Darja Fišer
Filozofska fakulteta Univerze v Ljubljani

Vloga jezikovnih tehnologij pri zasnovi slovarja

Prispevek predstavi pomen in možnosti jezikovnih tehnologij pri zasnovi, izdelavi in objavi načrtovanega novega slovarja slovenskega jezika s poudarkom na leksikalni semantiki. V prispevku argumentiramo, da jezikovnotehnološke vsebine sodobnega leksikografskega projekta ne smejo biti zgolj v podporni funkciji in podrejene ostalim delovnim fazam, kot je bilo v odzivih na objavljen predlog novega slovenskega slovarja večkrat namignjeno, temveč mora biti tak projekt jezikovnotehnološko že zasnovan, nato pa jezikovne tehnologije igrajo ključno vlogo tudi pri vseh nadaljnjih leksikografskih postopkih.

The role of language technologies in dictionary design

The paper discusses the importance and possibilities of language technologies in the design, production and publication of the planned new dictionary of Slovene language with an emphasis on lexical semantics. The paper argues that language technologies should not play only a support role and be inferior to other work packages as has been suggested in feedback to the published proposal for the new dictionary but that the dictionary should be based on the state-of-the-art language technologies which should then continue to play a key role in all further stages of dictionary production.

Ključne besede: jezikovne tehnologije, računalniška leksikografija, leksikalna semantika

Keywords: language technologies, computational lexicography, lexical semantics

1 Uvod

Računalniške tehnologije so se v leksikografiji začele uporabljati na začetku osemdesetih let prejšnjega stoletja, vse odtlej pa njihov pomen samo še narašča. Izdelava slovarjev postaja vse bolj računalniško usmerjena (prim. Clear 1987 in Glassman idr. 1992) in to se jasno odraža tudi v danes eni najživahnejših vej sodobne leksikografije, računalniški leksikografiji

(Boguraev in Briscoe 1989), ki se je porodila iz slovarskega projekta COBUILD English Language Dictionary (Sinclair 1987). S pomočjo informacijskih in jezikovnih tehnologij ter inženiringa znanja je slovaropisje učinkovitejše in bolj konsistentno, končni izdelek pa kvalitetnejši (Kruyt 1995). A pri tem nimamo v mislih le izboljšav pri izdaji in uporabi slovarjev; tehnologije so revolucionarno vplivale tudi na delo leksikografa, tako pri sestavi kot posodabljanju slovarjev (Raymond 1986).

Namen informacijskih in jezikovnih tehnologij ni, da bi leksikografa nadomestile ali celo izpodrinile, temveč da bi mu z avtomatizacijo ponavljajočih se in duhamornih opravil omogočile, da se posveti zahtevnejšim nalogam (Rundell 2012). Tehnološka podpora omogoča tudi sprejemanje metodološko utemeljenih in sistematičnih odločitev v vseh fazah projekta (Grefenstette 1998), to pa zagotavlja preverljive in ponovljive postopke (Sinclair 1991) ter prinaša konsistentne leksikografske odločitve skozi celotno trajanje projekta pri posameznem leksikografu, izboljšuje pa tudi ujemanje med različnimi leksikografi.

V prispevku predstavljamo možnosti in vlogo jezikovnih tehnologij pri analizi in interpretaciji korpusnih podatkov za leksikografske namene. Osnovna teza prispevka je, da jezikovne tehnologije v uspešnem leksikografskem projektu niso uporabljene le kot podpora ostalim delovnim fazam, ki so jim podrejene, temveč morajo biti sodobni leksikografski projekti jezikovnotehnološko že zasnovani, tehnologije pa morajo igrati ključno vlogo tudi pri vseh nadaljnjih korakih. Zgolj na tak način lahko namreč dosežemo objektivizacijo leksikografskega dela in posledično zadostno mero učinkovitosti in natančnosti ter zagotovimo sledljivost, reproducibilnost, povezljivost in dolgoročno ter večnamensko uporabnost rezultatov.

