AVTOMATIZACIJA LEKSIKOGRAFSKIH POSTOPKOV

Iztok KOSEM
Trojina, zavod za uporabno slovenistiko

Polona GANTAR
Inštitut za slovenski jezik Frana Ramovša ZRC SAZU

Simon KREK
Institut "Jožef Stefan", Laboratorij za umetno inteligenco
Univerza v Ljubljani, Fakulteta za družbene vede

Kosem, I. Gantar, P. in Krek, S. (2013): Avtomatizacija leksikografskih postopkov. Slovenščina 2.0, 1 (2): 139–164.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2013/2/Slo2.0_2013_2_07.pdf.

V razpravi opisujemo poskus uvajanja postopkov avtomatizacije v proces izdelave slovarjev, ki smo ga uporabili v končni fazi izdelave leksikalne baze za slovenščino (LBS). Gre za avtomatizacijo dela leksikografskega procesa, pri katerem računalnik na podlagi vnaprej določenih parametrov izbere, izloči ter neposredno v program za izdelavo slovarja ali slovarske baze prenese vrsto leksikalnogramatičnih podatkov o konkretni lemi, ki jih leksikograf nato uporabi za pomensko analizo in končno izdelavo slovarskega gesla. Optimalnost avtomatsko izluščenih podatkov smo določali s sprotnim prilagajanjem parametrov glede na različne frekvenčne skupine lem po posameznih besednih vrstah in s postopnim prilagajanjem hevristik v aplikaciji GDEX za izbor dobrih korpusnih zgledov. Prispevek zaključujemo s prikazom vključitve postopka avtomatskega luščenja leksikalnih podatkov v predlagan slovar sodobnega slovenskega jezika.

Ključne besede: avtomatsko luščenje jezikovnih podatkov, leksikalna podatkovna baza za slovenščino, predlog za izdelavo slovarja sodobnega slovenskega jezika, besedne skice, GDEX

1 PREDSTAVITEV

Zadnje desetletje smo bili na področju leksikografije priča vrsti prelomnih

dogodkov, ki so posledica velikega tehnološkega napredka in s tem povezanih uporabnikovih predstav ter pričakovanj o slovarju. Sodobna tehnologija namreč omogoča izgradnjo obsežnih korpusov, ki leksikografom zagotavljajo dostop do resnično obsežnih jezikovnih podatkovnih baz. Poleg tega prisotnost elektronskega medija in še posebno spletnega formata, ki se je v leksikografiji uveljavil kot osrednji medij za prenos slovarskih vsebin, pomeni, da lahko te dosežejo uporabnike hitreje, kot so jih kadarkoli doslej. Splet med drugim pomeni tudi odsotnost kakršnekoli prostorske omejitve, ki je (bila) značilna za knjižni medij, s tem pa tudi povsem druge možnosti obvladovanja količine in upoštevanje različnih vrst jezikovnih podatkov.

Posledično je napredek informacijsko-komunikacijskih tehnologij prinesel marsikateri izziv tudi za leksikografe: jezikovnih podatkov je s tehnološkim razvojem na voljo veliko več kot v preteklosti, hkrati pa leksikografom, upoštevajoč vse bolj zahtevne uporabnike in tempo jezikovnih sprememb, ostaja vedno manj časa za njihovo analizo.

Možnosti optimiziranja leksikografskega procesa, ki so z novimi tehnologijami postale vse bolj realne, so spodbudile razmišljanje leksikografov v smeri, ali je mogoče leksikografovo delo kakorkoli pospešiti oz. ali je mogoče znotraj celotnega leksikografskega procesa določiti tiste postopke, ki jih je mogoče avtomatizirati, ne da bi se pri tem izgubile relevantne jezikovne informacije, njihova zanesljivost in kakovost. V ta namen so bila izdelana različna orodja za analizo korpusov, med katerimi je potrebno izpostaviti zlasti Sketch Engine z aplikacijami besedne skice (ang. Word Sketches), t. i. kliksikografijo[1] (ang. Tick-box Lexicography) in modulom za izbor dobrih zgledov GDEX (ang. Good Dictionary EXamples). Čeprav je temeljni namen omenjenih aplikacij opraviti sintezo korpusnih podatkov in tako pospešiti slovarski proces, je njihova izhodiščna zasnovanost (še vedno) namenjena "ročnemu" selekcioniranju in rutinskemu prenašanju relevantnih korpusnih podatkov v

[1] Delovno različico slovenskega izraza smo uvedli pri izdelavi LBS (Gantar in dr. 2008: 39).

program za izdelavo slovarjev.

Proces avtomatizacije leksikografskih postopkov je bil iniciiran prav z razvojem orodja GDEX (Kilgarriff in dr. 2008), ki je omogočil racionalizacijo v slovarskem procesu zelo zamudnega postopka, tj. izbire dobrih korpusnih zgledov. Program je bil razvit za izdelavo elektronske različice MacMillanovega slovarja za učenje angleščine (Rundell 2002; 2007), z njim pa je bil leksikografom omogočen izbor npr. 20 zgledov, določenih na podlagi predhodnih leksikografskih smernic prepoznavanja dobrih korpusnih zgledov (Atkins, Rundell 2008). Na podlagi omenjenega programa je bil nato v okviru projekta ForBetterEnglish izdelan GDEX Demo Dictionary – poskusni avtomatsko generirani kolokacijski slovar za angleščino (Kilgarriff in dr. 2013) (Slika 1), ki je prosto dostopen na spletu.[2] Slovar temelji na prikazu kolokacij, ki so avtomatsko generirane prek slovničnih relacij, skupaj s pripadajočim korpusnim zgledom.

