LUŠČENJE DEFINICIJSKIH KANDIDATOV IZ SPECIALIZIRANIH KORPUSOV

Senja POLLAK
Institut "Jožef Stefan", Odsek za tehnologije znanja

Pollak, S. (2014): Luščenje definicijskih kandidatov iz specializiranih korpusov. Slovenščina 2.0, 1 (2): 1–40.

URL: http://www.trojina.org/slovenscina2.0/arhiv/2014/1/Slo2.0_1_02.pdf.

Predstavljamo metodo za luščenje definicij iz specializiranih korpusov. Metoda je bila razvita za slovenščino in angleščino, sestavljajo pa jo trije pristopi: v prvem definicije luščimo z leksikoskladenjskimi vzorci, drugi uporablja avtomatsko izluščeno terminologijo, tretji pa lušči stavke, v katerih se nahajata pojem in njegova nadpomenka iz semantičnega leksikona wordnet. Metodologijo smo preizkusili na primeru področja jezikovnih tehnologij. Za namene modeliranja izbranega področja smo zgradili primerljivi slovensko-angleški *Korpus jezikovnih tehnologij,* izluščene kandidate pa smo uporabili pri gradnji *Glosarja jezikovnih tehnologij*. Celotno metodologijo smo strnili v prosto dostopen delotok, implementiran v spletnem okolju za gradnjo delotokov Clowdflows. V delotok lahko uporabnik prek spleta naloži korpus v različnih formatih, ga jezikoslovno označi, izlušči terminologijo in kandidate za definicije ter rezultate vizualizira ali shrani.

Ključne besede: luščenje definicij, spletni delotoki, jezikovne tehnologije, procesiranje naravnega jezika, luščenje znanja iz korpusov, avtomatizacija terminografskih postopkov

1 UVOD

Človeško znanje je dostopno v strokovnih besedilih, terminoloških slovarjih in enciklopedijah, v zadnjem času pa tudi v računalniku razumljivih predstavitvah področnega znanja, kot so taksonomije in ontologije. Ker je ročno modeliranje področnega znanja časovno in finančno zahtevno, so raziskovalci s področja jezikovnih tehnologij začeli razvijati (pol)avtomatske metode in orodja za luščenje strokovnega znanja iz nestrukturiranih besedil. Med njihove naloge

prištevamo na primer luščenje terminologije, definicij ali semantičnih relacij, kot tudi (pol)avtomatske pristope h gradnji taksonomij in ontologij. Luščenji terminologije in definicij sta pomembna koraka modeliranja strokovnega znanja in lahko služita različnim ciljem. Uporabljamo ju lahko kot korak pri izgradnji ontologij ali drugih uporabnih virov za nadaljnje računalniške aplikacije ali pa samostojno, za jezikoslovne, prevajalske ali terminografske namene. Do sedaj razvite metode in orodja za luščenje definicij so večinoma prilagojena za posamezne jezike, a le redko za manj razširjene jezike, kot je slovenščina.

Izhajamo iz hipoteze, da je mogoče – tudi kadar določena znanstvena veja ne razpolaga s strukturiranimi specializiranimi viri, kot so terminološki slovarji ali tezavri – s pomočjo računalniških metod samodejno izluščiti del področnega znanja iz nestrukturiranih specializiranih besedil. Naš cilj je izluščiti znanje v obliki strokovnega izrazja in kandidatov za definicije. Kot obravnavano področje smo si izbrali področje jezikovnih tehnologij.

Naš glavni doprinos je izdelava metodologije za luščenje definicijskih kandidatov iz slovenskih nestrukturiranih besedil, saj je to za razliko od marsikaterih drugih jezikov še neraziskano področje. Metodologijo pa se zlahka prilagodi za druge jezike, kar ilustriramo na primeru angleščine. Predlagano metodologijo implementiramo v obliki spletno dostopnega delotoka[1] (angl. *workflow*). V našem primeru uporabimo metodologijo v terminografske namene, natančneje pri gradnji nastajajočega *Glosarja jezikovnih tehnologij*.[2]

Članek obsega naslednja poglavja. V drugem poglavju na kratko predstavimo filozofske in leksikografske poglede na definicije. V tretjem poglavju se posvetimo področju avtomatskega luščenja znanja iz besedil, pri čemer je glavni

[1] http://www.clowdflows.org/workflow/1380/

[2] http://kt.ijs.si/senja_pollak/jt_glosar/
Glosar razumemo kot najbolj preprosto eno- ali večjezično zbirko izrazov določenega področja, skupaj z njihovih definicijami oz. razlagami. Služi lahko kot osnutek terminološkega slovarja, vendar bi bilo glosarju potrebno dodati še vrsto drugih podatkov o terminih, termine povezati med seboj, vnose pregledati ter poskrbeti za širši konsenz stroke.

poudarek na luščenju definicij. V četrtem poglavju predlagamo metodologijo za luščenje definicij iz slovenskih (in angleških) besedil. Metodologijo preizkusimo na študijskem primeru področja jezikovnih tehnologij, za potrebe katerega smo zgradili tudi primerljivi slovensko-angleški korpus (glej peto poglavje). V šestem poglavju predstavimo eksperimente in rezultate, medtem ko je sedmo poglavje namenjeno razlagi implementacije metodologije v obliki spletnega delotoka. Prispevek zaključimo s pregledom glavnih doprinosov, ugotovitvami ter smernicami za nadaljnje delo.

2 O DEFINICIJAH

V tem poglavju na kratko obravnavamo definicije v filozofski ter leksikografski literaturi, razpravljamo pa tudi o principih tvorjenja dobrih leksikografskih definicij. Številni pomembni filozofi – od antične filozofije s Platonom in Aristotelom pa do predstavnikov vseh pomembnejših smeri zahodnih filozofskih tradicij, kot so na primer Blaise Pascal, Benedict de Spinoza, John Locke, Gottfried Wilhelm Leibniz, George Berkeley, Immanuel Kant, John Stuart Mill in Heinrich Rickert – so se v svojih razmišljanjih posvečali definicijam (glej npr. Rey 2000). Vsaj kratek uvod v obravnavo definicij v filozofiji oz. natančneje logiki je na mestu, saj se velik del kasnejše leksikografske literature nanaša nanjo. Čeprav se nimamo namena ukvarjati s samo teorijo definicij, se je tudi za potrebe avtomatizacije iskanja definicijskih kontekstov iz korpusov pomembno zavedati, da definicija definicije ni nekaj univerzalnega in samoumevnega ter da je definicijo sámo smiselno definirati glede na končno aplikacijo.

2.1 Vrste definicij glede na njihov namen

V filozofiji se z definicijami ukvarja predvsem logika, avtorji pa glede na namen ločijo več vrst definicij (npr. Copi, Cohen 2009; Parry, Hacker, 1991; Hurley 2012).

Leksikalne definicije (angl. *lexical definitions*) se uporabljajo v slovarjih in

razlagajo že uveljavljeni pomen definienduma (tj. definirani pojem). To so tudi definicije, s katerimi se bomo v večji meri ukvarjali v nadaljevanju članka. Te so resnične ali neresnične, saj točno opisujejo konvencionalno rabo besede ali pa ne.

Stipulativne definicije (angl. *stipulative definitions*) so tiste, v katerih je definiendum nov ali obstoječ izraz, ki se mu pripiše poljuben pomen, ne glede na njegov morebitni že obstoječi dejanski pomen. Za te definicije ne moremo trditi, da so resnične ali neresnične.

Izostritvene definicije (angl. *precising definitions*) uporabljamo zato, da natančneje opredelimo pomen nekega izraza, vendar za razliko od stipulativnih definicij pri teh ne gre za nove, temveč za obstoječe izraze, prav tako njihov obstoječi konvencionalni pomen le zožijo, izostrijo, saj pojem podrobneje definirajo, vendar z že uveljavljenim pomenom niso v kontradikciji.

Teoretične definicije (angl. *theoretical definitions*) so razumljivi strnjeni povzetki določene teorije. *Prepričevalne definicije* (angl. *persuasive definitions*) se uporabljajo predvsem v politični argumentaciji z namenom vplivanja na obnašanje drugih.

2.2 Vrste definicij glede na način definiranja

V nadaljevanju se ukvarjamo z leksikalnimi oz. slovarskimi definicijami, ki jih kategoriziramo glede na različne načine definiranja, tj. glede na uporabljene definicijske strategije. Pojem, ki ga definiramo, se imenuje *definiendum*, del, ki definira njegov pomen, je *definiens,* oba dela pa sta lahko povezana z *zglobom* (angl. *hinge*). Glavna razlika glede načina definiranja definienduma je že pri Aristotelu postavljena med *intenzionalnimi definicijami* (angl. *intensional definitions*) in *ekstenzionalnimi definicijami* (angl. *extensional definitions*). Prve definirajo tako, da se osredotočajo na lastnosti (bistvena določila), ki so značilne za razred, ki ga definiendum opisuje (ne pa za entitete ostalih razredov), druge pa se osredotočajo na ekstenzijo oz. obseg definienduma, kar pomeni da navajajo vse možne oz. najbolj tipične realizacije definiranega pojma

(gre torej za naštevanje vseh ali tipičnih pripadajočih elementov razreda) (Copi, Cohen 2009; Svensen 1993; Zgusta 1971; Geeraerts 2003). V slovenščini so različne tipe definicij obravnavali npr. Vidovič Muha (2000 v Gantar, Krek 2009), Žagar Karer (2011), Krek (2004), Kosem (2006) ter Gantar in Krek (2009), ki glavno razliko postavljata predvsem med klasično analitično slovarsko definicijo ter celostavčno definicijo, ki se je uveljavila s slovarji Collins COBUILD (Sinclair 1987). V nadaljevanju obravnavamo različne podtipe intenzionalnih in ekstenzionalnih definicij, kategorizacija na posamezne podtipe pa ni pri vseh avtorjih enaka (glej npr. Borsodi 1967; Robinson 1972; Jackson 2002; Westerhout 2010; Kosem 2006; Geeraerts 2003).