2 Zamujene priložnosti

Pri večini leksikografskih projektov Inštituta za slovenski jezik, ki je bil ustanovljen z namenom, da sistematično zbira, spremlja in uporablja jezikovno gradivo za izdelavo temeljnih del, med katere sodijo tudi slovarji in glosarji, so po tehnološki podpori posegli šele po zaključku projektov, pa še ta je bila le delna. Glede na okoliščine tako niti ni presenetljivo, da številnih funkcionalnosti, ki so v sodobni leksikografiji že dolgo *de facto* standard, tehnologi v izdelkih Inštituta niso mogli udejanjiti, saj tega strukture slovarskih baz preprosto niso omogočale. Večkrat izpostavljen argument pomanjkanja finančnih sredstev v tem

primeru ne zdrži, saj bi bilo z jezikovnotehnološko osveščenim pristopom in drugačnim načinom dela z istimi sredstvi mogoče doseči bistveno bolj uporabne rezultate.

Stanje ilustriramo na nekaterih strukturnih pomanjkljivostih SSKJ, ki bi jih bilo z ustrezno tehnološko podporo mogoče na enostaven način odpraviti (glej Sliko 1):

Slika 1: Primer slovarskega gesla iz SSKJ

1. Za krajšave, ki so bile v tiskanih slovarjih potrebne zaradi stiske s prostorom (npr. *ipd.*), bi pričakovali, da bodo v eletronskem slovarju zaradi lažjega razumevanja razvezane, saj prostorskih omejitev v elektronskem mediju ni.
2. S kazalkami nakazane navzkrižne povezave med gesli (npr. *gl.*) v elektronski različici slovarja niso avtomatizirane, čeprav bi to bila razmeroma rutinska programerska naloga. Tako mora uporabnik namesto klika na hiperpovezavo (npr. *avtoštopar -> avtostopar*) opraviti novo iskanje, s čimer bo slovar nedvomno izgubil marsikaterega uporabnika.
3. Implicitno izražene semantične relacije med besedami v elektronski različici slovarja niso eksplicitno izražene (npr. sinonimi *avtostopar, avtoštopar, stopar* in *štopar*).
4. Slovar ne vsebuje povezav na druge leksikografske, enciklopedične in korpusne vire, ki bi nudili dodatne informacije (npr. katera od ortografskih različic je najpogostejša).

Te pomanjkljivosti zmanjšujejo uporabno vrednost slovarja, saj so informacije, ki v slovarju so, a uporabnik do njih zaradi slabe implementacije ne more priti ali jih ne razume, izgubljene in ne upravičujejo časovne in finančne investicije. Ker temeljni slovenski slovarski projekt ne trpi zgolj zaradi pomanjkljive jezikovnotehnološke podpore, temveč se zaplete že pri informacijskotehnološki podpori, smo se za primere, po katerih se pri pripravi zasnove novega slovarja lahko zgledujemo, ozrli v tujino. Mednje sodijo angleška lekskikalna baza

DANTE[1], Splošni spletni nizozemski slovar[2], Leksikalna podatkovna zbirka za francoščino in Poljski spletni slovar[3]. Omenjenim izdelkom je skupno, da za izdelavo in vizualizacijo uporabljajo napredne informacijske in jezikovne tehnologije ter prinašajo hitro dostopne, učinkovite in zanesljive informacije, ki jih tudi redno posodabljajo.