benefit (n)

pp_of	hindsight :	The list of mistakes is a long one , even without the benefit of hindsight .
object_of	reap :	But online the same sort of behaviour surely is n't going to reap any benefit at all .
	bring :	Whilst this has brought economic benefits , there are signs that other aspects of the city 's culture now need attention .
	maximise :	This mix of techniques has proven to increase the learning of the individual , encourages teamwork and maximises the benefits to the organisation .
	derive :	You can derive benefit from an elective procedure , but it may be better to wait a while before having it done .
	deliver :	A key benefit delivered by Revision 8 is the support of Application Packs .
a_modifier	added :	In most cases researchers will not need the added benefit of a stereo recording .
	maximum :	Rent Almost two-thirds of tenants interviewed were in receipt of maximum housing benefit .
	mutual :	Our intentions are simple : to bring together the community and the football club for the mutual benefit of all .
	potential :	There may also be some potential tax benefits depending on your individual circumstances .
	means-tested :	But there was no comparable drop in the number of households with children on means-tested benefits .
	economic :	Whilst this has brought economic benefits , there are signs that other aspects of the city 's culture now need attention .
modifies	fraud :	Help Basingstoke & Deane crack down on benefit fraud .
	entitlement :	Benefit entitlements can also be checked through the use of computer software .
	package :	In return our client is able to offer an excellent remuneration package with class leading benefits packages .
n_modifier	incapacity :	Crucially , those already receiving an incapacity benefit can participate voluntarily .
	housing :	Rent Almost two-thirds of tenants interviewed were in receipt of maximum housing benefit .
	welfare :	The taxes go in part to pay the welfare benefits of the workers that you have thrown out of work .
	retirement :	Effective dates of allocation An allocation takes effect on the day on which the person making it becomes entitled to payment of retirement benefits .
	unemployment :	The main reason is almost incontrovertible ; it is the long duration for which unemployment benefits are payable .

Slika 1: Prikaz korpusnih zgledov za kolokacije v slovničnih relacijah samostalnika *benefit* (korist) v poskusnem GDEX Demo Dictionary.

[2] GDEX Demo Dictionary: http://forbetterenglish.com/.

Drugi pristop v procesu izdelave slovarja, na katerega se sklicujemo v prispevku, je opis avtomatizacije postopkov v Rundell in Kilgarriff (2011), pri katerem je leksikograf predvsem ocenjevalec izbir, ki jih predhodno opravi računalnik. Kot poudarjata avtorja, je veliko bolj učinkovito, če leksikograf zgolj uredi podatke in izloči računalnikove napake, kot pa če opravi celotni proces analize in izbire podatkov od začetka do konca. Tak pristop ne določa na novo le leksikografovega dela, kjer je mogoče ločevati rutinske od zahtevnejših leksikografskih postopkov, ampak tudi vlogo korpusov v procesu izdelave slovarjev na sploh.

V prispevku opisujemo tudi poskus uporabe omenjenih pristopov pri izdelavi LBS (Gantar, Krek 2011). Najprej na kratko predstavimo vsebino leksikalnogramatičnih podatkov v bazi in njeno zgradbo. Nato se osredotočamo na metodo avtomatskega pridobivanja podatkov iz korpusa, izpostavimo elemente, ki so potrebni za oblikovanje algoritma za luščenje podatkov, ter opišemo vsebino dobljenega "izvoza".

Prispevek zaključujejo z razmišljanjem o možnostih nadaljnjega izboljšanja opisanega postopka in vključitve drugih postopkov avtomatizacije, ki bi jih bilo mogoče kombinirati s predstavljenim. V zvezi s tem je potrebno poudariti, da je v prispevku opisani postopek v celoti prilagojen strukturi in vrsti leksikalnogramatičnih podatkov v LBS ter lastnostim korpusa, iz katerega so bili podatki črpani. Tudi nastavitev parametrov v API skripti je vedno smiselno prilagajati robustnosti oz. podrobnosti želenih informacij, ki jih določa tip slovarja in njegov naslovnik. Osnovni namen prispevka zato ni predstavitev posameznih nastavitev parametrov za luščenje podatkov, pač pa predstavitev metodologije, ki omogoča, npr. tudi s pomočjo drugih formaliziranih na korpusu procesljivih vzorcev, kot jih denimo določa analiza korpusnih vzorcev v projektu CPA (ang. Corpus Pattern Analysis)[3] Patricka Hanksa (Hanks, Pustejovsky 2005; Hanks 2013) in razdvoumljanje glagolskih

[3] http://nlp.fi.muni.cz/projects/cpa/

pomenov s pomočjo kolokacij (ang. DVC: Disambiguation of Verbs by Collocation),[4] avtomatsko ekstrakcijo in prenos v slovarski urejevalnik.

2 LEKSIKALNA BAZA ZA SLOVENŠČINO

LBS je eden od rezultatov projekta Sporazumevanje v slovenskem jeziku,[5] katerega cilj je izdelati jezikovne vire in orodja za za računalniško obdelavo slovenščine ter sekundarne jezikovne vire, namenjene sodobnim slovarskim, slovničnim in slogovnim (pravopisnim) opisom slovenščine. LBS ima dva osnovna namena: predstavlja vir podatkov, ki jih bo mogoče izrabiti pri izdelavi sodobnih slovarskih priročnikov za slovenščino, zato ima njena zgradba in vsebina primarno slovarski značaj, ter vir podatkov za jezikovnotehnološke potrebe in izboljšavo orodij za računalniško obdelavo slovenščine.

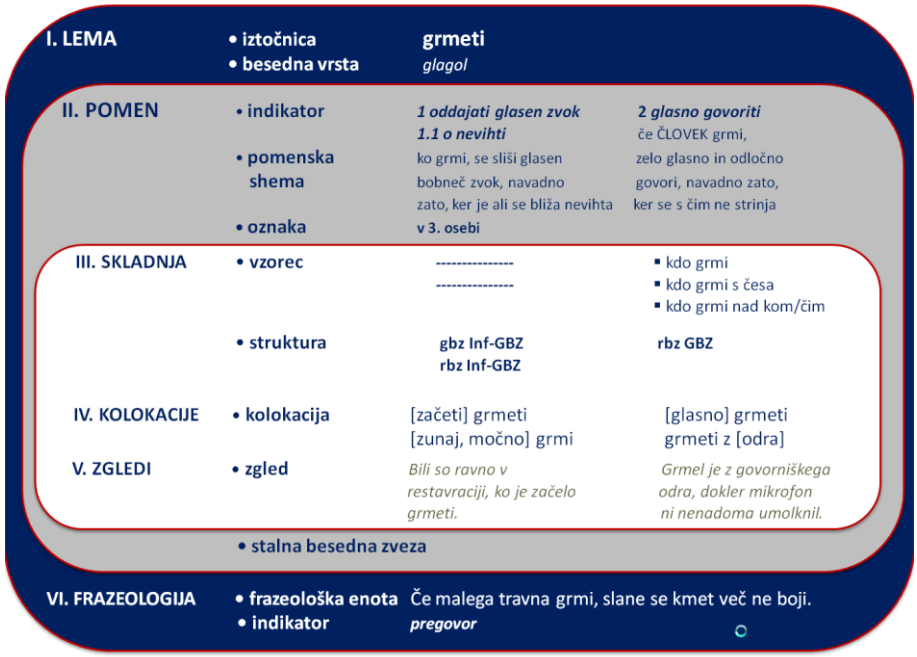
Glede na svojo dvonamenskost vsebuje LBS dva tipa podatkov: pomenske opise v t. i. pomenskih shemah, ki so izhodišče za oblikovanje razlag stavčnega tipa, kolokacijske podatke, vezane na posamezni pomen besede, ter korpusne zglede, kar je primarno namenjeno človeškemu uporabniku. Na drugi strani vključuje LBS vrsto podatkov, katerih narava in zapis sta primarno namenjena računalniški obdelavi jezika. V tem segmentu predvideva LBS zahtevnejšega jezikoslovno in/ali računalniško usposobljenega uporabnika, ki je sposoben podatke ustrezno sprocesirati oziroma jih jezikoslovno interpretirati in kombinirati. Sem sodi kodifikacija skladenjskih struktur na besednozvezni in stavčni ravni ter kodiranje vezljivostnih mest ter njihovih semantičnih tipov v argumentni zgradbi glagolskih ter nekaterih pridevniških in samostalniških pomenov.