2.2.1 INTENZIONALNE DEFINICIJE

Najbolj tipične – in po mnenju nekaterih avtorjev najbolj prestižne – so torej leksikografske definicije z obliko *genus et differentiae*. V njih je definiendum definiran z nadpomenko oz. najbližjim rodom (*genus proximum*) in vrstnimi razlikami (*differentiae specificae*) oz. vsaj eno bistveno značilnostjo, ki definiendum (oz. razred definienduma) ločuje od ostalih pripadnikov rodu (npr. Svensen 1993; Béjoint 2000). Ker se pri definicijah z obliko *genus et differentiae* pomen definienduma analizira, se imenujejo tudi *analitične definicije*. Primer analitične definicije, ki jo podata Uršič in Markič (1997) je:

i. Krog je ravninski lik, pri katerem je vsaka točka enako oddaljena od ene, poljubno izbrane točke v ravnini.

V zgornjem primeru je definiran pojem 'krog', ki je torej *definiendum,* del, ki ga definira (*definiens*), pa sestavljata *genus proximum* 'ravninski lik' ter *differentia specifica* 'pri katerem je vsaka točka /.../ v ravnini', ki krog loči od drugih ravninskih likov. Poseben podtip analitičnih definicij so *razvrstitvene definicije* (angl. *classificatory definitions*), ki podajajo le nadpomenko, ne pa vrstnih razlik.

Med intenzionalne definicijske strategije uvrščamo tudi definiranje s *parafrazo* ali s *sinonimi* (sintetične definicije) oz. širše razumljeno *relacijske definicije*, ki

pojme definirajo v odnosu do drugih pojmov, na primer z njihovimi antonimi.

V *funkcijskih definicijah* (angl. *functional definitions*) je definiendum definiran s svojo rabo, namenom oz. funkcijo. Še en podtip je definiranje s pomočjo *tipičnih lastnosti (*angl. *typifying definitions*). Slednja dva tipa sta zelo pogosta v celostavčnih definicijah (cf. Sinclair 1987).

Pogosto se različne oblike definicij med seboj kombinirajo in predstavljajo hibridne tipe. Svensen (1993: 131) in Zgusta (1971: 260) tako omenjata, da je pogosto, da ima na primer analitična definicija z obliko genus-differentiae ali pa parafraza zraven podane tudi sinonime. Funkcijske definicije ter definicije s tipičnimi lastnostmi se običajno uporabljajo v kombinaciji z že omenjenimi analitičnimi definicijami genus-differentiae, poleg tega pa lahko na primer funkcijske definicije razumemo kot podvrsto definiranja s tipičnimi lastnostmi. Za razliko od avtorjev, ki priznavajo le analitične definicije, smo mnenja, da je vse odvisno od same narave aplikacije, pri avtomatskem luščenju pa je tudi zaradi omejene količine besedil določene stroke (predvsem pri manj razširjenih jezikih) na mestu sprejeti širši pogled na definicije.

2.2.2 EKSTENZIONALNE DEFINICIJE

Druga strategija za definiranje pojmov je z njeno *ekstenzijo* oz. *obsegom*. Za razliko od intenzionalnih definicij, ki se osredotočajo na bistvene lastnosti, s katerimi je pojem definiran, ekstenzija zajema množico stvari, na katere se pojem nanaša. Naštejemo lahko vse stvari, ki jih pojem zajema, ali pa le najbolj reprezentativne. Tudi pri *ekstenzionalnih definicijah* je v literaturi omenjenih več podtipov (cf. Parry, Hacker 1991; Copi, Cohen 2009; Zgusta 1971; Svensen 1993; Westerhout 2010). Najbolj razširjen tip so *navedbene definicije* (angl. *citational definitions*), ki so tudi to, na kar mislimo, če ne specificiramo podtipa ekstenzionalnih definicij. Pri teh definicijah definirani pojem ni zaznavno prisoten, temveč se nanj nanašamo z besedami, tako da naštejemo predstavnike opisanega razreda (npr. za razlago pojma *germanski jeziki* naštejemo jezike, ki spadajo v to skupino). Za razliko od navedbenih definicij se pri *ostenzivnih*

definicijah uporabljajo zunajjezikovne strategije, kot je npr. kazanje na elemente v prostoru.

Poleg te glavne razdelitve ekstenzionalnih definicij na *navedbene* in *ostenzivne* pa zgoraj omenjena literatura navaja še nekaj tipov ekstenzionalnih definicij, ki se po navadi (a ne izključno) nanašajo na ekstenzionalne navedbene definicije. *Naštevalne definicije* (angl. *enumerative definitions*) so poseben podtip ekstenzionalnih definicij, v katerih naštejemo vse predstavnike definiranega razreda. V *definiciji s paradigmatskim primerom* (angl. *definition by paradigm example*), ki je sicer lahko navedbena ali ostenzivna, pojem definiramo z enim reprezentativnim primerom namesto naštevanja vseh ali tipičnih predstavnikov razreda. Definicije lahko tvorimo tudi z *definiranjem sestavnih delov pojma* (angl. *partitive concept definition*), npr. *Benelux* tvorijo *Belgija, Nizozemska in Luksemburg*.

2.3 Principi tvorjenja dobrih leksikografskih definicij

Aristotelska logika določa formalne pogoje za tvorjenje definicije, ki je sestavljena iz definienduma in definiensa. Definicija mora biti relevantna (definiendum in definiens morata biti pomensko povezana), koordinirana (definiendum in definiens morata imeti enak obseg), ne sme biti krožna ali nikalna, mora pa biti jasna, enoznačna in natančna (Uršič, Markič 1997: 38–39).

V leksikografiji za razliko od formalnih definicij ni treba, da je naveden ravno najbližji rod, ampak lahko bolj na splošno govorimo o definicijah z nadpomenko in razlikovalnimi značilnostmi. Vendar pa genus ne sme biti ne presplošen ne prespecifičen (Ayto 1983; Kosem 2006), osredotočiti na pa se je treba na bistvene značilnosti. Zgusta (1971: 252–253) opiše razliko med pomenom definicije v logiki in v leksikografiji tako:

> medtem ko mora logična definicija nedvoumno identificirati predmet (*definiendum*) tako, da ga postavi v očiten kontrast napram vsemu drugemu, kar se da definirati, ter ga pozitivno in nedvoumno kategorizirati kot del

> najbližjega razreda, pa leksikografska definicija našteje samo najpomembnejše semantične značilnosti definirane leksikalne enote, ki zadostujejo za njeno razlikovanje od ostalih enot.

V leksikografiji velja, da mora biti definiendum definiran z izrazi, ki so splošnejši od njega (prim. jasna definicija), izogibati pa se je treba krožnosti v definicijah, in kolikor je mogoče, tudi v zbirkah. Glede sloga se morajo definicije ogibati dvoumnega in metaforičnega izražanja, atipičnih in marginalnih pomenov, in če je le možno, uporabljati trdilno obliko (seveda so vse to priporočila in je nekatere pojme zaradi njihove narave smiselno nikalno definirati, na primer pridevnik *plešast*) (prim. Jackson 2002; Zgusta 1971; Béjoint 2000; Svensen 1993; Atkins, Rundell 2008: 412). Glede sloga Atkins in Rundell (ibid.) ter Gantar in Krek (2009) zagovarjajo, da morata ubeseditev in zgradba definicij čim bolj ustrezati naravnemu diskurzu, s čimer merijo predvsem na celostavčne definicije. Seveda pa mora biti definicija prilagojena potrebam, sposobnostim in predpostavljenemu tehničnemu znanju uporabnika (ibid.), kar je tudi načelo, zaradi katerega so bistvene razlike med splošnimi slovarji, pedagoškimi slovarji in terminološkimi slovarji. Terminografija ima kar nekaj specifičnih lastnosti, saj vsebujejo terminološki slovarji bolj podrobne (specifične) definicije, ker je med termini potrebno bolj natančno ločevanje kot v leksikografiji (Svensen 1993: 3, 22, 122–123; Zgusta 1971: 251–255).

Sodobna leksikografija in terminografija omogočata avtomatizacijo postopkov, pri luščenju podatkov iz korpusov pa je tako glede na končno aplikacijo mogoče določiti, kateri kriteriji so najpomembnejši pri definiciji definicije.

3 AVTOMATSKO LUŠČENJE ZNANJA IZ KORPUSOV

V tem poglavju se odmaknemo od filozofskih in leksikografskih pogledov ter se posvetimo avtomatskim pristopom modeliranja področnega znanja iz korpusov. Luščenje terminov kot osnovnih nosilcev znanja v specializiranih korpusih je že relativno dobro znano področje računalniškega jezikoslovja. Samodejne metode so bile razvite za različne jezike, npr. za angleščino Sclano

in Velardi (2007), Ahmad in dr. (2007), Frantzi in Ananiadou (1999), Kozakov in dr. (2004) ter Vintar (2010) za slovenščino. Za dvojezično luščenje terminologije pa so na voljo komercialna (npr. SDL MultiTerm)[3] in nekomercialna orodja (npr. Lefever in dr. 2009; Macken in dr. 2013; Vintar 2010).