3 Potencial jezikovnih tehnologij

V nadaljevanju predstavljamo jezikovne tehnologije za računalniško leksikografijo. Pri tem se osredotočamo na področje leksikalne semantike, saj je ta ena osrednjih tem vsakega leksikografskega projekta, hkrati pa je zaradi izmuzljivosti besednega pomena, pomenskih odtenkov in semantičnih povezav med besedami tudi ena najtežavnejših. Uporaba jezikovnih tehnologij za leksikalno semantiko ima za leksikografske projekte velik potencial, saj algoritmi zagotavljajo temeljito in sistematično obravnavo jezikovnih fenomenov v gradivu, ki je za ročni pregled preobsežno. Vendar je pri tem treba opozoriti na dvoje. Prvič, jezikovne tehnologije lahko dajejo dobre rezultate le, če jih uporabimo na kvalitetnih korpusnih podatkih, ki morajo biti dovolj obsežni za statistične metode, pa tudi dovolj reprezentativni za posploševanje identificiranih pojavov. In drugič, predstavljene tehnologije so uporabne le, če je gradivo ustrezno zbrano in pripravljeno ter integrirano v slovarsko bazo, algoritmi pa prilagojeni konkretnemu slovarskemu projektu glede na vire, ki so na voljo, potrebe, ki jih v projektu imamo, in cilje, ki jih s projektom želimo doseči. Zato morajo biti jezikovni tehnologi vpeti v celoten projekt od zasnove do objave slovarja ter v sodelovanju z leksikografi in uredniki algoritme sproti prilagajati in dopolnjevati.

Jezikovnotehnološki paket za avtomatizirano izdelavo slovarjev si je zamislil že Sinclair (1991), ki je v njem predvidel module za luščenje kolokacij, večbesednih leksemov, določanje pomena in besedilnih tipov. Podobno vizionarski je bil Grafenstette (1998), ki je napovedal, da se bo leksikografski poklic v novem tisočletju radikalno spremenil, saj bo zbiranje in gručenje primerov rabe kmalu popolnoma avtomatizirano, leksikograf pa bo odgovoren le še za prepoznavanje lastnosti primerov v isti gruči in oblikovanje slovarskih definicij zanje. Precejšen del Sinclairovih in Grafenstettejevih vizij je že uresničenih in se uporabljajo v sodobnih leksikografskih projektih. Te opisujemo v razdelku 2.1, razdelek 2.2 pa posvečamo

[1] http://www.webdante.com.
[2] http://anw.inl.nl/search.
[3] http://www.wsjp.pl.

tehnologijam, ki sicer že obstajajo in veliko obetajo, a bodo svoj polni potencial dosegle šele z nadaljnjim razvojem.

a. Zrele jezikovne tehnologije

Ne glede na vrsto leksikografskega projekta in tip slovarja, ki ga izdelujemo, glavnina leksikografskega dela temelji na korpusnih podatkih, tudi pri nalogah, ki (še) niso avtomatizirane. Zato je kvaliteta predprocesiranja korpusnih podatkov za učinkovito in uspešno delo ključnega pomena. Najprej je treba poskrbeti za natančno in robustno **segmentacijo** besedil na stavke ter **tokenizacijo** oz. prepoznavanje pojavnic in ločil v njih, s čimer določimo elemente, ki jih bomo v korpusu opazovali in primerjali. Pri stavčni segmentaciji navadno predpostavljamo, da so meje med stavki končna ločila. Do napak pri segmentiranju prihaja predvsem zaradi dvoumnosti pike, ki ni vedno uporabljena kot končno ločilo, zato je treba segmentacijska pravila napisati glede na potrebe projekta. Tudi na videz trivialna tokenizacija se pri delu z resničnimi besedili hitro zaplete, saj so ločila včasih del pojavnic (npr. *4., itd., tetra-hidro-kanabinol*), veliko besedil pa vsebuje tudi napake (*biba gre.Biba mala*). Visoka stopnja natančnosti pri tokenizaciji je pomembna, ker predstavlja temeljno fazo računalniške obdelave besedil in so od nje odvisni vsi nadaljnji postopki.

V drugem koraku pojavnicam glede na njihovo okolico pripišemo oblikoslovne lastnosti oz. jih **oblikoslovno označimo** (npr. *samostalnik ženskega spola ednine v dajalniku*), s čimer omogočimo posploševanje identificiranih jezikovnih pojavov s posameznih pojavnic na celotno besedno vrsto. Orodja za označevanje lahko temeljijo na slovničnih pravilih, statističnih metodah ali metodah strojnega učenja. Do težav prihaja pri dvoumnih besednih oblikah (npr. *brata – rodilnik/tožilnik, hotel – samostalnik/glagol*), probleme pa povzroča tudi nestandarden zapis besed (npr. *jest – jaz/jesti*).

