijo za vsako besedno vrsto.

- cela poved
- ne vsebuje pojavnico s frekvenco manj kot 3
- poved mora biti daljša od 7 pojavnic
- poved mora biti krajša od 60 pojavnic
- poved ne sme vsebovati ponovitve leme
- vsebuje elektronski ali spletni naslov
- optimalna dolžina (med X in Y pojavnic)
- vsebuje redke leme
- vsebuje pojavnice, daljše od 12 znakov
- število ločil v zgledu (brez vejic)
- število vejic v povedi
- pojavnice z velikimi začetnicami
- pojavnice z mešanimi simboli (npr. črke in številke)
- lastna imena
- zaimki
- položaj leme v povedi
- seznam prepovedanih besed na začetku povedi
- seznam prepovedanih besednih zvez na začetku povedi
- tretji kolokator
- Levenshteinova razdalja[3]

Tabela 2: Hevristika konfiguracij orodja GDEX za slovenščino za ALLP

Drugi del analize za določanje najprimernejših GDEX konfiguracij je vključeval evalvacijo zgledov predhodne konfiguracije v orodju SkE na vzorčnem izboru lem s seznama za ALLP, sledilo je prilagajanje nastavitev oz. izdelava nove različice konfiguracije ter ponovna evalvacija. Postopek smo ponavljali, dokler nismo izoblikovali optimalne končne verzije konfiguracije GDEX za postopek ALLP. Pomemben rezultat tega dela analize je oblikovanje več novih klasifikatorjev, ki jih prvotna verzija GDEX ni vključevala. Zlasti npr. oblikovanje seznama prepovedanih besed ali zvez na

[3] http://en.wikipedia.org/wiki/Levenshtein_distance

začetku povedi in upoštevanje t. i. tretjega kolokatorja. Predvsem zadnje prinaša pri izboru korpusnih zgledov v postopku ALLP dobre rezultate, saj posredno upošteva merilo koligacijske tipičnosti določene kolokacije. Npr. pri kolokaciji *klavrn + podoba* klasifikator višje točkuje zglede s statistično pomembnim tretjim kolokatorjem *kazati*. Izbrana konfiguracija pa posledično ponudi zglede, ki vsebujejo tipično širšo strukturo kolokabilne okolice: *kazati klavrno podobo česa*.

3.5. Priprava API skripte

Prilagoditev slovnice besednih skic in konfiguracije GDEX sta bila predpogoja za pripravo API skripte, ki je zahtevala tudi usklajevanje oz. posodabljanje orodja SkE. API skripta je napisana v programu Python in omogoča luščenje podatkov s povezavo na strežnik, kjer je nameščen SkE, ter določitev ukaznih parametrov, kot so:

- korpus
- lema (za več lem je potrebna datoteka s seznamom)
- slovnična relacija (za več relacij je potrebna datoteka s seznamom)
- GDEX konfiguracija
- število zgledov na kolokator
- število kolokatorjev na slovnično relacijo
- minimalna frekvenca kolokatorja
- minimalna frekvenca slovnične relacije
- minimalna jakost kolokatorja (salience)
- minimalna jakost slovnične relacije (salience).

Za izdelavo API skripte je bilo potrebno pripraviti XML predlogo, ki smo jo nato uporabili pri izvozu podatkov. Da bi bilo avtomatsko izluščene podatke mogoče uvoziti v slovarski program iLex, je bilo potrebno predlogo ustrezno poenotiti z DTD strukturo LBS. Zaradi lažjega pregledovanja izvoženih podatkov smo v DTD dodali atribute pri elementih <kolokacija> in <zgled>, in sicer identifikacijsko številko za kolokator (v oba elementa zaradi možnosti identifikacije povezave med zgledom in kolokatorjem), indeksno številko pojavnice pri elementu <zgled>, kar bi omogočilo identifikacijo zgledov v korpusu, ter zaporedno številko zgleda za vsak kolokator v GDEX-ovi razvrstitvi zgledov.

3.5.1. Določanje parametrov

Prvi test ALLP smo izvedli s privzetimi nastavitvami: 10 kolokatorjev na relacijo, 6 zgledov na kolokator, minimalna jakost relacije ali kolokatorja = 0, minimalna frekvenca kolokatorja = 0, minimalna frekvenca relacije = 25, vendar so prve evalvacije pokazale, da ni mogoče uporabiti enakih nastavitev za vse relacije in kolokatorje, saj je izpis pri nekaterih lemah pokazal veliko nerelevantnih relacij in pripadajočih kolokatorjev, pri drugih pa nekatere relevantne relacije in kolokatorji niso bili zabeleženi. Izkazalo se je tudi, da je izluščenih zgledov za končno urejanje gesla občutno preveč.