Leksikalnogramatični podatki so v LBS organizirani v šest med seboj

[4] http://clg.wlv.ac.uk/projects/DVC/

[5] Operacijo sta financirala Evropska unija iz Evropskega socialnega sklada ter Ministrstvo za izobraževanje, znanost in šport Republike Slovenije.

povezanih nivojev (Slika 2). Hierarhična ureditev temelji na semantičnem izhodišču, kar pomeni, da so podatki na posameznem nivoju podrejeni pomenskim lastnostim besede.



Slika 2: Zgradba leksikalne baze za slovenščino.

a) Hierarhično najvišja je lema, tj. iztočnica v osnovni obliki, ki zastopa vse pripadajoče leksikalne enote, registrirane v bazi.

b) Na pomenskem nivoju so osnovni pomeni in podpomeni obravnavane besede zabeleženi s t. i. pomenskimi indikatorji, ki so primarno namenjeni oblikovanju pomenskega menija, ki služi uporabniku za hitro navigacijo po večpomenskem geslu. Drugi del pomenske informacije predstavlja pomenska shema, ki je teoretično blizu pomenskim shemam, kot jih predvideva projekt FrameNet (Fillmore, Atkins 1992; Baker in dr. 2003), in sistemu analize korpusnih vzorcev (Hanks 2013). Stalne zveze, ki zastopajo večbesedne leksikalne enote z lastnim pomenom, so za razliko od frazeoloških enot, ki so

od posameznih pomenov iztočnice neodvisne, obravnavane znotraj posameznega (pod)pomena pri samostalnikih, pridevnikih in prislovih.

c) Na skladenjskem nivoju so zabeležene skladenjske strukture, ki so formalizacija tipične besednozvezne realizacije obravnavane besede in so primarno namenjene računalniški obdelavi. V LBS predstavljajo skladenjsko ogrodje kolokacijam znotraj posameznih registriranih pomenov besede v iztočnici.

č) Na kolokacijskem nivoju so vzorci in skladenjske strukture potrjeni s tipičnimi realizacijami na leksikalni ravni. Na naslednji stopnji so kolokacije, hkrati pa tudi vse predhodne informacije (vzorci, strukture in pomenski opis v pomenski shemi, vključno s semantičnimi tipi udeležencev) potrjene še s korpusnimi zgledi.

3 IZDELAVA SLOVARSKIH GESEL S POMOČJO AVTOMATSKO IZLUŠČENIH PODATKOV

Odločitev za izvedbo postopka avtomatskega luščenja leksikalnih podatkov (ALLP) iz korpusa izhaja najprej iz narave LBS, v katero je bilo potrebno glede na njen namen in dejstvo, da za slovenščino na korpusu temelječi opisi leksike ne obstajajo, vključiti veliko količino jezikovnih podatkov, hkrati pa je bilo treba glede na omejeni čas izdelave postopek pridobivanja in analize jezikovnih podatkov čim bolj skrajšati in optimizirati. Pri izdelavi gesel smo že v začetni »ročni« fazi uporabljali orodje Sketch Engine (Kilgarriff in dr. 2004) in omenjene (leksikografske) aplikacije, zlasti besedne skice in kliksikografijo,[6] vendar pa je bilo kljub temu za kopiranje in umeščanje relevantnih kolokacij ter pripadajočih zgledov pod ustrezne skladenjske strukture v program za izdelavo slovarjev potrebnega izredno veliko časa, zaradi česar je bilo leksikografsko delo tudi veliko finančno breme. Zamudnost

[6] Prvotna različica konfiguracije GDEX, na kateri je delovala kliksikografija, še ni bila izdelana za proces ALLP, je pa vključevala prilagoditev modela za slovenščino (Kosem in dr. 2011).

bolj ali manj rutinskih postopkov je negativno vplivala tudi na leksikografovo učinkovitost in kakovost drugih opravil. Leksikograf je namreč pri pomensko kompleksnejših iztočnicah in številnih skladenjskih strukturah, ki so zahtevale prenos velike količine kolokacijskih podatkov ter zgledov v slovarski program, manj časa namenil zahtevnejšim opravilom, kot je identifikacija (pod)pomenov, večbesednih leksikalnih enot, oblikovanje pomenskih indikatorjev ter shem za posamezni pomen ali podpomen besede.

3.1 Metodologija

S postopkom ALLP smo iz korpusa Gigafida pridobili leksikalne podatke, vezane na skladenjske strukture, ki smo jih predhodno zabeležili v leksikalni bazi. V končni vsebuje LBS pribl. 450 različnih skladenjskih struktur po posameznih besednih vrstah, ki ustrezajo gramatičnim relacijam v slovnici besednih skic. Podatke smo izluščili v formatu XML in jih uvozili v iLex – program za izdelavo slovarjev (Erlandsen 2004). Kot relevantne podatke smo upoštevali skladenjske strukture, pripadajoče kolokacije in korpusne zglede. Priprava postopka avtomatizacije je vključevala naslednje:

a) izbor lem za luščenje podatkov;

b) slovnico besednih skic, ki je bila izdelana posebej za namene avtomatskega luščenja na podlagi v LBS ugotovljenih skladenjskih struktur;

c) GDEX konfiguracijo;

č) pripravo API skripte za luščenje podatkov prek funkcije besedne skice v orodju Sketch Engine in

d) določitev parametrov, kot je npr. minimalna frekvenca kolokatorja, minimalna jakost slovnične relacije in kolokacije.