Sledi še **lematizacija** oz. pripisovanje osnovne besedne oblike pojavnicam v korpusu (npr. *pisal – pisati*), ki je za visoko pregibne jezike, kot je slovenščina, nujna, saj omogoči opazovanje rabe neke besede v vseh njenih oblikah. Orodja za lematizacijo lahko temeljijo na metodah strojnega učenja na osnovi ročno označenega korpusa ali z uporabo pravil. Tudi pri lematizaciji težave povzročajo dvoumne besedne oblike, še trši oreh pa so t. i. neznane besede, torej besede, ki jih v učnem korpusu ni in se jih lematizator ni mogel naučiti lematizirati. Mednje sodijo odprte besedne vrste, predvsem samostalniki in lastna imena. Naprednejša orodja lemo neznanih besed skušajo uganiti, vendar je natančnost teh rezultatov nižja.

Z orodji za jezikoslovno označevanje korpusov je slovenščina razmeroma dobro opremljena, saj sta na voljo kar dva prosto dostopna označevalnika in lematizatorja, in sicer ToTaLe (Erjavec idr. 2005) in Obeliks (Grčar idr. 2012). Trenutno dosegata 90-odstotno natančnost za cele oznake in 97-odstotno za označevanje besedne vrste. Ob upoštevanju dejstva, da ima slovenščina skoraj 2000 oblikoskladenjskih oznak, je jasno, da je označevanje slovenščine zahtevna naloga, vendar je nadaljnji razvoj označevalnikov nujen, saj se za potrebe slovaropisja ne smemo sprijazniti z dejstvom, da je v korpusu vsaka deseta pojavnica označena narobe.

Naj na tem mestu izpostavimo, da ti **algoritmi za normalizacijo** niso pomembni samo za označevanje korpusnih podatkov, temveč jih vse več elektronskih slovarjev vključuje tudi v uporabniški vmesnik, to pa poenostavi iskanje v primerih, ko uporabniki kot iskalni pogoj uporabijo besede oz. besedne zveze v neosnovni obliki ali se zatipkajo (Měchura 2008). Lematizatorji in črkovalniki imajo v tem primeru še težjo nalogo kot pri korpusih, saj iskalni pogoji nimajo sobesedila, na katero bi se algoritmi lahko oprli.

Jezikovne tehnologije so lahko učinkovit pripomoček tudi pri pripravi **nabora gesel** za slovar in odločitvah, v kakšnem vrstnem redu je izbrana gesla najbolj smiselno obdelati; to je eno ključnih vprašanj, ki se kljub večstoletni slovaropisni tradiciji vedno znova zastavlja pri vsakem leksikografskem projektu. Tako je danes besedišče za vključitev v slovar mogoče motivirati po ključnosti (Kilgarriff in Rundel 2002). Poleg pogostosti rabe, za katero se je izkazalo, da ni zanesljiv indikator (de Schryver idr. 2006), je tako pomembno ugotavljati še, ali se beseda uporablja v vseh besedilnih tipih in virih ali samo v nekaterih, ali je pogosta zgolj v kratkem časovnem obdobju, ali njena raba narašča ipd. Parametre določamo za vsak projekt sproti, saj bodo odločitve neposredno odvisne od tipa slovarja, ki ga pripravljamo, in njegovih ciljnih uporabnikov.

Pri posodabljanju slovarjev imajo elektronski slovarji pomembno prednost pred knjižnimi, saj lahko z **analizo zgodovine iskanj** ugotavljamo, kaj uporabnike dejansko zanima, katera gesla iščejo in jih ne najdejo, ter izsledke upoštevamo pri pripravi nove različice (Prószéky in Balázs 2002).

Naslednji nepogrešljivi del sodobnih leksikografskih projektov je avtomatsko **luščenje kolokacij, večbesednih zvez in terminov**. Raziskovanje kolokacijskega vedenja besed je za več kot 50 jezikov omogočilo orodje Sketch Egine (Kilgarriff idr. 2010), ki s pomočjo vnaprej

pripravljenih slovničnih vzorcev iz korpusa ustvari besedne skice, to pa močno olajša in pospeši leksikografsko delo. Na podlagi besednih skic so bili avtomatsko izdelani slovarji kolokacij za 11 jezikov.[4] Za slovenščino sta slovnične vzorce za luščenje kolokacij (npr. *osvojiti medaljo*) izdelala Krek in Kilgarriff (2006).