Izhodiščne nastavitve smo v nadaljevanju izboljšali tako, da smo iz besednih skic vseh lem z našega seznama pridobili statistične podatke o relacijah in kolokatorjih, nato pa za vsako relacijo (v okviru skupine lem iste besedne vrste) analizirali vrednosti, pri čemer smo iskali optimalne minimalne frekvence in jakosti relacije. Pomagali smo si tudi s podatkom o deležu pojavitev leme v določeni relaciji. Statistično analizo smo kombinirali z ročnim pregledovanjem besednih skic, saj se je pri nekaterih lemah, zlasti tistih, kjer se je relacija pojavljala redkeje, izkazalo, da relacija za luščenje ni relevantna. Dodatna korist ročnega pregledovanja besednih skic je bila identifikacija nekaterih pomanjkljivosti v slovnici besednih skic (npr. napačno opredeljena ali klasificirana relacija), ki smo jih pred izvedbo končnega postopka odpravili.

Pri določanju minimalne vrednosti frekvence in jakosti kolokatorjev smo se oprli na podatke, ki smo jih pridobili z ročnim pregledovanjem besednih skic pri posamezni besedni vrsti in pri različnih slovničnih relacijah. Pri določanju minimalnih statističnih vrednosti na kolokator smo v besedni skici upoštevali kolokatorje, ki so predstavljali še smiselne kombinacije ter uporabili njihove statistične parametre kot osnovo za določitev vrednosti.

Pregled izluščenih podatkov na podlagi izhodiščnih nastavitev je med drugim pokazal, da je nastavitev števila kolokatorjev na slovnično relacijo za končni rezultat zelo pomemben parameter. Če namreč med prvimi desetimi kolokatorji (privzeta nastavitev) ni takih, ki bi presegali minimalno frekvenco in jakost, se relacija pri luščenju ne izpiše, četudi je zelo pogosta. Zato smo minimalno število kolokatorjev na relacijo dvignili na 25, luščenje relevantnih kolokatorjev pa 'prepustili' parametroma za minimalno frekvenco in jakost kolokatorja. Število zgledov na kolokator smo znižali na 3, tudi zaradi tega, ker je evalvacija testnih izpisov pokazala, da je v veliki večini primerov med njimi vsaj en dober zgled (pogosto pa kar vsi trije).

3.6. Od izpisa do gesla

Najboljši pokazatelj učinkovitosti ALLP je čas, v katerem iz izluščenih podatkov izdelamo končno verzijo gesla v LBS. Pri tem je dovolj zgovoren podatek, da leksikograf na podlagi ročne analize korpusnih podatkov (tj. z uporabo Besednih skic) izdela v eni uri slabo četrtino gesla oz. povprečno 0,23 gesla/uro, medtem ko je mogoče na podlagi ALLP izdelati geslo v dveh urah oz. povprečno 0,5 gesla/uro. V enakem času je torej razmerje v prid na podlagi ALLP izdelanih gesel 2 : 1. Ko so leksikalni podatki izluščeni in uvoženi v slovarsko orodje, jih leksikograf pregleda in ustrezno pomensko razčleni oz. združi. Prav tako (za zdaj) ostaja v leksikografovi pristojnosti identifikacija stalnih zvez, frazeoloških enot in pragmatičnih lastnosti pomena. Osnovni vsebinski doprinos leksikografa je še vedno tudi ubesedenje pomenskih indikatorjev in izdelava definicij ter dodajanje stilnih in področnih oznak.[4]

4. Sklep

Velik delež časa pri oblikovanju gesla na podlagi izluščenih podatkov ostaja namenjen razvrščanju in selekcioniranju podatkov, na primer združevanju kolokacij pod posamezne pomene in hkratno razvrščanje ustreznih kolokacijsko povezanih korpusnih zgledov, kopiranju kolokacij in razvrščanju zgledov v primerih, ko se kolokacija pojavi pri več kot enem (pod)pomenu. Poleg

[4] Slovnične oznake, kot denimo pogosta raba tretjeosebne konstrukcije, je, kot rečeno, avtomatično izluščena s kombinacijo direktiv CONSTRUCTION+UNARY.

tega zahteva končna ureditev gesla tudi ustrezno prečrkovanje struktur, kolokacij in skladenjskih zvez. Vendar pa je končna izdelava gesel pokazala, da je vsaj del teh opravil mogoče avtomatizirati in da obstaja še veliko možnosti za izboljšanje postopka. Menimo, da je v prispevku opisani postopek ALLP koristna pridobitev za slovensko leksikografijo, zlasti če upoštevamo stanje v trenutni slovenski leksikografski praksi, ki ne obeta sodobnega slovarskega priročnika v doglednem času. Prikazana metodologija avtomatskega luščenja je jezikovno neodvisna, posamezni parametri na vseh stopnjah ALLP pa so prilagodljivi tudi za druge jezike, kar je pomembno tudi z vidika sodobne leksikografske prakse, kjer se zagovarjajo in preizkušajo metode, ki bi v čim večji meri avtomatizirale slovarsko delo (prim. Rundell in Kilgarriff, 2011).