V specializiranih besedilih se poleg terminov skrivajo še drugi dragoceni deli znanja, med njimi tudi definicije, katerih luščenje predstavlja osrednjo temo pričujočega prispevka. Metode luščenja definicij so bile razvite za več jezikov, kot so angleščina (Navigli, Velardi 2010; Borg in dr. 2010), nizozemščina (Westerhout 2010), francoščina (Malaisé in dr. 2004), nemščina (Fahmi, Bouma 2006; Storrer, Wellinghoff 2006), portugalščina (Del Gaudio, Branco 2007), romunščina (Iftene in dr. 2007), poljščina (Degórski in dr. 2008), pa tudi za druge slovanske jezike (Przepiórkowski in dr. 2007). Luščenje informacij iz slovanskih jezikov je težka naloga, saj so to morfološko bogati jeziki z relativno prostim besednim redom (Przepiórkowski 2007). Za slovenščino smo začeli razvijati metodologijo v Fišer in dr. (2010) in Pollak in dr. (2012a). Poleg luščenja definicij je pomembno področje modeliranja področnega znanja tudi luščenje semantičnih relacij, ne le nadpomenk in podpomenk, temveč tudi sinonimov, antonimov, meronimov ali vzročnih relacij (Meyer 2001; L'Homme, Marchman 2006).

Dosedanji pristopi k samodejnemu luščenju definicij in semantičnih relacij iz specializiranih korpusov ali s spleta se v grobem delijo na dve veji: prva temelji na (v glavnem ročno zgrajenih) pravilih oz. vzorcih, druga na strojnem učenju, pojavljajo pa se tudi kombinacije obeh pristopov.

Na pravilih temelječi pristopi skušajo do definicij priti predvsem prek njihovih skladenjskih in leksikalnih značilnosti (vzorcev). Takšno metodo je uporabil že Hearst (1992), a tudi v novejših raziskavah se uporabljajo različice metode z vzorci (npr. Muresan, Klavans 2002; Walter, Pinkal 2006; Storrer, Wellinghoff

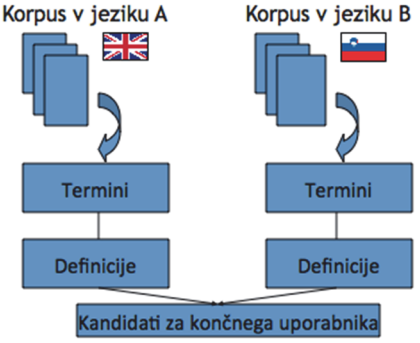
3 http://www.translationzone.com/products/sdl-multiterm/desktop/

2006). Tudi med pristopi, uporabljenimi v naši raziskavi, apliciramo metodo s pravili.

Drugi sklop raziskav se poslužuje metod strojnega učenja, pri čemer je odkrivanje definicij mogoče razumeti kot problem razvrščanja; algoritem se skuša iz učnega korpusa definicij, v nekaterih primerih pa tudi iz negativnih primerov nedefinicij, naučiti pravil za razlikovanje med pravimi in nepravimi definicijami. Z običajnimi klasifikacijskimi algoritmi, kot so naivni Bayes, odločitvena drevesa in metoda podpornih vektorjev, je različnim avtorjem uspelo razlikovati med dobro in slabo oblikovanimi definicijami (Del Gaudio, Branco 2007; Velardi in dr. 2008; Fahmi, Bouma 2006; Westerhout 2010; Kobyliński, Przepiórkowski 2008; Del Gaudio in dr. 2013 ter tudi naš poizkus v Fišer in dr. 2010), za popolnoma avtomatske pristope pa so bili uporabljeni tudi genetski algoritmi (Borg in dr. 2010) ter mreže besednih vrst (Navigli, Velardi 2010; Faralli, Navigli 2013).

4 PREDLAGANI PRISTOP ZA LUŠČENJE DEFINICIJSKIH KANDIDATOV IZ SPECIALIZIRANIH KORPUSOV

Najprej napravimo kratek shematski prikaz metodologije (Slika 1). Naš cilj je iz primerljivih korpusov, to je korpusov, ki pokrivajo določeno področje v dveh ali več jezikih (v našem primeru v slovenščini in angleščini), izluščiti terminologijo in kandidate za definicije, ki jih nato posredujemo uporabniku za namene izgradnje glosarja oz. osnutka terminološkega slovarja. Glede na naš namen puščamo definicijo definicije zelo odprto ter se ne omejimo s formalnim tipom, temveč s primernostjo definicije za vključitev v glosar. Naš pristop je polavtomatski, saj uporabnik izbere korpus, nato uporabi orodja za predprocesiranje besedil, luščenje terminologije in na novo razvite metode luščenja definicij, na koncu pa je ključna vloga spet na strani uporabnika, ki nabor terminoloških in definicijskih kandidatov uredi.



Slika 1: Polavtomatski pristop k luščenju znanja v obliki terminologije in definicij iz primerljivih specializiranih korpusov.

V nadaljevanju se posvetimo predvsem razviti metodologiji za luščenje definicijskih kandidatov (razdelek 4.1), v razdelku 4.2. pa na hitro predstavimo že obstoječa orodja, ki jih v prestavljeni metodologiji uporabljamo.

4.1 Razvita metodologija za luščenje definicij

Predlagana metodologija temelji na treh različnih pristopih in njihovih kombinacijah. Prvi sledi tradicionalnemu pristopu luščenja z uporabo leksikoskladenjskih vzorcev, drugi uporablja informacije, pridobljene z avtomatskim razpoznavanjem terminov, tretji pa temelji na luščenju stavkov, ki vsebujejo termin skupaj s svojo nadpomenko (iz semantičnega leksikona tipa wordnet). Metodologija je nadgradnja luščenja kandidatov, ki smo ga prvič predstavili v Fišer in dr. (2010).

4.1.1 LUŠČENJE DEFINICIJ NA PODLAGI LEKSIKOSKLADENJSKIH VZORCEV

Vzorce prvega pristopa smo določili za vsak jezik posebej, na podlagi analize manjšega vzorca definicijskih stavkov, uporabljajo pa leme, besedne oblike ter oblikoskladenjske oznake, kot so npr. skloni samostalnikov (za slovenščino), glagolska oseba itd. Najbolj osnoven vzorec je npr.:

[samostalniška bes. zveza v imenoval.] + je/so + [samostalniška bes. zveza v imenoval.]

Vendar pa uporabljamo veliko širši nabor vzorcev, z različnimi glagoli za definiranje ali drugimi tipičnimi strukturami.

4.1.2 LUŠČENJE DEFINICIJ NA PODLAGI IZLUŠČENIH TERMINOV

Naša druga hipoteza predpostavlja, da so stavki, pri katerih se pojavita dva strokovna izraza, dobri kandidati za definicije. Temu smo dodali dodatne pogoje, npr. da mora biti vsaj eden (ali več) izmed terminov v imenovalniku, da mora biti med dvema terminoma glagol ipd. Z različnimi nastavitvami lahko izbiramo med manjšim naborom kakovostnejših kandidatov ali pa širšim pogledom, ki vodi do več kandidatov, a je med njimi tudi več šuma.

Za prepoznavanje terminološko relevantnih enot v besedilu smo uporabili in prilagodili luščilnik terminov LUIZ (Vintar 2010), ki na podlagi oblikoskladenjskih vzorcev in izračuna terminološkosti predlaga eno- in večbesedne terminološke izraze.

Seveda niso vsi stavki, ki vsebujejo najmanj dva termina, definicije, so pa to pogosto pomensko bogati konteksti oz. okolja, bogata z znanjem in informacijami o terminu, v katerih se definicije nahajajo (angl. *knowledge-rich contexts*, Meyer 2001). Ta metoda je tudi primerna za luščenje funkcijskih definicij, definicij s tipičnimi lastnostmi ali npr. ekstenzionalnih definicij, ki so s prvim pristopom zajete le v manjši meri.

4.1.3 LUŠČENJE DEFINICIJ NA PODLAGI WORDNETA

Tretja metoda meri na tip definicij *genus et differentiae* in lušči stavke, ki vsebujejo dva izraza, od katerih je eden nadpomenka drugega. Poleg analitičnih definicij želimo s to metodo zaobjeti tudi ekstenzionalne definicije. Za luščenje stavkov s pojmi v hierarhičnem odnosu smo uporabili semantični leksikon WordNet (Fellbaum 1998) za angleščino ter sloWNet (Fišer in Sagot 2008) za slovenščino.

Tri metode lahko med sabo poljubno kombiniramo in iščemo stavke, ki so

izluščene z vsaj eno od treh metod, stavke na presečišču dveh ali treh metod ali pa uporabimo bolj kompliciranе kombinacije z različnimi nastavitvami parametrov posameznih metod.

4.2 Uporabljene obstoječe tehnologije

V tem razdelku predstavimo že obstoječe vire, orodja in programe, ki smo jih uporabili v raziskavi. Za predprocesiranje korpusa smo uporabili jezikoslovni označevalnik ToTrTaLe (Erjavec 2011), s katerim angleška in slovenska besedila segmentiramo, lematiziramo in označimo z oblikoskladenjskimi oznakami. Obstoječe orodje smo implementirali v obliki spletnega servisa, kar nam je omogočilo, da smo ga lahko vključili v razviti delotok (glej sedmo poglavje). V nadaljnjem delu (glej zadnje poglavje) bi lahko to orodje nadomestili z novejšim orodjem Obelisk za slovenščino (Grčar in dr. 2012), za angleščino pa uporabili bolj razširjeni Tree Tagger (Schmid 1994).

Za luščenje terminologije za slovenščino in angleščino smo uporabili enojezični del sistema LUIZ (Vintar 2010), ki deluje na podlagi oblikoskladenjskih vzorcev ter relativne pogostosti besed v danem korpusu v primerjavi z referenčnim korpusom. Za referenčni korpus orodje uporablja korpusa FidaPLUS (Arhar Holdt, Gorjanc 2007) za slovenščino ter BNC (2001) za angleščino, v nadaljnjih verzijah pa bi lahko orodje posodobili z novimi korpusi za slovenščino, na primer z Gigafido (Logar Berginc in dr. 2012). Orodje smo implementirali kot gradnik delotoka ter ga uporabili v našem delotoku za luščenje terminologije in pri eni izmed metod za luščenje definicij.