Luščenje skladenjsko-avtomatskih podatkov iz korpusa, predvsem za namene identifikacije vezljivostnih podatkov (npr. *[kdo] ponazori [komu] [kaj] [s čim]*) ter stalnih besednih zvez (npr. *sončna očala*), je omogočeno s **skladenjskim razčlenjevalnikom**. Za slovenščino je bil naučen MSTParser (McDonald idr. 2006, Dobrovoljc idr. 2012).

Pomemben delež večbesednih zvez predstavljajo tudi **imenske entitete**, ki k vsebini besedila prispevajo več informacij, kot bi bilo moč razbrati zgolj iz njihovih posameznih elementov, zato jih je treba obravnavati kot celoto (npr. *Evropska komisija*). Prepoznavanje imenskih entitet v slovenskih besedilih deluje po metodi nadzorovanega strojnega učenja na osnovi pogojnih naključnih polj, ki je bil naučen na označenem korpusu ssj500k (Štajner idr. 2013).

Luščenje terminologije (npr. *dimeljska kila*) omogoča Sketch Engine, orodje, ki iz enojezičnih in vzporednih korpusov slovenske termine lušči s pomočjo statističnih pristopov in oblikoskladenjskih vzorcev, pa je razvila tudi Vintar (2008).

Del rutinskih leksikografskih postopkov so tudi aplikacije za avtomatsko **luščenje slovarskih zgledov** iz korpusov, ki omogočajo racionalizacijo v slovarskem procesu zelo zamudnega postopka. Ena najbolj znanih tovrstnih aplikacij je sistem GDEX (Kilgarriff idr. 2008), ki na podlagi predhodnih leksikografskih smernic prepoznavanja dobrih korpusnih zgledov leksikografom ponudi v dokončen izbor npr. 20 zgledov iz korpusa, s čimer jim skrajša in olajša izbiro. Razvit je bil za sestavo učnega slovarja Macmillan English Dictionary, medtem pa uspešno prilagojen še za slovarske projekte v drugih jezikih, tudi za slovenščino (Kosem idr. 2013).

[4] https://www.sketchengine.co.uk/documentation/wiki/Website/LanguageResourcesAndTools.

b. Jezikovne tehnologije v razvoju

Avtomatsko prepoznavanje pomena večpomenskih besed in z njim povezana **pomenska členitev slovarskih gesel** je ena najbolj zaželjenih jezikovnotehnoloških aplikacij za številna področja procesiranja jezika, od luščenja informacij do strojnega prevajanja. Od tovrstnih aplikacij pričakujemo, da bodo s primerjavo podobnosti sobesedila in metodami za nenadzorovano strojno učenje s hierarhičnim razvrščanjem v skupine (Purandare in Pedersen 2004) konkordance neke besede razvrstile v več skupin (npr. *krilo-organ, krilo-letalo, krilo-stavba, krilo-obleka*). Kljub temu da je področje v zadnjem desetletju močno napredovalo, je za potrebe leksikografskega dela zaenkrat le pogojno uporabno, saj sistemi trenutno dosegajo dobre rezultate za prepoznavanje homonimije in grobo členitev pomenov polisemnih besed (npr. *krilo-organ* in *krilo-obleka*), za nadrobno pa še ne (npr. *krilo-letalo* in *krilo-stavba*). Zaradi tega se avtomatsko prepoznavanje pomenov zaenkrat uporablja pri avtomatski gradnji leksikonov za potrebe računalniškega jezikoslovja (Lau 2009), za katerega groba pomenska členitev v največjem številu primerov zadošča, za uporabo v leksikografskih projektih pa bi jih bilo treba še izboljšati.