3.2 Izbor lem

Zaradi lažje obvladljivosti količine podatkov pri evalvaciji na podlagi izluščenih podatkov izdelanih gesel – in posledično zaradi možnosti

postopnega izboljševanja nastavitev orodja GDEX in API skripte – smo pri naboru lem upoštevali tri merila:

a) Relevantno frekvenco leme za posamezno besedno vrsto, kar pomeni predvsem dovolj obsežno besedno skico. Prvo testiranje je namreč pokazalo, da besedne skice za manj frekventne leme (manj kot 600 pojavitev v Gigafidi) večinoma ne zagotavljajo dovolj relevantnih podatkov. Zato smo 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 dovolj »razvejano« besedno skico.

b) Potencialno enopomenskost glede na sloWNet – slovensko različico Wordneta (Fišer 2009), pri kateri smo izbrali leme z enim ali dvema sopomenskima nizoma (sinsetoma), in iztočnice z izkazanim enim ali največ dvema pomenoma v Slovarju slovenskega knjižnega jezika (SSKJ).

c) Vključenost leme v sloWNet zaradi možnosti nadaljnjih povezav in hkrati nevključenost v SSKJ zaradi možnosti identifikacije potencialnih novih izrazov in pomenov.

Končni izbor je vseboval 515 samostalnikov, 260 glagolov, 275 pridevnikov in 177 prislovov,[7] pri čemer so prevladovale leme s 1.000 in 10.000 pojavitvami v korpusu Gigafida. Manjše število lem z izkazano nižjo frekvenco smo vključili za namene dodatnega testiranja avtomatskega postopka, nekatere pogostejše leme pa z namenom, da bi preverili delovanje API skripte za vse relacije v slovnici besednih skic.

3.3 Slovnica besednih skic

Za potrebe postopka ALLP je bila izdelana nova slovnica besednih skic,[8] ki

[7] V celoti so za izbrane leme avtomatsko izluščeni podatki na voljo v datoteki XML po licenci CC na spletni strani projekta: http://www.slovenscina.eu/spletni-slovar/prenos. Del avtomatsko izluščenih podatkov je sicer vključen tudi v končno izdelanih 198 gesel v LBS.

[8] Slovnica besednih skic, ki je bila uporabljena pri ALLP, in posamezne konfiguracije za (opomba se nadaljuje na naslednji strani)

izkorišča tudi nekatere elemente, ki so bili v orodje Sketch Engine dodani v novejših različicah. Med njimi so predvsem t. i. direktive[9] (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) v novem zavihku, kar omogoča podrobnejši prikaz posamezne relacije, ki vsebuje predlog (npr. samostalnik-predlog-samostalnik, glagol-predlog-samostalnik, pridevnik-predlog-samostalnik), zaradi česar je po novem mogoče upoštevati tudi kolokacije, ki pri pridevniških in samostalniških elementih razločujejo sklone glede na predlog,[10] npr. gbz med SBZ4/SBZ6: [vpisati se] med legende, [kaditi, zrediti se] med nosečnostjo. Nova slovnica je bila izdelana z upoštevanjem vseh struktur, registiriranih v LBS 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 relacij po direktivah je navedeno v Tabeli 1.

Direktive	**Število slovničnih relacij**
SEPARATE PAGE + TRINARY	36
DUAL	23
UNARY	2
CONSTRUCTION	13

GDEX so dostopne prek orodja SketchEngine podjetja Lexical Computing Limited (LCL).

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

[10] Ta možnost v prejšnji slovnici besednih skic ni bila upoštevana, zato je bilo za ročno leksikografsko analizo pridobljenih preveč relacij oz. stolpcev v besedni skici.

CONSTRUCTION + UNARY	6
COLLOC	3
SYMMETRIC	2
brez direktive	18
skupaj	**103**

Tabela 1: Slovnične relacije po direktivah.

Kot je razvidno iz Tabele 1, so vse direktive s tremi elementi (*TRINARY) uporabljene v kombinaciji z izpisom na novi spletni 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 avtomatsko generirati tudi opozorila, ki bi jih v klasičnih slovarjih pričakovali v t. i. slovničnih kvalifikatorjih, npr. *pogosto zanikano*, *pogosto v 3. os. ednine*. 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, s pomočjo katerih je mogoče identificirati neposredno povezavo med relacijo in elementom v LBS. Primer:

```
*DUAL
=S v rodil-s/S s-koga-česa
```

Struktura, s pomočjo katere luščimo kombinacije samostalnika v kateremkoli sklonu in samostalnika v rodilniku (npr. *delovanje motorja*, *valovanje morja*) se v LBS 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 ALLP, saj je za človeškega uporabnika razmeroma težko berljiva zaradi velikega števila relacij in njihovih kompleksih poimenovanj. Posledično pa je glede na ročno analizo, ki je temeljila na manj razčlenjeni besedni skici, tudi avtomatsko izluščenih podatkov več, kot jih je ročno izločil leksikograf.

3.4 Konfiguracija GDEX

Vključevanje korpusnih zgledov je v LBS pomemben del 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 oz. 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 (Killgarriff in dr. 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, zaključena stavčna oblika, preprosta ali manj kompleksna skladenjska zgradba povedi, prisotnost ali odsotnost redkih besed, spletnih in elektronskih naslovov ipd.

Prva različica orodja GDEX za slovenščino (Kosem in dr. 2011) je bila razvita za namene ročne izdelave LBS in je precej olajšala delo leksikografom, vendar pa za ALLP ni bila ustrezna zaradi razlik v konceptu računalniško-leksikografskega dela. Pri običajnem ročnem postopku namreč leksikograf s

pomočjo korpusnih orodij pregleduje jezikovne podatke, jih selekcionira in vnese v program za izdelavo slovarjev. V tem primeru je bila naloga konfiguracije GDEX zagotoviti vsaj tri dobre zglede med desetimi ponujenimi za vsak kolokator v besedni skici.