WordNet (Fellbaum 1998) in sloWNet (Fišer in Sagot 2008) sta leksikalni bazi oz. mreži, v katerih so besede (*literali*) združene v skupine sopomenk (*sinseti*), vsak sinset pa predstavlja svoj koncept. Sinseti oz. koncepti so v mreži organizirani z relacijami, kot sta nad- in podpomenskost, protipomenskost, meronimija (del – celota). sloWNet je podoben WordNetu, a je avtomatsko izdelan vir, sinseti pa so povezani z originalnim angleškim WordNetom. WordNet in sloWNet uporabljamo pri tretji metodi luščenja definicij.

5 ŠTUDIJA PRIMERA: LUŠČENJE DEFINICIJSKIH KANDIDATOV S PODROČJA JEZIKOVNIH TEHNOLOGIJ

V našem primeru smo si za domeno izbrali področje jezikovnih tehnologij. V slovenskem prostoru je bila prva konferenca s tega področja organizirana leta 1998, v zadnjih letih pa se področje zelo hitro razvija. Kljub temu področje še nima strukturiranih virov oz. priročnikov, v katerih bi bila terminologija področja definirana, kar za raziskovalce, ki pogosto pišejo v angleščini, predstavlja veliko težav pri iskanju pravih izrazov oz. prevodov za bolj uveljavljeno angleško terminologijo. Naš osrednji cilj je bil razviti metodologijo za luščenje definicij iz slovenskih strokovnih besedil ter jo skupaj z luščenjem terminologije implementirati v delotok, ki je preprost za uporabo. Dodatni cilj pa je bil metodologijo uporabiti na področju, ki rezultate lahko koristno uporabi. Rezultati v obliki terminologije in definicij tako predstavljajo kandidate za *Glosar jezikovnih tehnologij,* ki ga na podlagi izluščenih kandidatov ročno obdelujemo.

Za namene modeliranja izbranega področja smo zgradili primerljivi slovensko-angleški *Korpus jezikovnih tehnologij*. V prvem koraku smo zajeli članke konference Jezikovne tehnologije, ki v Sloveniji od leta 1998 dalje poteka vsako drugo leto. Zajeli smo vse članke od začetkov pa do vključno leta 2010. Članki konferenčnih zbornikov so glede na jezik razvrščeni v slovenski ali angleški del korpusa. Članke smo iz formata PDF pretvorili v format TXT, ta osrednji del korpusa pa je bil tudi dodatno prečiščen (npr. popravljene napake ob pretvorbi formatov, prelomu strani, mesta opomb ipd.). Korpus konferenčnih zbornikov smo v drugi fazi nadgradili z drugimi tipi besedil, kot so diplomske, magistrske in doktorske naloge, članki v znanstvenih revijah, poglavja iz knjig ter prispevki v Wikipediji. Tudi ta del korpusa smo pretvorili v format TXT, zaradi njegove obsežnosti in heterogenosti pa smo manj napora posvetili dodatnemu čiščenju korpusa. Korpus smo nato segmentirali, lematizirali in besedam pripisali oblikoslovne oznake z orodjem ToTrTaLe (Erjavec 2011). Vsak članek v korpusu ima svojo identifikacijsko oznako, prav

tako pa smo unikatne identifikacijske oznake pripisali tudi vsakemu stavku v korpusu, saj se je tako ne glede na to, za kaj korpus uporabljamo, vedno mogoče vrniti na izvorno besedilo.

Velikost slovenskega dela korpusa je 903.189 različnic, velikost angleškega dela *Korpusa jezikovnih tehnologij*, ki je bil zgrajen kot primerljiv slovenskemu delu, pa je 909.606 različnic (brez ločil).

Velikost korpusa	**Slovenski del**	**Angleški del**	**Skupaj**
Stavki	44.749	43.018	87.767
Različnice (brez ločil)	903.189	909.606	1.812.795
Različnice (z ločili)	1.089.968	1.073.470	2.163.438

Tabela 1: Velikost Korpusa jezikovnih tehnologij.

6 EKSPERIMENTI IN REZULTATI

To je osrednje poglavje prispevka, saj v njem obravnavamo eksperimente in rezultate luščenja definicij iz izdelanega *Korpusa jezikovnih tehnologij*. Najprej predstavimo mere za vrednotenje metod luščenja definicij (natančnost in priklic), nato predstavimo eksperimente in rezultate. Rezultati niso neodvisni od uspešnosti obstoječih tehnologij, ki jih uporabljamo, zato na kratko povzamemo tudi rezultate luščilnika terminologije LUIZ (Vintar 2010) ter omenimo nekatere napake označevalnika ToTrTaLe (Erjavec 2011). Nadaljujemo s predstavitvijo kvalitativnih oznak, s katerimi smo označili različne tipe definicijskih kandidatov v korpusu, končamo pa z eksperimentom izračuna strinjanja med ocenjevalci.

6.1 Metodologija vrednotenja

Najprej predstavimo metodologijo vrednotenja sistema za luščenje definicij. Za kvantitativni del evalvacije uporabimo meri *natančnost* in *priklic*. Natančnost označuje odstotek definicij izmed vseh izluščenih stavkov, ki jih sistem predlaga kot definicijske kandidate. Priklic meri, koliko izmed vseh definicij iz korpusa

sistem pravilno zazna. V večini izvedenih eksperimentov podamo dejansko natančnost, saj evalviramo vse izluščene kandidate, a le oceno priklica, saj ne poznamo dejanskega števila vseh definicij v korpusu. V ta namen uporabimo nabor 150 definicij, na katerih merimo priklic.

Pri kvantitativnem vrednotenju torej vsakemu stavku, ki je izluščen kot definicijski kandidat, pripišemo binarno oznako, glede na to, ali stavek *je* ali *ni* definicija. Kriterij vrednotenja izhaja predvsem iz namena aplikacije, torej v našem primeru luščenja kandidatov za vključitev v glosar. Kot definicije označimo kandidate, ki neki pojem definirajo tako, da so primerni za vnos oz. potrebujejo manjšo ročno spremembo. Definicij ne omejimo na poseben podtip in so lahko analitične, ekstenzionalne, funkcijske ali kakršnega koli drugega tipa, vendar pa ne smejo biti izven domene, preširoke in tako dalje. Seveda sama odločitev o oznaki ni vedno lahka, na kar kažejo tudi eksperimenti ocene strinjanja med označevalci.

Poleg kvantitativnih rezultatov je zanimiva tudi analiza definicijskih kandidatov, kar predstavimo v razdelku 6.3. Nabor definicijskih kandidatov smo označili tudi z natančnejšimi oznakami. Poleg že omenjenih binarnih kategorij, ki jima dodamo še kategoriji *mejna definicija* in *mejna definicija,* kandidatom pripišemo še vrsto oznak, ki se nanašajo na obliko definicije, njeno vsebino, na definiendum ali pa na za luščenje iz korpusov specifične probleme segmentacije.

6.2 Eksperimenti in kvantitativni rezultati

Prvi del zajema luščenje definicij s tremi metodami in njihovimi osnovnimi kombinacijami iz slovenskega dela korpusa, drugi del pa luščenje iz angleškega dela. V Tabelah 2 in 3 podamo rezultate osnovnih treh metod, osnovnih kombinacij (unije in preseka, pri čemer s presekom mislimo na presek vsaj dveh metod) ter izbranih kompleksnejših kombinacij. Celoten set eksperimentov in podrobnejša razlaga nastavitve parametrov izbranih kombinacije je predstavljen v Pollak (2014).

6.2.1 LUŠČENJE DEFINICIJ IZ SLOVENSKEGA DELA KORPUSA

Najprej podamo rezultate, pridobljene z uporabo vsake od treh metod. V Tabeli 2 vrstica *Vzorci* označuje metodo luščenja z vzorci, vrstica *Termini* luščenje z uporabo izluščene terminologije, oznaka sloWNet pa se nanaša na tretjo metodo iskanja stavkov s terminom sloWNet in njegovo nadpomenko. Sledijo kombinacije metod, ki so razložene spodaj. Za vsako metodo podamo število kandidatov (z vzorci smo na primer izluščili 1728 stavkov), natančnost metode pomeni odstotek definicij izmed kandidatov (v oklepaju navedemo število pozitivno ocenjenih definicij) ter oceno priklica, ki smo ga izračunali na podlagi testnega nabora 150 definicij.

Izbor metod na slovenskem korpusu	**Število izluščenih kandidatov**	**Natančnost (št. definicij)**	**Priklic**
Vzorci	1728	0,2251 (389)	0,5867
Termini[4]	721	0,1747 (126)	0,0467
sloWNet	4670	0,0570 (270)	0,2533
Unija	6606	0,0978 (646)	0,7000
Presek	489	0,2638 (129)	0,1800
Komb. A[5] za priklic	2382	0,2040 (486)	0,6130
Komb. B[6] za natanč.	336	0,3180 (107)	0,1730

Tabela 2: Izbor rezultatov metod in njihovih kombinacij na slovenskem delu.

a) *Luščenje z vzorci*

[4] Izmed številnih eksperimentov, predstavljenih v disertaciji (Pollak 2014), tu podajamo rezultate le za izbrane kombinacije. Pri terminih navajamo nastavitev, kjer upoštevamo zgornji 1 % izluščenih terminov, stavki imajo glagol med dvema terminoma, en termin se pojavlja na začetku stavka, stavek pa ima ali 5 terminov ali pa 4, vendar vsaj enega v nominativu, 2 večbesedna termina in enega na začetku stavka.