Pri leksikografskih projektih prav tako veliko pričakujemo od **avtomatskega iskanja semantičnih relacij** med besedami (npr. *kos-ptica*). S tem bi lahko besede v slovarju povezali v semantično mrežo ter preverjali sistematičnost razlag in drugih delov semantično povezanih gesel. Eden najuspešnejših pristopov temelji na vnaprej določenih leksikalno-sintaktičnih vzorcih med nad- in podpomenkami (Hearst 1992), ki je bil nato še nadgrajen z avtomatskim odkrivanjem vzorcev, v katerih se hiper- in hiponimi pojavljajo. Z identifikacijo implicitno izraženih semantičnih relacij v slovarju je ta slovar mogoče obogatiti z dragocenimi informacijami, ki jih že vsebuje (Nakamura in Nagao 1988), obstajajo pa tudi pristopi za luščenje sinonimov, meronimov in drugih semantično povezanih besed iz korpusov, ki jih že s pridom izkoriščajo za izdelavo različnih tipov leksikalnih baz (npr. Derwojedowa idr. 2008, Richardson idr. 1998), ki za slovenščino še niso bili preizkušeni.

V prihodnosti bosta leksikografija in terminografija veliko pridobili tudi od **avtomatskega luščenja razlag**, saj je ravno pisanje razlag ena najzahtevnejših in najbolj kompleksnih leksikografskih nalog. Luščenje razlag iz korpusov je trenutno zelo živahno raziskovalno področje in je že prineslo prve spodbudne rezultate za specializirane slovarje (Navigli in Velardi 2010), za slovenščino pa jih je na specializiranem korpusu jezikovnotehnoloških besedil preizkusila Pollak (2014).

Ker se jezik nenehno spreminja, slovarji pa posledično hitro zastarijo, računalniški jezikoslovci razvijajo tudi sisteme za **avtomatsko posodabljanje slovarja**. Z njihovo pomočjo bi lahko zagotovili učinkovito in sprotno objavo posodobljenega slovarja v rednih presledkih, tako da problema z zastarelim slovarjem, ob katerem je zraslo celo več generacij slovenskih govorcev, ne bi več bilo. Pri posodabljanju slovarja nas po eni strani zanima identifikacija nove rabe besed, za katere gesla v slovarju že obstajajo (Cook in Hirst 2011), po drugi pa identifikacija neologizmov, ki jih je v slovar treba vključiti na novo (Halskov in Jarvad 2010). Predpogoj za obe aplikaciji je seveda redno posodabljan ali, še bolje, spremljevalni korpus. Evert idr. (2004) so razvili sistem za posodabljanje slovarjev na podlagi korpusnih podatkov, ki so ga že uspešno preizkusili na nemščini in nizozemščini.

4 Zaključek

V prispevku smo predstavili potencial jezikovnih tehnologij pri zasnovi, izdelavi in objavi načrtovanega novega slovarja slovenskega jezika. Na podlagi analize stanja v slovenski leksikografiji in s predstavitvijo primerov dobrih praks v tujini smo argumentirali, da jezikovnotehnološke vsebine ne smejo biti zgolj v podporni funkciji, temveč mora biti slovarski projekt jezikovnotehnološko zasnovan in izvajan.

Avtorji novega slovarja bodo morali zagotoviti hitro in postopno objavo rezultatov takoj, ko bo posamezni delovni sklop zaključen. Zagotoviti bo treba diagnostične postopke za zagotavljanje ujemanja med različnimi projektnimi fazami in med različnimi leksikografi. Slovar bo moral biti bogat z navzkrižnimi povezavami med gesli, pa tudi s korpusom in drugimi zunanjimi viri, kot je na primer Wikipedija. Prav tako pa je že zdaj treba načrtovati redne posodobitve slovarja in predvideti njegove prilagoditve za druge potrebe, s čimer bo enkratni finančni, časovni in strokovni vložek večstransko izkoriščen.