Pri postopku avtomatizacije se podatki avtomatsko izvozijo iz korpusa neposredno v program za izdelavo slovarjev, kjer jih leksikograf pregleda in selekcionira po tem, ko je računalnik selekcijo že opravil. Osnovni namen tega postopka je torej skrajšati proces ročnega polnjenja elementov geselske zgradbe, hkrati pa razbremeniti proces odstranjevanja nerelevantnih ali napačnih podatkov. Zato je bil naš cilj izdelati konfiguracijo GDEX, pri kateri bi bili prvi trije ponujeni zgledi že dovolj dobri za pojasnitev predhodno registriranih kolokacij.

Na podlagi izkušenj pri izdelavi prvotne konfiguracije GDEX smo predvidevali, da je mogoče rezultate izboljšati z izdelavo samostojne konfiguracije za vsako besedno vrsto, ki je zastopana v LBS, tj. za samostalnik, glagol, pridevnik in prislov, pri čemer se konfiguracije niso razlikovale v merilih, naštetih v Tabeli 2, temveč v posameznih nastavitvah. Pri določanju izhodiščnih statističnih vrednosti za klasifikatorje, na podlagi katerih smo izdelali konfiguracijo za vsako besedno vrsto, smo upoštevali analizo zgledov, ki so bili v LBS že ročno izbrani na podlagi meril dobrih korpusnih zgledov.

- cela poved
- ne vsebuje pojavnice s frekvenco manj kot 3
- poved mora biti daljša od 7 pojavnic
- poved mora biti krajša od 60 pojavnic
- poved ne sme vsebovati ponovitve iskane 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[11]
- pojavnice z velikimi začetnicami
- pojavnice z mešanimi simboli (npr. črke in številke)
- lastna imena
- zaimki
- položaj iskane leme v povedi
- seznam prepovedanih besed na začetku povedi
- seznam prepovedanih besednih zvez na začetku povedi
- tretji kolokator
- Levenshteinova razdalja[12]

Seznam 1: Hevristika konfiguracij orodja GDEX za slovenščino za avtomatsko luščenje podatkov.

Drugi del analize za določanje najprimernejših GDEX konfiguracij je vključeval evalvacijo zgledov predhodne konfiguracije v orodju Sketch Engine 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 različice konfiguracije GDEX za postopek ALLP. Pomemben rezultat tega dela analize je oblikovanje več novih klasifikatorjev, ki jih prvotna različica GDEX ni vključevala. Zlasti npr. oblikovanje seznama prepovedanih besed ali zvez na začetku povedi in upoštevanje t. i. tretjega kolokatorja (tj. kolokatorja kolokacije). 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

[11] Več vejic predpostavlja skladenjsko kompleksnejši in s tem za branje zahtevnejši zgled.
[12] http://en.wikipedia.org/wiki/Levenshtein_distance

pomembnim tretjim kolokatorjem *kazati*. Izbrana konfiguracija posledično ponudi zglede, ki vsebujejo tipično širšo strukturo kolokabilne okolice: *kazati klavrno podobo česa*.

3.5 Priprava API skripte

API skripta[13] za ALLP je napisana v programu Python, za ustrezno delovanje pa je bila potrebna predhodna prilagoditev slovnice besednih skic in konfiguracije GDEX v orodju Sketch Engine 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 izpostavljenost kolokatorja (*salience*)[14]
- minimalna izpostavljenost 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

[13] Program, ki omogoča prenos izluščenih podatkov iz korpusa v slovarsko bazo, t. i. API, so razvili sodelavci podjetja LCL in prav tako deluje v orodju Sketch Engine. Parametri, ki so v programu določeni za luščenje leksikalnih podatkov, so odvisni od velikosti korpusa in želene robustnosti dobljenih podatkov, zato jih je smiselno na podlagi smernic, kot so predstavljene v prispevku, oblikovati specifično za konkretni slovarski projekt. O prilagajanju parametrov za luščenje terminoloških kandidatov s pomočjo opisane metode pri projektu Termis je mogoče prebrati v prispevku Logar Berginc in Kosem (2013).

[14] Izpostavljenost (ang. *salience*) je statistična vrednost, ki daje podatek o tem, kako močna je vez med kolokatorjem in izbrano iztočnico (Yerošina Pobirk in dr. 2009: 415).

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 evalvacije pokazale, da ni mogoče uporabiti enakih nastavitev za vse relacije in kolokatorje, saj je izpis pri nekaterih lemah pokazal veliko nerelevantnih relacij ter 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, pri katerih 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 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 Evalvacija končanih gesel

Pri preverjanju zanesljivosti postopka ALLP je potrebno v izhodišču ločiti dva tipa evalvacije. V času izdelave pričujočega prispevka je bil opravljen zgolj prvi tip evalvacije, ki ugotavlja optimalnost parametrov, določenih v API skripti, ki se nanašajo na vrednosti v besedni skici (število kolokatorjev na relacijo, število zgledov na kolokator, minimalna jakost relacije in kolokatorja, minimalna frekvenca kolokatorja in relacije). Ti parametri omogočajo luščenje relevantnih skladenjskih struktur in za posamezno skladenjsko strukturo relevantnih kolokatorjev ter zgledov. Kot že poudarjeno, so parametri odvisni od tega, čemu so izluščeni podatki namenjeni (gradnja slovarske baze, izdelava spletnega slovarja, učna gradiva ipd.), ker je s tem povezana tudi njihova količina oz. podrobnost. Ta evalvacija je bila za opisani postopek ALLP že opravljena in je opisana v predhodnem poglavju.

Drugi tip preverjanja uspešnosti ALLP postopka pa je evalvacija leksikografskega dela, ki za omenjeni postopek na geslih LBS še ni bila opravljena. V primeru take "vsebinske" evalvacije nameravamo v prihodnje primerjati identična (tj. za isto lemo) končno izdelana slovarska gesla, in sicer: gesla, ki so izdelana na podlagi avtomatsko izluščenih podatkov, in gesla, ki so izdelana na podlagi ročne analize besedne skice za določene leme. V izhodišču imata torej avtomatsko in ročno izdelano geslo isto besedno skico, le da so podatki iz nje v primeru avtomatizacije izluščeni avtomatsko, pri ročni analizi pa leksikograf sam pregleduje celotno besedno skico, ugotavlja relevantnost posamezne skladenjske strukture, odbira kolokatorje znotraj nje in izbira zglede po lastni presoji. Pri avtomatsko izluščenih podatkih se za razliko od ročno izdelanega gesla dodatno pregledovanje konkordanc ne načrtuje.