[5] Stavki, izluščeni ali z vzorci ali s termini.

[6] Stavki, izluščeni z vzorci, ki so hrati izluščeni še s temini ali sloWNetom.

Prva metoda, tj. *luščenje z vzorci,* v slovenščini zajema dvanajst vzorcev. Najbolj preprost vzorec išče stavke, v katerih sta dve samostalniški besedni zvezi v imenovalniku povezani z glagolom *biti* v tretji osebi. Pogosto samostalniški besedni zvezi sledi angleški prevod termina ali pa sta v slovenščini podani dve samost. besedni zvezi kot alternativna poimenovanje. Primer definicije iz našega korpusa, ki jo lahko izluščimo s prvim vzorcem, je podan v primeru (ii). Podčrtani del označuje vzorec »N je N«, na podlagi katerega je stavek izluščen.

ii. Lematizacija je postopek pripisovanja osnovne oblike besedam v korpusnem besedilu.

Ostali vzorci vsebujejo glagole, kot so *definirati, opisati, poimenovati* ter vrsto drugih. Nekateri vzorci tudi ciljajo na parafraze in sinonime (npr. »NP imenujemo tudi NP«, kjer NP označuje samostalniško besedno zvezo, angl. *noun phrase*) ali funkcijske definicije (»naloga NP je …«), ponazorjena s korpusnima primeroma spodaj:

iii. Inventar jezikovnih poimenovanj pojmov neke stroke imenujemo tudi terminologija, na primer geološka, medicinska, planinska terminologija.

iv. Naloga oblikoslovnih označevalnikov besedil je določevanje besednih vrst (angleško "part-of-speech") ali še natančnejših oblik znotraj besednih vrst besedam v besedilu.

Z osnovnim vzorcem, ki uporablja samostalniški besedni zvezi v imenovalniku, povezani z glagolom *biti,* smo izluščili 259 definicij z 20-odstotno natančnostjo. Če pa vzamemo vseh dvanajst tipov vzorcev, ki smo jih ročno definirali, izluščimo iz korpusa 389 definicij z 22,5-odstotno natančnostjo, kar je tudi rezultat, podan v Tabeli 2. Na testnem korpusu 150 definicij smo izračunali priklic (v našem primeru nekaj pod 60 %), kar je tudi ocena odstotka vseh izluščenih definicij proti vsem definicijam v korpusu.

b) *Luščenje s termini*

Drugi pristop izhaja iz osnovne hipoteze, da definicije vsebujejo vsaj dva terminološka izraza. Temu osnovnemu pogoju dodamo še vrsto drugih pogojev,

s katerimi omejimo izbor kandidatov, saj je jasno, da število terminov še ni zadosten kriterij za luščenje definicij, omogoča pa zaznavo z informacijami bogatih jezikovnih okolij. Luščenje s pomočjo terminov omogoča uporabo različnih nastavitev glede na to, ali si uporabnik želi višjo natančnost (manjše število izluščenih stavkov, a večji odstotek definicij med njimi) ali višji priklic (predvsem, kadar metodo uporabljamo v kombinaciji z drugima dvema metodama) ali pa najboljši kompromis med obema.

Določili in preizkusili smo različne nastavitve parametrov, ki jih lahko uporabimo za reguliranje višje natančnosti. Natančnost je večja, če upoštevamo le termine z višjo terminološko vrednostjo. Boljši kandidati za definicije imajo več terminoloških izrazov, ne le dveh. Boljša natančnost je, če uporabimo pogoj, da med dvema terminoma stoji glagol. Naslednji pogoj, s kateri izboljšamo natančnost, je termin na začetku stavka (začetek na prvi ali drugi besedi stavka). Natančnost je boljša, če je prvi termin večbesedni terminološki izraz ter tudi če je več izmed zaznanih terminov večbesednih izrazov. Zadnji, in morda za slovenščino najbolj pomemben, pa je pogoj, koliko terminov mora biti v imenovalniku, s čimer rahlo ciljamo na tipe definicij »X je Y«, vendar brez omejevanja glagola na glagol *biti* oz. ostale vnaprej določene glagole kot pri pristopu z vzorci.

Z dokaj strogimi pogoji dosežemo okoli 26-odstotno natančnost, a z njimi izluščimo le 27 definicij. Z izbrano zmernejšo kombinacijo pogojev pa izluščimo 126 definicij z natančnostjo okoli 17,5 % (metoda, katere rezultate povzamemo tudi v Tabeli 2). Z nastavitvami z zelo nizko natančnostjo pa lahko izluščimo tudi nad 85 % definicij.

Na splošno imamo pri luščenju z vzorci boljše sorazmerje med natančnostjo in priklicem, vendar ima luščenje z uporabo terminov tudi nekaj prednosti. Je bolj ohlapno in omogoča luščenje definicijskih stavkov, ki uporabljajo glagole, ki jih nismo vnaprej definirali, kar je še posebej pomembno pri definiranju termina z njegovo rabo (glej funkcijski definiciji spodaj) ali tipičnimi lastnostmi. V spodnjih primerih so podčrtani vsi termini, od nastavitev pa je odvisno, koliko

odstotkov terminov upoštevamo. V obeh primerih vidimo tudi, da je prvi termin večbesedni ter da je glagol med prvima dvema terminoma.

v. Pomnilnik prevodov hrani prevodne enote, tj. segmente (ponavadi povedi) nekega originala in njihove prevode.

vi. Besedna skica prikazuje leksikalni profil izbrane iztočnice s podatki o njenem tipičnem sobesedilnem okolju (Gantar in dr. 2009: 33).

Metoda omogoča tudi luščenje kompleksnejših stavkov, v katerih je vsebovana definicija, poleg tega pa je prednost tudi to, da je metoda veliko manj odvisna od napak pri jezikoslovnem označevanju v predprocesiranju.

c) *Luščenje s sloWNetom*

V tretji metodi luščimo definicije s pomočjo semantičnega leksikona tipa wordnet. Za luščenje definicijskih kandidatov iz slovenskih besedil smo uporabili semantični leksikon sloWNet (Fišer, Sagot 2008), s pomočjo katerega smo iz korpusa izluščili vse tiste stavke, v katerih se pojavita najmanj dva pojma iz sloWNeta in je hkrati eden direktna nadpomenka drugega. Ta metoda je najmanj natančna, saj izluščimo 270 definicij z manj kot šestodstotno natančnostjo, kar lahko pripišemo temu, da pari nad- in podpomenk iz wordneta niso specifični za področje jezikovnih tehnologij, ki je v sloWNetu slabo pokrito. Ta metoda je zato primerna predvsem v kombinaciji z drugimi metodami, saj kot samostojna metoda lušči predvsem stavke izven domene, kot na primer spodnji, v katerem sta para besed *zaključek* in *poglavje* (glej podčrtane pare iz leksikona sloWNet v spodnjih primerih).

vii. Zaključek bomo podali v petem poglavju.

Metoda deluje bolje, če so teksti splošnejši. To se vidi predvsem v eksperimentih, ki smo jih izvedli v Fišer in dr. (2010), v našem korpusu pa se razlika vidi že v stavkih, ki obravnavajo bolj jezikoslovne kot jezikovnotehnološke teme. V primeru, podanem spodaj, sta v sloWNetu direktni nadpomenki besedi *homonim* in *beseda*. V nadaljnjem delu bomo tudi preverili, ali na metodo vpliva upoštevanje vseh nadpomenk in ne le direktnih.

viii. Homonimi ali enakozvočnice so besede, ki sicer imajo enako glasovno podobo in več pomenov, med njimi pa ni videti kake metonimične ali metaforične povezanosti, niti si v danem trenutku ne moremo misliti, da bi bila taka zveza kdaj obstajala.

d) *Kombinacije metod*

Poleg dveh najbolj očitnih kombinacij, to je *unije* metod (stavki, izluščeni z vsaj eno izmed treh metod) ali *preseka*, s čimer mislimo na stavke, ki so izluščeni vsaj z dvema metodama, lahko uporabimo vrsto drugih kombinacij, saj lahko uporabljamo različne nastavitve parametrov pri metodi za luščenje s termini. Različne kombinacije metod so naravnane k boljši natančnosti ali priklicu. Z 20-odstotno natančnostjo iz slovenskega korpusa izluščimo 486 definicij, medtem ko jih z nekaj manj kot 32-odstotno natančnostjo izluščimo 107.

6.2.2. LUŠČENJE DEFINICIJ IZ ANGLEŠKEGA DELA KORPUSA

Naslednje podpoglavje obravnava prilagoditev metode na angleščino in definicij iz angleških dokumentov *Korpusa jezikovnih tehnologij*.

Izbor metod na angleškem korpusu	**Število izluščenih kandidatov**	**Natančnost (št. definicij)**		**Priklic**
Vzorci				
(i) začetek stavka	562	0,3292	(185)	0,2533
(ii) variab. zač.st.	702		0,2849	0,2733
(iii) kjer koli v st.	2283		(200)	0,3730
		0,1196	(273)	
Termini[7]	1092	0,0824	(90)	0,1333
WordNet	3258	0,0430	(140)	0,2667
Unija[8]	4727	0,0728	(344)	0,5400

[7] Izbrana nastavitev: 5 terminov iz zgornjih 10 % izluščenih terminov, glagol med prvima dvema terminoma, termin začetku stavka, prvi termin je večbesedna zveza. Za podrobnost izbora glej Pollak (2014).

[8] Unija in presek uporabljata nastavitev metode z vzorci (ii).