Za učinkovitost gradnje slovarja in visoko kakovost rezultatov so ključni napredni postopki za analizo in interpretacijo vse obsežnejših korpusov, ki za slovenščino večinoma še ne obstajajo ali pa še niso dovolj natančni, zato bo v njihov razvoj nujno treba vlagati tudi v okviru projekta novega slovarja slovenskega jezika. Pri tem bo treba zagotoviti izboljšave tako v njihovi natančnosti kot priklicu, saj slaba natančnost zmanjšuje uporabno vrednost orodij, slab priklic pa ne omogoča realnega vpogleda v jezikovno rabo. Ker številne aplikacije za zanesljive rezultate potrebujejo velike količine korpusnih podatkov, bo treba nadgraditi tudi korpus, redke jezikovne pojave pa najverjetneje še vedno analizirati ročno. Nezanemarljiv

zalogaj bo tudi integracija vseh obstoječih aplikacij v enotno delovno okolje, saj vsaka zahteva drugače pripravljene vhodne podatke in računalniško okolje.

S temi izzivi se bo lahko slovarska ekipa uspešno spopadla le tako, da bodo z možnostmi in omejitvami jezikovnih tehnologij seznanjeni vsi, od vodje projekta do urednikov slovarja in leksikografov, in da bodo svoje interdisciplinarno znanje uporabili pri pripravi zasnove slovarja, gradnji slovarske baze ter razvoju leksikografskih orodij in uporabniških vmesnikov. Zasnova in izdelava slovarja brez jezikovnotehnoloških kompetenc z naknadno računalniško implementacijo bi namreč vodili v že videne neuspešne leksikografske poskuse, objava knjižne slovarske oblike na digitalnem mediju pa bi imela podoben učinek, kot da bi pred najnovejši avto vpregli konja.

Literatura

AGIRRE, E., EDMONDS, P. G., 2006. *Word sense disambiguation: Algorithms and applications*. Springer.

BOGURAEV, B., BRISCOE, T., 1989. *Computational Lexicography for Natural Language Processing*. Longman.

CLEAR, J., 1987. Computing. *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary.* HarperCollins Publishers Limited.

COOK, P., HIRST, G., 2011. Automatic identification of words with novel but infrequent senses. *Zbornik konference PACLIC*.

de SCHRYVER, G., JOFFE, D., JOFFE, P., HILLEWAERT, S., 2006. Do dictionary users really look up frequent words? On the overestimation of the value of corpus-based lexicography. *Lexikos*, 16(1).

DERWOJEDOWA, M., PIASECKI, M., SZPAKOWICZ, S., ZAWISŁAWSKA, M., BRODA, B., 2008. Words, concepts and relations in the construction of Polish WordNet. *Zbornik konference GWC*, 162–177.

DOBROVOLJC, K., KREK, S., RUPNIK, J., 2012. Skladenjski razčlenjevalnik za slovenščino. *Zbornik konference Jezikovne tehnologije*.

ERJAVEC, T., IGNAT, C., POLIQUEN, B., STEINBERGER, R., 2005. Massive multilingual corpus compilation: Acquis Communautaire and ToTaLe. *Zbornik konference LTC*, 32–36.

EVERT, S., HEID, U., SÄUBERLICH, B., DEBUS-GREGOR, E., SCHOLZE-STUBENRECHT, W., 2004. Supporting corpus-based dictionary updating. *Zborik conference EURALEX*.

FIŠER, D., 2012. Language resources and tools for semantically enhanced processing of Slovene. *Multilingual processing in Eastern and Southern EU languages: low-resourced technologies in translation.* Cambridge Scholars, 92–118.

GLASSMAN, L., GRINBERG, D., HIBARD, C., MEEHAN, J., GUARINO REID, L., VAN LEUNEN, M., 1992. Hector: Connecting words with definitions. *UW Centre for the New OED and Text Research*, 37–74.

GRANGER, S., PAQUOT, M. (ur.), 2012. *Electronic lexicography*. Oxford University Press.

GRČAR, M., KREK, S., DOBROVOLJC, K., 2012. Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. *Zbornik konference Jezikovne tehnologije*.

GREFENSTETTE, G., 1998. The Future of Linguistics and Lexicographers: Will there be Lexicographers in the year 3000? *Zbornik konference EURALEX*.