Predvidevamo, da bomo z vsebinsko evalvacijo dobili predvsem odgovor na vprašanje, ali avtomatsko izluščeni podatki leksikografu omogočajo izdelavo slovarskega gesla, ki vsebuje vse ključne leksikalnogramatične informacije. Ali torej ti podatki zadoščajo za zanesljivo pomensko razčlenitev in pomenski opis besede? Je na njihovi podlagi mogoče registrirati in pomensko opisati večbesedne leksikalne enote in frazeologijo? Ali so na podlagi avtomatsko izluščenih podatkov ustrezno oz. enakovredno kot pri ročni analizi zaznane stilne posebnosti pri posameznih (pod)pomenih, stalnih zvezah in frazeoloških enotah? Ali je pri avtomatsko izluščenih podatkih leksikografu nedostopna katera od bistvenih slovarskih informacij, zaradi katere je smiselno korpus v določeni meri analizirati ročno? In nenazadnje, ali je na ta način izdelavo gesel mogoče pospešiti.

4 INTEGRACIJA POSTOPKA AVTOMATIZACIJE V SLOVARSKI PROJEKT

Eden od osnovnih namenov izdelave LBS je bil zagotoviti podatke in preizkusiti metode, na podlagi katerih bo mogoče izdelati sodobni slovarski priročnik za slovenski jezik (Gantar 2009). Potrebno je izpostaviti, da je zadnji

obsežni enojezični slovar slovenskega jezika (SSKJ) izšel leta 1991. Ker je nastajal 20 let in ker je letnico 1991 treba razumeti tudi kot družbenopolitično prelomnico v statusu slovenskega jezika, so številne besede v njem zastarale, številne nove besede, kot tudi podatki o sodobnem besedišču na različnih ravneh (kolokacije, stalne zveze, frazeologija) pa vanj niso vključeni. Novo izdajo SSKJ je mogoče pričakovati leta 2014 (Tomažič 2013), vendar pa je glede na to, da je vključitev novih podatkov predvidena le kot nadgradnja že obstoječih, mogoče predvideti, da bo novi slovar vseboval tudi vse pomanjkljivosti svojega predhodnika. Poleg tega bo nova izdaja po vsej verjetnosti na voljo le v tiskani obliki, kar je nekoliko presenetljivo, če izhajamo iz podatkov raziskav (Rozman 2010), ki kažejo, da slovenski slovarski uporabniki, še posebno mlajše digitalno osveščene generacije, zelo redko ali skoraj nikoli ne uporabljajo tiskanih slovarjev.

V slovenskem jezikovnopriročniškem prostoru se že dolgo ustvarja potreba po popolnoma novem, celovitem opisu besedišča, ki bo odražal, kako besede in njihove pomene govorci dojemamo z vidika sodobnega sveta. Poleg tega bi moral biti tak opis redno posodabljan, če bi želel resnično zadovoljiti potrebe jezikovnih uporabnikov. Ker smo torej uporabniki slovenskega jezika z vidika podatkov o sodobnem slovenskem besedišču precej podhranjeni, je bil izdelan Predlog za izdelavo Sodobnega slovarja slovenskega jezika – SSSJ (Krek in dr. 2013), ki sledi prav tem zahtevam, in sicer s pomočjo metod, opisanih v tem prispevku. Proces izdelave predlaganega slovarja predvideva pet samostojnih in hkrati medsebojno povezanih faz, in sicer:

a) Rdečo fazo, ki predvideva avtomatsko luščenje leksikalnih podatkov, kot so skladenjske strukture, kolokacije in zgledi, iz korpusa.

b) Oranžno fazo, v kateri se predvideva, da bodo avtomatsko izluščeni podatki iz rdeče faze pregledani, neustrezni ter nerelevantni podatki pa izločeni iz

slovarske baze (in slovarja) z uporabo množičenja.[15]

c) Rumeno fazo, ki je z vidika zanesljivosti in povednosti slovarskih informacij ključna. V tej fazi se predvideva, da bodo leksikografi opravili večino analitično-sintetičnega slovarskega dela, kamor sodi pomenska členitev in pomenski opis besed, identifikacija stalnih zvez in dodajanje manjkajočih podatkov, kot so slovnične omejitve, besedilne posebnosti ipd. V tej fazi je prav tako predvidena vključitev množičenja, in sicer za ustrezno distribucijo kolokacij pod posamezne predhodno identificirane (pod)pomene ali stalne zveze.

č) V modri fazi je predvideno, da se podatkom iz rumene faze dodajo podatki o izgovoru, etimologiji, normativnih posebnostih, govornih oblikah, ki jih prispevajo specializirani jezikoslovci (terminologi, etimologi, fonetiki itd.), in podatki, ki jih je mogoče pridobiti iz drugih obstoječih ter prosto dostopnih jezikovnih baz na spletu.[16]

d) Za zeleno fazo je predvideno, da se v njej opravi končno redakcijsko delo, v katerem se preveri ustreznost vseh navedenih podatkov.