Presek	318	0,2579	(82)	0,1333
Komb. A[9] za natanč.	107	0,5420	(58)	0,0867
Komb. B[10] prik. & nat.	1022	0,2250	(230)	0,3333

Tabela 3: Izbor rezultatov metod in njihovih kombinacij na angleškem delu.

a) *Luščenje z vzorci*

Pri prvi metodi, tj. metodi luščenja z uporabo leksikoskladenjskih vzorcev, smo za angleščino prilagodili vzorce za luščenje definicij iz slovenskih besedil. Preizkusili smo različne nastavitve, in sicer eno (i), v kateri se vzorec začne na začetku stavka, ter drugo (iii), v kateri vzorce iščemo kjer koli v stavku. Za to dodatno opcijo pogoja začetka stavka smo se odločili zato, ker v angleščini skloni niso izraženi na enak način kot v slovenščini in tako v vzorcih ni mogoče uporabiti enakih restriktivnih pogojev kot v slovenščini z uporabo imenovalnika. Začetek stavka se pokaže kot dober pogoj za višjo natančnost, ki pa gre na račun nižjega priklica. Z izbranim pogojem začetka stavka se natančnost poviša za 10 %, saj brez tega pogoja izluščimo 273 definicij z natančnostjo okrog 12 %, z dodatnim pogojem pa 185 definicij s približno 33-odstotno natančnostjo. Testiramo tudi vmesno rešitev (ii), pri kateri dopuščamo bolj raznolike začetke stavkov, s čimer izluščimo 200 definicij z 28,5-odstotno natančnostjo.

b) *Luščenje s termini*

Pri luščenju definicij s pomočjo terminoloških kandidatov iz angleških besedil preverjamo enake hipoteze kot v slovenščini z izjemo pogoja terminov v imenovalniku. Slednje je verjetno tudi glavni razlog za slabše delovanje sistema

[9] Vzorci (i), ki so hkrati izluščeni tudi s termini ali WordNetom.

[10] Vzorci (iii), ki so hkrati izluščeni tudi s termini ali WordNetom, ter Vzorci (ii), ki so jim dodani stavki izluščeni z najbolj restriktivnimi nastavitvami metode luščenja s termini (stavki s 5 termini, od katerih so vsaj trije večbesedni izrazi, z vrednostjo nad mejo zgornjega 1 % terminoloških kandidatov, glagol med prvima dvema terminoma, večbesedni termin na začetku stavka).

v angleškem jeziku (natančnost večinoma pod 10 %).

c) *Luščenje z WordNetom*

Pri luščenju s pomočjo WordNeta (Fellbaum 1998) velja podobno kot pri slovenščini, da luščimo preveč splošne pare nad- in podpomenk in ne parov, ki so specifični za domeno. Natančnost je podobna kot pri slovenščini (le 4 %), kar pomeni, da testirana metoda ni uporabna za samostojno rabo, kljub večjemu obsegu in natančnosti angleškega semantičnega leksikona.

d) *Kombinacije metod*

Poleg *unije,* ki ima dober priklic (nad 50 %), in *preseka* smo tudi za angleščino preizkusili različne kombinacije metod. Glede na naše izbire lahko dosežemo 54-odstotno natančnost (a tako izluščimo le 58 definicij, glej kombinacija A), medtem ko z drugo predstavljeno kombinacijo (B) z 22,5-odstotno natančnostjo izluščimo 230 definicij.

6.1.1 REZULTATI VKLJUČENIH OBSTOJEČIH TEHNOLOGIJ

Na koncu na kratko evalviramo označevalnik ToTrTaLe ter luščilnik terminov LUIZ, saj sta to tehnologiji, ki imata bistven vpliv na rezultate luščenja definicij.

Označevalnika ToTrTaLe nismo sistematično ovrednotili, vendar smo opazili različne vrste napak. Tiste napake, ki se sistematično pojavljajo, lahko delno odpravimo s kratkim, na osnovi pravil zasnovanim programom, ki ga lahko zaženemo z izbiro dodatnega parametra v delotoku oz. gradniku ToTrTaLe. Napake izhajajo iz napačne segmentacije stavkov, napačnega pripisovanja oblikoskladenjskih oznak ter napačne lematizacije. Napačna segmentacija, vezana na razpoznavanje kratic (npr. *et al.*) lahko močno vpliva na luščenje definicij iz znanstvenih besedil, napačne oblikoskladenjske oznake pa predvsem na luščenje z vzorci, pa tudi na pogoj imenovalnika v metodi luščenja z vzorci.

Pri evalvaciji luščilnika terminologije LUIZ podamo natančnost in priklic te komponente, izračunamo pa tudi oceno strinjanja med ocenjevalci. Natančnost

ocenimo od 1 do 5 in ocenimo 200 najboljših kandidatov, kjer dva označevalca za okrog 20 % izluščenih terminov navedeta, da gre za popoln termin, tj. za polno leksikalizirano besedno zvezo, ki označuje koncept s področja jezikovnih tehnologij, okrog 90 % kandidatom pa kandidata podata pozitivno oceno. Priklic se meri tako, da je na manjšem podkorpusu strokovnjak označil vse termine, za katere smo nato izmerili, koliko jih zazna sistem LUIZ. Če upoštevamo zgornjih 200 terminoloških kandidatov, je priklic približno 25 %, če pa vse, je okoli 71 % za slovenščino ter 82 % za angleščino. Napake sistema LUIZ vplivajo na luščenje s pomočjo terminov in v nadaljnjem delu bi bilo smiselno, da po fazi luščenja terminologije uporabnik najprej ovrednoti oz. prefiltrira avtomatsko izluščeno terminologijo.

6.3 Kvalitativna analiza

Poleg kvantitativnih binarnih kategorij, ali je izluščeni stavek definicija ali ne, pri evalvaciji na podmnožici kandidatov pripišemo tudi dodatne oznake. Te so kvalitativne narave in označujejo mejne primere, preveč splošne ali preveč specifične definicije ipd.

Analizirali smo 3404 stavke s 716 definicijami (skupno za slovenščino in angleščino, za slovenščino smo vzeli stavke iz kombinacije A, za angleščino pa tiste iz kombinacije B). Te stavke smo po razvrstitvi v glavni kategoriji glede na to, ali je stavek definicija ali ne, označili s podkategorijami. Poleg nedvoumnih stavkov, ki jasno pripadajo kategoriji definicij in nedefinicij (brez dodatnih oznak), ter oznak za manj jasne primere (označene z vprašajem) smo kandidatom pripisali oznake, vezane na *obliko definicije* (npr. ekstenzionalne definicije, definicije brez hipernima, kot so funkcijske definicije, definicije samo z nadpomenko (razvrstitvene definicije), oznake, vezane na *vsebino definicije* (npr. presplošne ali prespecifične), *definiendum* (označili smo, ali gre za lastna imena, kratice, termine, ki niso iz obravnavane domene), *segmentacijo* (kjer so označeni kandidati, ki imajo napako pri segmentaciji, ter tisti primeri, v katerih se ena definicija razteza čez več stavkov oz. en stavek vsebuje več definicij).

Zadnja kategorija vsebuje oznake, ki se nanašajo na evalvacijo (npr. stavke, ki smo jih sami ali pa različni ocenjevalci v zaporednih vrednotenjih različno označili).

Z namenom ponazoritve kategorije oznak, vezanih na vsebino definicije, analizirajmo spodnja dva stavka, ki smo ju označili kot nedefinicije.

ix. Vreča besed (angl. bag of words ali BOW) je preprosta tehnika preoblikovanja besedila za potrebe klasifikacije.

x. Lematizacija pa je proces, kjer odstranimo besedam sklanjatev in množino.

Prvi stavek sicer terminu *vreča besed* doda angleško poimenovanje ter nadpomenko (*tehnika preoblikovanja besedila*), vendar je stavek po eni strani presplošen, da bi bil definicija, saj ne pove, za kakšno vrsto predprocesiranja oz. preoblikovanja besedila gre. Po drugi strani je termin definiran z uporabo *za potrebe klasifikacije,* vendar tudi to ni dovolj, saj se za potrebe klasifikacije uporabljajo tudi druge metode, poleg tega pa se format besedila v obliki *vreče besed* uporablja tudi za druge potrebe, npr. za gručenje. Drugi stavek, ki smo ga prav tako označili za nedefinicijo, pa *lematizacijo* definira prespecifično, saj je spodnji stavek zelo nenatančen, veljaven pa na primer le za samostalnike.

Ta primera tudi zgovorno pričata o tem, kako zahtevna je naloga avtomatskega luščenja definicij iz korpusa. Po drugi strani pa vidimo, da so lahko tudi izluščeni stavki, ki niso definicije, zelo informativni in uporabni, če za izbrane termine nimamo boljših kandidatov.

Odločitev, ali je neki stavek definicija ali ne, ni vedno očitna, kar se kaže tudi v izračunu strinjanja med ocenjevalci. V eksperimentu z 21 ocenjevalci smo ocenili 15 stavkov ter izračunali Randolphovo variacijo statistike kappa[11] (Randolph 2008), v kateri 0 pomeni naključno strinjanje, –1 in 1 pa popolno nestrinjanje oz. popolno strinjanje. Rezultati strinjanja med označevalci so 0,36, kar je veliko manj kot 0,7, kar označuje dober rezultat strinjanja med

[11] justusrandolph.net/kappa/

ocenjevalci. Razlike med ocenjevalci glede na tip kandidata za definicijo ponazorimo s stavkoma spodaj, kjer se pri prvi, analitični definicij s strukturo *genus et differentiae* vsi ocenjevalci strinjajo, da gre za definicijo, v drugem primeru pa se mnenja ocenjevalcev najbolj razhajajo.