HALSKOV, J., JARVAD, P., 2010. Automated extraction of neologisms for lexicography. *Zbornik konference ELex*.

HEARST, M. A., 1992. Automatic acquisition of hyponyms from large text corpora. *Zbornik konference ACL*.

HEID, U., WORSCH, W., EVERT, S., DOCHERTY, V., WERMKE, M., 2000. Computational linguistic tools for semi-automatic corpus-based updating of dictionaries. *Zbornik konference EURALEX*, 183–195.

KILGARRIFF, A., HUSÁK, M., McADAM, K., RUNDELL, M., RYCHLÝ, P., 2008. GDEX: Automatically finding good dictionary examples in a corpus. *Zbornik konference EURALEX*, 425–432.

KILGARRIFF, A., KOVÁŘ, V., KREK, S., SRDANOVIĆ, I., TIBERIUS, C., 2010. A quantitative evaluation of word sketches. *Zbornik konference EURALEX*.

KILGARRIFF, A., RUNDELL, M., 2002. Lexical Profiling Software and its lexicographic applications: a case study. *Zbornik konference EURALEX*.

KOSEM, I., HUSAK, M., MCCARTHY, D., 2011. GDEX for Slovene. *Zbornik konference eLex*, 151–159.

KREK, S., KILGARRIFF, A., 2006. Slovene word sketches. *Zbornik konference Jezikovne tehnologije*.

KRUYT, J. G., 1995. Technologies in Computerized Lexicography. *Lexikos 5(1).*

LAU, J. H., COOK, P., MCCARTHY, D., NEWMAN, D., BALDWIN, T. (2012). Word sense induction for novel sense detection. *Zbornik conference EACL*, 591–601.

LIN, D., 1998. Extracting collocations from text corpora. *Zbornik delavnice Computational terminology*.

MAGAY, T., ZIGANY, J., 1988. BudaLEX'88. *Zbornik konference EURALEX*.

MCDONALD, R., LERMAN, K., PEREIRA, F., 2006. Multilingual Dependency Parsing with a Two-Stage Discriminative Parser. *Zbornik konference CoNLL*.

MĚCHURA, M. B., 2008. Giving them what they want: search strategies for electronic dictionaries. *Zbornik konference EURALEX*.

NAKRAMURA, J., NAGAO, M., 1988. Extraction of semantic information from an ordinary English dictionary and its evaluation. *Zbornik konference ACL*.

NAVIGLI, R., VELARDI, P., 2010: Learning word-class lattices for definition and hypernym extraction. *Zbornik konference ACL*.

POLLAK, S., 2014. *Polavtomatsko modeliranje področnega znanja iz večjezičnih korpusov: doktorska disertacija*. Ljubljana.

PRÓSZÉKY, G., BALÁZS, K., 2002. Development of a Context-Sensitive Electronic Dictionary. *Zbornik konference EURALEX*.

PURANDARE, A., PEDERSEN, T., 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. *Proceedings of Conference on Computational Natural Language Learning*.

RAYMOND, D. R., WARBURTON, Y., 1986. Computerization of lexicographical activity on the New Oxford English Dictionary. *UW Centre for the New OED and Text Research.*

RICHARDSON, S. D., DOLAN, W. B., VANDERWENDE, L., 1998. MindNet: acquiring and structuring semantic information from text. *Proceedings of 17th international conference on Computational linguistics*, 1098–1102.

RUNDELL, M., 2012. The road to automated lexicography: an editor's viewpoint. *Electronic Lexicography*, 15–30.

SINCLAIR, J. M., 1987. *Looking Up: an account of the COBUILD project in lexical computing*. Collins.

SINCLAIR, J. M., 1991. *Corpus, concordance, collocation*. Vol. 1. Oxford: Oxford University Press.

ŠTAJNER, T., ERJAVEC, T., KREK, S., 2012. Razpoznavanje imenskih entitet v slovenskem besedilu. Zbornik 8. *konference Jezikovne tehnologije*.

VINTAR, Š., 2008. *Terminologija: terminološka veda in računalniško podprta terminografija.* Znanstvena založba Filozofske fakultete.