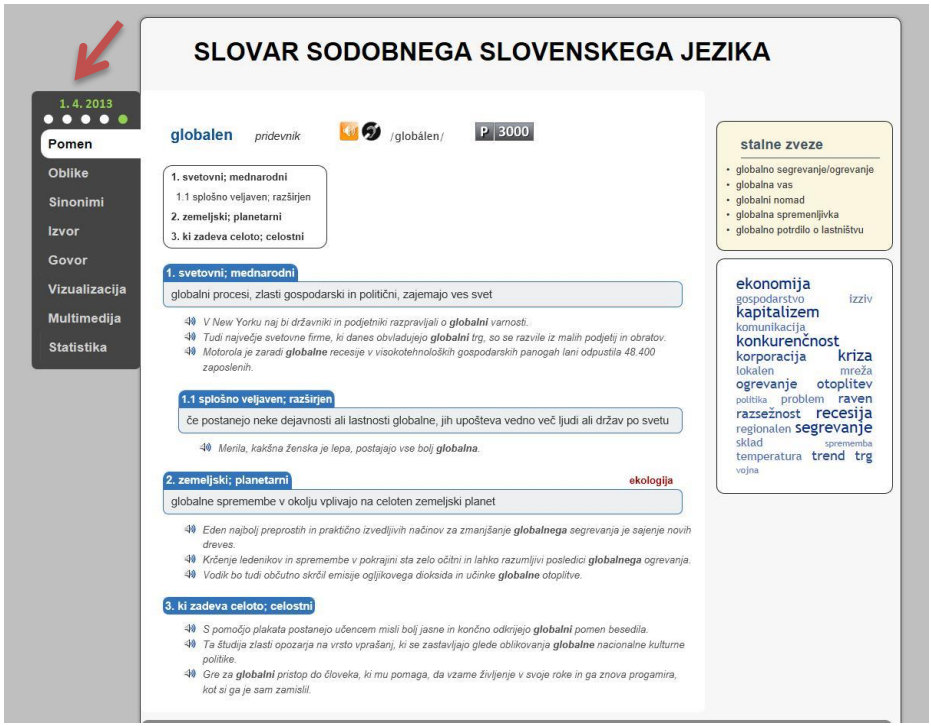
Glede na izkazano zanesljivost avtomatske metode predlagani SSSJ ne bi bil uporabnikom na voljo le ob koncu projekta, ko bodo vsa gesla izdelana do

---

[15] Avtorji prispevka menimo, da je za angleški izraz *crowdsourcing,* ki poudarja sodelovanje množice ljudi z namenom reševanja širšega problema, v slovenščini najprimernejši izraz *množičenje*, za vzporedni izraz *crowdfunding,* tj. zbiranje denarja s pomočjo prispevkov posameznikov, pa izraz *množicanje*. Oba izraza sta se, sicer še ob nekaterih drugih, po mnenju avtorjev manj uspešnih poskusih slovenjenja, kot so npr. *množično zunanje izvajanje, moč množic* in *množično izvajanje* (prim. »Terminologišče« Inštituta za slovenski jezik Frana Ramovša ZRC SAZU: http://isjfr.zrc-sazu.si/sl/terminologisce/svetovanje?keys=crowdsourcing#v), uveljavila tudi v javni komunikaciji (npr. Razvezani jezik, Mladina ipd.).

[16] V mislih imamo predvsem podatke iz Slovenskega oblikoslovnega leksikona Sloleks: http://www.slovenscina.eu/sloleks, in Slogovnega priročnika: http://slogovni.slovenscina.eu/, ki sta bila izdelana pri projektu Sporazumevanje v slovenskem jeziku, ter podatke iz Jezikovnih virov starejše slovenščine: http://nl.ijs.si/imp/, Slovenskega semantičnega leksikona sloWNet: http://nl.ijs.si/slowtool/, ter spleta na splošno.

konca, ampak že takoj po avtomatski ekstrakciji podatkov za vse predvidene leme, tj. v t. i. rdeči fazi. Gesla bi se nato dopolnjevala in posodabljala v posameznih zaporednih fazah. V zvezi s tem je predviden podatek o tem, v kateri fazi se geslo nahaja, ki uporabniku sporoča, kolikšno količino podatkov vsebuje in kako zanesljivi so. To bi lahko uporabnik razbral iz datuma zadnje posodobitve in iz barvne oznake faze, v kateri se geslo nahaja (gl. Sliko 2).



Slika 2: Podatki in informacija o stopnji dokončanega gesla v predlaganem Slovarju sodobnega slovenskega jezika.

Predlog za izdelavo SSSJ predvideva prioritetno obravnavo tematsko specifičnega in tipičnega besedišča ter terminologije, ki najbolj opazno prodira v splošni jezik – četudi le za krajše obdobje. Detektiranje tematsko specifičnega besedišča je predvideno z avtomatiziranim strojnim pregledovanjem spletnih novičarskih portalov, časopisov in drugih virov. Na

ta način je mogoče nove besede in pomene redno dodajati bodisi na podlagi strojne analize korpusa bodisi na podlagi odzivov uporabnikov.

Predlagani način izdelave slovarjev z integriranim postopkom ALLP predvideva le avtomatizacijo izhodiščnih postopkov, medtem ko leksikografska analiza, glede na to da so podatki avtomatsko selekcionirani na podlagi korpusa, še vedno temelji na (popolnem ali delnem) korpusnem pristopu. Leksikograf je odgovoren za to, da podatke pregleda, oceni, dopolni ter oblikuje končno različico slovarskega gesla. Prednosti takega pristopa k izdelavi slovarjev pridejo še posebej do izraza pri jezikih, pri katerih je potrebno izdelati slovarski opis od začetka, hkrati pa jezikovna skupnost potrebuje podatke takoj in ne z zamikom petih, desetih ali celo več let, in sicer tudi za izdelavo jezikovnotehnoloških aplikacij.

5 ZAKLJUČEK

V prispevku smo predlagali nekoliko modificiran pristop k leksikografskemu delu, ki sta ga predstavila Rundell in Kilgarriff, in sicer pristop, ki izhaja iz (1) strojnega oz. računalniškega dela, kjer ugotavljamo, da je s pomočjo računalnika mogoče v izhodišču zagotoviti zadostno količino relevantnih leksikalnogramatičnih podatkov, (2) predvideva vmesno fazo čiščenja in preurejanja podatkov, ki jo glede na jezikoslovno nezahtevnost postopkov lahko opravijo neleksikografi oz. nejezikoslovci (t. i. množičenje), ter (3) v končni fazi zahteva delo leksikografov, ki podatke oblikujejo v končni slovarski izdelek.

Stanje sodobne leksikografije v tem trenutku ni daleč od vizije, ki sta jo v svojem prispevku predstavila Rundell in Kilgarriff (2011). Avtomatizacijo postopkov je namreč mogoče implementirati v različne procese leksikografskega dela ter s tem prihraniti mnogo dragocenega časa in financ. Kljub vsemu pa naloge, ki so povezane z analiziranjem pomenskih informacij, ostajajo, vsaj za zdaj, v pristojnosti leksikografov.