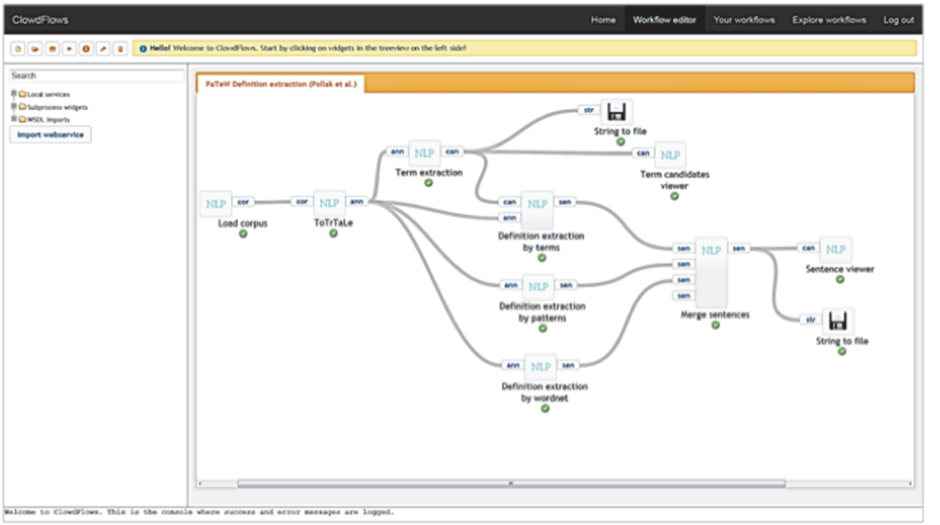
xi. Lombardov efekt je pojav, pri katerem govorec poveča glasnost govora ob povečanju glasnosti šuma ozadja.

xii. Google Translate je tipični pripadnik sistemov statističnega strojnega prevajanja (Statistical Machine Translation-SMT), ki je predstavljena v Razdelku 3.

7 IMPLEMENTACIJA METODOLOGIJE

V tem poglavju predstavimo delotok, ki implementira našo metodologijo in omogoča njeno enostavno uporabo. Podrobnosti implementacije posameznih gradnikov so bile že predstavljene ob prvi verziji delotoka (glej Pollak in dr. 2012a, 2012b), v novi verziji pa je dostopnih več parametrov, spodnji opis pa je manj tehnično naravnan in pisan bolj z vidika uporabnika. Za implementacijo delotoka smo izbrali okolje ClowdFlows (Kranjc in dr. 2012), aplikacijo v oblaku, ki omogoča, da brez namestitev katerih koli programov dostopamo do že zgrajenih delotokov ali gradimo nove delotoke iz poljubnega brskalnika.

V nadaljevanju predstavimo posamezne gradnike prosto dostopnega *delotoka* (angl. *workflow*), prikazanega na Sliki 2. *Gradniki* (angl. *widgets*) so vizualno predstavljeni deli programov, ki imajo definirane vhodne in izhodne oblike podatkov, parametre pa lahko uporabnik ročno izbere. Implementirajo lahko operacije spletnih servisov (angl. *web services*) ali pa so implementirani lokalno v okolju ClowdFlows. Posamezne gradnike lahko uporabnik na kanvasu kombinira s principom primi-odloži in sestavlja nove delotoke.



Slika 2: Izdelan delotok za luščenje terminologije in definicij, dostopen na http://clowdflows.org/workflow/1380/.

S prvim gradnikom *Load corpus* uporabnik naloži poljubni korpus, ki je lahko v različnih formatih: PDF, DOC, DOCX, TXT ali HTML, poleg tega pa gre lahko za samostojne dokumente ali datoteke ZIP.

Gradnik *ToTrTaLe* kliče spletni servis za označevanje besedil. Uporabnik izbere med jezikoma slovenščina ali angleščina, dodatne opcije (parametri), ki jih lahko izbere, pa so: postprocesiranje, s katerim popravimo nekatere napake segmentacije in oblikoskladenjskega označevanja, izvozni format XML, za slovenščino pa je na voljo tudi označevanje stare slovenščine (ta del procesa, nalaganje korpusa in označevanje z označevalnikom ToTrTaLe, je na voljo tudi v samostojnem delotoku (glej Pollak in dr. 2012b): http://clowdflows.org/workflow/228/.

Naslednji gradnik delotoka *Term extraction* implementira nekoliko prilagojen luščilnik terminologije LUIZ (Vintar 2010). Uporabnik izbira med slovenščino in angleščino.

Sledi glavni del, spletni servis za luščenje definicijskih kandidatov. Spletni servis ima tri operacije, prva je implementirana v gradniku *Definition*

extraction by patterns, kjer uporabnik določi jezik in na podlagi vnaprej določenih leksikoskladenjskih vzorcev izlušči stavke, ki vzorcem ustrezajo. Dodatni parameter omogoča, da uporabnik izbere, ali se mora vzorec obvezno nahajati na začetku stavka ali kjer koli (trenutno ta parameter uporabljamo le za angleščino).

Drugi gradnik za luščenje definicij (*Definition extraction by terms*) omogoča luščenje informacijsko bogatih stavkov, kjer se uporabnik lahko odloči med različno strogimi parametri, predvsem v odvisnosti od tega, ali želi metodo uporabljati samostojno ali v kombinaciji z drugimi metodami. Razpoložljivi parametri implementirajo hipoteze, ki smo jih omenili v opisu prejšnjega poglavja, in sicer število terminov, terminov v imenovalniku, večbesednih terminov, termin na začetku stavka, glagol med dvema terminološkima izrazoma.

Zadnji gradnik za luščenje definicij *Definition extraction by wordnet* implementira luščenje kandidatov z uporabo wordneta.

Sledi še nekaj dodatnih gradnikov. *Merge sentences* omogoča kombiniranje izluščenih definicijskih kandidatov na različne načine. Vzamemo lahko vse kandidate in izbrišemo le dvojnike (unija), lahko pa izberemo tiste stavke, ki se pojavljajo v vsaj dveh oz. v vseh treh metodah. Ker je delotok modularen, pa lahko uporabnik naredi poljubno kombinacijo različnih metod. Dodatni gradniki, ki niso implementirani kot spletni servisi, temveč kot lokalni gradniki, so *Term viewer, Sentence viewer* in *String to file,* od katerih prva dva omogočata ogled izluščenih terminov (z lemami ter v kanonični obliki) ter definicijskih kandidatov, tretji pa se uporablja za shranjevanje rezultatov.

8 ZAKLJUČKI IN NADALJNJE DELO

Glavni cilj prispevka je predstavitev razvite metodologije, ki uporabniku omogoča (pol)avtomatsko izluščiti model področnega znanja iz nestrukturiranih besedil v obliki terminologije in definicij. Osrednji del metodologije predstavlja luščenje definicij s kombinacijo treh metod (luščenja

z vzorci, luščenja z uporabo terminov in luščenja z uporabo parov pod- in nadpomenk iz wordneta). Metodologija je dostopna za luščenje iz slovenskih in angleških besedil, vendar luščenju za slovenščino posvetimo več pozornosti, saj zanjo podobne metode še ne obstajajo. Dodatni prispevek je implementacija celotnega procesa – od nalaganja korpusa do pregleda izluščene terminologije in definicij – v obliki javno dostopnega delotoka, ki je preprost za uporabo v prevajalske, jezikoslovne ali terminografske namene. Posamezne komponente delotoka – med njimi tudi orodje za jezikoslovno označevanje korpusov v slovenskem in angleškem jeziku – pa so na voljo za vključevanje v druge delotoke procesiranja naravnega jezika.

Uporabnost metodologije smo preizkusili na primeru v ta namen zgrajenega primerljivega *Korpusa jezikovnih tehnologij*[12] v angleškem in slovenskem jeziku. Izluščili in ovrednotili smo veliko število definicijskih kandidatov, končni izbor definicijskih kandidatov pa uporabljamo pri nastajajočem *Glosarju jezikovnih tehnologij*.[13]

V primerjavi z delom drugih avtorjev ima naša metoda primerljive rezultate kot metode drugih slovanskih jezikov (npr. Przepiórkowski in dr. 2007), vendar slabše kot najbolje delujoče metode za angleščino (npr. Navigli, Velardi 2010 poročata skoraj 100-odstotno natančnost). Vendar, če rezultate primerjamo z metodami, ki nalogo prav tako definirajo z luščenjem stavkov iz podobnih korpusov znanstvenih besedil (npr. Reiplinger in dr. 2012 luščijo definicije iz antologije člankov računalniškega jezikoslovja ACL Anthology)[14] in ne uporabljajo spleta, postanejo razlike bistveno manjše.

Naše delo ima tudi kar nekaj prednosti. Celotni proces je implementiran v obliki prosto dostopnega delotoka, ki je na voljo vsem uporabnikom brez potrebnega predznanja ali namestitve sistema, kar je – po našem védenju – edini tovrstni

[12] Del slovenskega korpusa, ki zajema konferenčne zbornike, je prosto dostopen prek konkordančnika na naslovu http://nl.ijs.si:3003/cuwi/sdjt_sl.

[13] http://kt.ijs.si/senja_pollak/jt_glosar/

[14] http://aclweb.org/anthology/

sistem. Poleg luščenja s pomočjo leksikoskladenjskih vzorcev, ki ima že dolgo tradicijo (prim. Hearst 1992), uvedemo tudi bolj ohlapni metodi, s pomočjo terminov in wordneta, pri katerih se ne omejujemo na posamezni tip definicije. Vse tri metode pa lahko med seboj tudi kombiniramo. Za razliko od nekaterih drugih pristopov z uporabo strojnega učenja (npr. Navigli, Velardi 2010) pa naša metodologija ne zahteva vnaprej ročno označenih korpusov.

Doslej smo metodo aplicirali le na en korpus, na domeno jezikovnih tehnologij, kar želimo v nadaljevanju razširiti. Relativno nizko natančnost in priklic delno pripisujemo dokaj zahtevnemu izražanju v akademskih člankih, ki le redko zajemajo najbolj tipične oblike definicij. Naša metoda omogoča, da najdemo tudi netipične definicije, vendar moramo pregledati dokaj veliko število kandidatov, da dobimo dober izbor pravih definicij.