V prihodnje načrtujemo izboljšavo postopka ALLP, pri čemer nameravamo podrobneje analizirati avtomatsko izluščene podatke in jih primerjati z ročno pridobljenimi v smislu relevantnosti in podrobnosti oz. robustnosti leksikografskega opisa. ALLP nameravamo preizkusiti tudi na pogostejših večpomenskih lemah, predvsem z namenom podrobnejšega testiranja konfiguracij GDEX, saj pri manj pogostih lemah zaradi manjše frekvence kolokatorjev in manjše izbire zgledov učinkovitost orodja dejansko ne pride do izraza.

Posvetili se bomo tudi izboljšavi postopka na ravni različnih nastavitev parametrov API skripte in na ravni izpisa. Pri izpisu izluščenih podatkov v slovarsko orodje nameravamo vključiti postopke, ki bodo dodatno skrajšali čas končnega urejanja gesla. V mislih imamo avtomatsko odstranjevanje kolokatorjev, ki ponudijo same enake zglede (gre za korpusni šum, ki neupravičeno izpostavi določeno besedo kot kolokator), in postavitev leme in/ali kolokatorja pri izpisu v slovarsko orodje v ustrezen sklon, spol in število. Raziskati nameravamo uporabo funkcije gručenja (clustering) kolokacij na podlagi podatkov iz tezavra v orodju SkE ter preizkusiti možnost uporabe funkcije povezav na večbesedne leksikalne enote (MWU links), kar bi, predvidevamo, omogočilo luščenje t. i. razširjenih kolokacij tipa: *[delovno] mesto → [prosto, novo] delovno mesto → [razpisati, objaviti] [prosto, novo] delovno mesto.* Dolgoročni načrti vključujejo pomensko razdvoumljanje s pomočjo podatkov iz sloWNeta in izdelavo geselskih predlog za ALLP na podlagi sistemske polisemije.

S prikazano zasnovo ALLP še zdaleč nismo izčrpali vseh možnosti, ki jih postopek ponuja, vsekakor pa izkušnje in prvi rezultati kažejo, da je z današnjega vidika in prihodnjih trendov v leksikografiji pomembno vlagati v razvoj temeljnih jezikovnih virov, ki so zasnovani z mislijo na avtomatsko izrabo podatkov, in jezikovnih orodij, ki znajo podatke v takšnih jezikovnih virih izkoristiti.

5. Literatura

Atkins, T. B. S., Rundell, M. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Erlandsen, J. 2004. iLex – new DWS. Third International Workshop on Dictionary Writing systems (DWS 2004). Brno, 6. – 7. september 2004. Dostopno na: http://nlp.fi.muni.cz/dws2004/pres/#15.

Fišer, D. 2009. SloWNet – slovenski semantični leksikon. V Stabej, M. (ur.): *Infrastruktura slovenščine in slovenistike (Obdobja 28)*. Ljubljana: Znanstvena založba Filozofske fakultete. 145–149.

Gantar, P., Krek, S. 2011. Slovene lexical database. V Majchraková, D., Garabík, R. (ur.): *Natural language processing, multilinguality: sixth international conference, Modra, Slovaška, 20-21. oktober 2011*. 72-80.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychly, P. 2008. GDEX: Automatically finding good dictionary examples in a corpus. V Bernal, E., DeCesaris, J. (ur.): Proceedings of the 13th Euralex International Congress. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra. 425-432.

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. 2004. The Sketch Engine. V: Williams, G., Vessier, S. (ur.): Proceedings of the 11th Euralex International Congress. Lorient: Universite de Bretagne-Sud. 105-116.

Kilgarriff, A., Rundell, M. 2002. 'Lexical profiling software and its lexicographic applications: a case study'. A. Braasch et al. (ur.) EURALEX 2002 Proceedings. Copenhagen: University of Copenhagen.

Kosem, I., Husák, M., McCarthy, D. 2011. GDEX for Slovene. V Kosem, I., Kosem K. (ur.): *Electronic Lexicography in the 21st Century: New applications for new users, Proceedings of eLex 2011, Bled, 10-12 November 2011*. Ljubljana: Trojina, zavod za uporabno slovenistiko. 151-159.

Krek, S. 2012. New Slovene sketch grammar for automatic extraction of lexical data. SKEW3, tretja mednarodna delavnica orodja Sketch Engine, Brno, Češka, 21-22. marec 2012. Dostopno na: https://trac.sketchengine.co.uk/attachment/wiki/SKEW-3/Program/Krek_SKEW-3.pdf?format=raw.

Lorentzen, H., Theilgaard, L. 2012. Online dictionaries – how do users find them and what do they do once they have? V Vatvedt Fjeld, R., Torjusen, J. M. (ur.): *Proceedings of the 15th EURALEX International Congress, Oslo, 7–11 August 2012*. Oslo: University of Oslo, Department of Linguistics and Scandinavian Studies. 654-660.

Rundell, M., Kilgarriff, A. 2011. Automating the creation of dictionaries: where will it all end? V Meunier, F., De Cock, S., Gilquin, G., Paquot, M. (ur.): *A Taste for Corpora. A tribute to Professor Sylviane Granger*. Amsterdam: John Benjamins.