Izkušnje, ki smo jih pridobili pri oblikovanju LBS, te trditve potrjujejo, obenem pa avtomatizacija postopkov izpostavlja tudi potrebo po drugačni razdelitvi človeškega dela in vključitvi novih udeležencev v celostni proces izdelave slovarja. V tej novi delitvi dela se leksikografi osredotočajo predvsem na zahtevnejše naloge, medtem ko se leksikografi z manj izkušnjami ali celo nejezikoslovci (v procesu množičenja) ukvarjajo z manj zahtevnimi nalogami in/ali rutinskimi opravili. Taka delitev dela pospeši proces izdelave slovarja in – predvidevamo – bo zlasti uspešna v dobi elektronske leksikografije, ko želijo uporabniki takojšen dostop do aktualnih leksikografskih informacij.

LITERATURA

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

Baker, C. F., Fillmore, C. J., in Cronin, B. (2003): The Structure of the FrameNet Database. *International Journal of Lexicography*, 16 (3): 281–296.

Erlandsen, J. (2004): iLex – New DWS. *Third International Workshop on Dictionary Writing systems (DWS 2004)*. Dostopno prek: http://nlp.fi.muni.cz/dws2004/pres/#15 (24. junij 2013).

Fillmore, C. J., in Atkins, S. B. T. (1992): Towards a Frame-based Organization of the Lexicon: the Semantics of RISK and its Neighbors. V A. Lehrer, E. Kittay (ur.): *Frames, Fields, and Contrasts: New Essays in Semantics and Lexical Organization:* 75–102. Hillsdale: Lawrence Erlbaum.

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

Gantar, P. (2009): Leksikalna baza: vse, kar ste vedno želeli vedeti o jeziku. *Jezik in slovstvo*, 54 (¾): 69–94.

Gantar, P., in Krek, S. (2011): Slovene Lexical Database. V D. Majchraková, R.

Garabík (ur.): *Natural Language Processing, Multilinguality: Sixth International Conference*: 72–80. Modra.

Gantar, P., Krek, S., Kosem, I., Šorli, M., Grabnar, K., Pobirk, O., Zaranšek, P., in Drstvenšek, N. (2012): *Leksikalna baza za slovenščino*. Dostopno prek: http://www.slovenscina.eu/spletni-slovar/leksikalna-baza, http://www.slovenscina.eu/spletni-slovar/prenos (20. junij 2013).

Hanks, P. (2013): *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press.

Hanks, P., in Pustejovsky, J. (2005): A Pattern Dictionary for Natural Language Processing. *Revue Française de Linguistique Appliquée*, 10 (2): 63–82.

Kilgarriff, A., Husak, M., in Jakubicek, M. (2013): Automatic Collocational Dictionaries. *Electronic Lexicography in the 21st Century: Thinking Outside the Paper, 17-19 October*. Tallinn.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., in Rychly, P. (2008): GDEX: Automatically Finding Good Dictionary Examples in a Corpus. *Proceedings of the 13th Euralex International Congress:* 425–432. Barcelona.

Kilgarriff, A., Rychly, P., Smrz, P., in Tugwell, D. (2004): *The Sketch Engine*. *Proceedings of the 11th Euralex International Congress:* 105–116. Lorient.

Kosem, I., Husák, M., in McCarthy, D. (2011): GDEX for Slovene. *Proceedings of eLex 2011:* 151–159. Ljubljana.

Krek, S., Kosem, I., in Gantar, P. (2013): *Predlog za izdelavo Slovarja sodobnega slovenskega jezika, v1.1*. Dostopno prek: http://trojina.org/slovar-predlog/datoteke/Predlog_SSSJ_v1.1.pdf (24. junij 2013).

Logar Berginc, N., in Kosem, I. (2013): TERMIS: A Corpus-driven Approach to Compiling an e-Dictionary of Terminology. *Proceedings of the eLex 2013 Conference:* 164–178. Ljubljana, Tallinn.

Tomažič, A. (2013): SKJ, d. o. o. *Pogledi*, 4 (11): 15. Dostopno prek:

http://www.pogledi.si/druzba/sskj-d-o-o (28. junij 2013).

Rozman, T. (2010): *Vloga enojezičnega razlagalnega slovarja slovenščine pri razvoju jezikovne zmožnosti: Doktorska disertacija*. Ljubljana: Filozofska fakulteta.

Rundell, M., ur. (2002): *Macmillan English Dictionary for Advanced Learners, 1st edition*. Oxford: Macmillan.

Rundell, M., ur. (2007): *Macmillan English Dictionary for Advanced Learners, 2nd edition*. Oxford: Macmillan.

Rundell, M., in Kilgarriff, A. (2011): Automating the Creation of Dictionaries: Where will it all End? V F. Meunier in dr. (ur.): *A Taste for Corpora: A tribute to Professor Sylviane Granger*: 257–281. Amsterdam: Benjamins.

Slovar slovenskega knjižnega jezika (1970–1991/spletna različica: 2000). Ljubljana: ZRC SAZU. Dostopno prek: http://bos.zrc-sazu.si/sskj.html (24. junij 2013).

Yerošina, O., Zaranšek, P., in Šuster, S. (2009): Besedne skice za slovenščino: Kritični pogled. V M. Stabej (ur.): *Infrastruktura slovenščine in slovenistike (Obdobja 28):* 413–422. Ljubljana: Znanstvena založba Filozofske fakultete.

AUTOMATIZATION OF LEXICOGRAPHIC WORK

A new approach to lexicographic work, in which the lexicographer is seen more as a validator of the choices made by computer, was recently envisaged by Rundell and Kilgarriff (2011). In this paper, we describe an experiment using such an approach during the creation of Slovene Lexical Database (Gantar, Krek, 2011). The corpus data, i.e. grammatical relations, collocations, examples, and grammatical labels, were automatically extracted from 1,18-billion-word Gigafida corpus of Slovene. The evaluation of the extracted data consisted of making a comparison between the time spent writing a manual entry and a (semi)-automatic entry, and identifying potential improvements in the extraction algorithm and in the presentation of data. An important finding was that the automatic approach was far more effective than the manual approach, without any significant loss of information. Based on our experience, we would propose a slightly revised version of the approach envisaged by Rundell and Kilgarriff in which the validation of data is left to lower-level linguists or crowd-sourcing, whereas high-level tasks such as meaning description remain the domain of lexicographers. Such an approach indeed reduces the scope of lexicographer's work, however it also results in the ability of bringing the content to the users more quickly.

Keywords: automatic extraction, Slovene Lexical Database, proposal for a dictionary of contemporary Slovene, Word Sketches, GDEX