V nadaljnjem delu bomo delo razširili na več ravneh. Ker je delotok modularen, lahko vključimo oz. zamenjamo nekatere gradnike delotoka, pod pogojem, da so alternativne komponente na voljo v obliki spletnih servisov. Za jezikoslovno označevanje besedil bi bilo smiselno preizkusiti vpliv orodja Tree Tagger (Schmid 1994) za angleščino ali nedavno razviti Obeliks za slovenščino (Grčar in dr. 2012), za luščenje terminologije pa sistema avtorjev Sclano in Velardi (2007) ali Macken in dr. (2013). Poleg tega bomo luščenje definicij poskusili izboljšati s strojnim učenjem (začetne eksperimente smo predstavili v Fišer in dr. (2010), v novih eksperimentih pa bomo značilke gradili tudi z uporabo atributov, ki smo jih predstavili v pričujočem delu). V metodologijo in implementacijo bomo dodali poravnavo terminov, izluščenih iz primerljivih korpusov z metodo, predstavljeno v Fišer in dr. (2011). Težišče nadaljnjega dela bo na preizkušanju metodologije na novih, tudi bolj poljudnih besedilih ter primerjavi metode z deli drugih avtorjev, ko bomo poskusili za slovenščino prilagoditi metodo, predstavljeno v Faralli in Navigli (2013). Preučili bomo tudi uporabnost predstavljene metode znotraj metod odkrivanja znanj iz besedil. Pri prikazani uporabi v terminografske namene pa bi morali metodologijo povezati z vmesnikom, primernim za urejanje vnosov, ter razviti možnosti za

kolaborativno urejanje zadetkov. Ko bomo zaključili z vnosom definicij v *Glosar jezikovnih tehnologij*, bomo razmislili o nadgradnji glosarja ter vključitvi zbirke v portal Termania[15] ali pa o povezavi zbirke z iSlovarjem.[16]

ZAHVALA

Zahvaljujem se mentorici doktorske disertacije izr. prof. dr. Špeli Vintar, študentkama Janji Sterle in Živi Malovrh za pomoč pri pripravi korpusa in glosarja ter Anžetu Vavpetiču in Nejcu Trdinu z Instituta Jožef Stefan za pomoč pri implementaciji delotoka.

LITERATURA

ACL Anthology. S. Bird in M.-Y. Kan (ur.). Dostopno prek: http://aclweb.org/anthology/ (12. februar 2014).

Ahmad, K., Gillam, L., in Tostevin, L. (2007): University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Rerieval (WILDER). *Proceedings of the Eight Text REtrieval Conference (TREC-8)*: 717–724. Gaithersburg.

Arhar Holdt, Š., in Gorjanc, V. (2007): Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo*, 52 (2): 95–110.

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

Ayto, J. (1983): On Specifying Meaning: Semantic Analysis and Dictionary Definitions. V R. Hartmann (ur.): *Lexicography: Principles and Practice*: 89–98. London: Academic Press.

Béjoint, H. (2000): *Modern Lexicography: An Introduction*. Oxford: Oxford University Press.

[15] http://www.amebis.si/termania
[16] http://www.islovar.org

BNC (2001): *The British National Corpus (version 2: BNC World, Mike Scott's lists)*. Dostopno prek: http://www.lexically.net/downloads/BNC_wordlists/downloading BNC.htm (11. januar 2011).

Borg, C., Rosner, M., in Pace, G. J. (2010): Automatic Grammar Rule Extraction and Ranking for Definitions. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*: 2577–2584. Valletta.

Borsodi, R. (1967): *The definition of definition*. Boston: Porter Sargent Publisher.

Copi, I. M., in Cohen, C. (2009): *Introduction to Logic*. Upper Saddle River, New Yersey: Pearson, Prentice Hall.

Degórski, L., Kobyliński, Ł., in Przepiórkowski, A. (2008): Definition extraction: Improving balanced random forests. *Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT 2008): Computational Linguistics – Applications (CLA'08)*: 353–357. Wisła.

Del Gaudio, R., in Branco, A. (2007): Automatic extraction of definitions in Portuguese: A rule-based approach. *Proceedings of the 13th Portuguese Conference on Artificial Intelligence (EPIA2007)*. LNAI 4874: 659–670. Berlin Heidelberg: Springer.

Del Gaudio, R., Batista, G., in Branco, A. (2013): Coping with highly imbalanced datasets: A case study with definition extraction in a multilingual setting. *Natural Language Engineering*, *FirstView*: 1–33.

Erjavec, T. (2011): Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. *Proceedings of the ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (ACL 2011)*: 33–38. Portland.

Fahmi, I., in Bouma, G. (2006): Learning to identify definitions using syntactic features. *Proceedings of the EACL workshop on Learning Structured Information in Natural Language Applications*: 64–71. Trento.

Faralli, S., in Navigli, R. (2013): A Java Framework for Multilingual Definition and Hypernym Extraction. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*: 103–108. Sofia.

Fellbaum, C. (1998): *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fišer, D., Ljubešić, N., Vintar, Š., in Pollak, S. (2011): Building and using comparable corpora for domain-specific bilingual lexicon extraction. *Proceedings of the 4th BUCC Workshop: Comparable Corpora and the Web*: 19–26. Portland.

Fišer, D., Pollak, S., in Vintar, Š. (2010): Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*: 2932–2936. Valletta.

Fišer, D., in Sagot, B. (2008): Combining Multiple Resources to Build Reliable Wordnets. *Proceedings of the 11th Text, Speech and Dialogue International Conference (TSD 2008), Brno (LNCS 5246)*: 61–68. Berlin Heidelberg.

Frantzi, K. T., in Ananiadou, S. (1999): The CValue/NCValue domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, *6* (3): 145–179.

Gantar, P., in Krek, S. (2009): Drugačen pogled na slovarske definicije: opisati, pojasniti, razložiti? V M. Stabej (ur.): *Infrastruktura slovenščine in slovenistike*: 151–159. Ljubljana: Znanstvena založba Filozofske fakultete.

Geeraerts, D. (2003): Meaning and definition. V P. van Sterkenburg (ur.): *A Practical Guide to Lexicography*: 83–93. Amsterdam, Philadelphia: John Benjamins Publishing.

Grčar, M., Krek, S., in Dobrovoljc, K. (2012): Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. *Proceedings of the 8th Language Technologies Conference*: 89–94. Ljubljana: Institut Jožef Stefan.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th conference on computational linguistics (COLING '92),* 2: 539–545. Nantes.

Hurley, P. J. (2012). *A concise introduction to logic*. Boston, MA: Wadsworth Cengage Learning.

Iftene, A., Trandaba, D., in Pistol, I. (2007): Natural language processing and knowledge representation for e-learning environments. *Proceedings of the RANLP 2007 workshop on Applications for Romanian*: 19–25. Iasi.

iSlovar. K. Puc (ur.). Dostopno prek: http://www.islovar.org (23. februar 2014).

Jackson, H. (2002): *Lexicography: An Introduction*. London, New York: Routledge.

Kobyliński, Ł., in Przepiórkowski, A. (2008): Definition Extraction with Balanced Random Forests. *Proceedings of the 6th International Conference on Natural Language Processing (GoTAL 2008)*. V B. Nordström in A. Ranta (ur.): *Advances in Natural Language Processing (LNCS 5221)*: 237–247. Berlin Heidelberg: Springer.

Kosem, I. (2006): Definicijski jezik v slovarju slovenskega knjižnega jezika s stališča sodobnih leksikografskih načel. *Jezik in slovstvo*, 51 (5): 25–45.

Kozakov, L., Park, Y., Fin, T., Drissi, Y., Doganata, Y., in Cofino, T. (2004): Glossary extraction and utilization in the information search and

delivery system for IBM technical support. *IBM Systems Journal*, 43 (3): 546–563.

Kranjc, J., Podpečan, V., in Lavrač, N. (2012): ClowdFlows: A Cloud Based Scientific Workflow Platform. *Proceedings of ECML/PKDD-2012 (2) (LNCS 7524)*: 816–819. Berlin Heidelberg.

Krek, S. (2004): Slovarji serije COBUILD in formalizacija definicijskega jezika. *Jezik in slovstvo*, 49 (2): 3–16.

Lefever, E., Macken, L., in Hoste, V. (2009): Language-independent bilingual terminology extraction from a multilingual parallel corpus. *Proceedings of the 12th Conference of the European Chapter of the ACL*: 496–504. Atene.

L'Homme, M.-C., in Marshman, E. (2006): Extracting Terminological Relationships from Specialized Corpora. V L. Bowker (ur.): *Lexicography, Terminology, Translation: Text-Based Studies in Honour of Ingrid Meyer*: 67–80. Ottawa: University of Ottawa Press.

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., in Krek, S. (2012): *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba.* Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Macken, L., Lefever, E., in Hoste, V. (2013): TExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-based Alignment. *Terminology*, 19 (1): 1–30.

Malaisé, V., Zweigenbaum, P., in Bachimont, B. (2004): Detecting Semantic Relations Between Terms In Definitions. *Proceedings of the 3rd International Workshop on Computational Terminology CompuTerm 2004* at *COLING 2004*: 55–62. Geneva.

Meyer, I. (2001): Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. V D. Bourigault, C.

Jacquemin in M.-C. L'Homme (ur.): *Recent Advances in Computational Terminology*: 279–302. Orsay: Université de Montréal.

Muresan, S., in Klavans, J. (2002): A Method for Automatically Building and Evaluating Dictionary Resources. *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*: 231–234. Las Palmas.

Navigli, R., in Velardi, P. (2010): Learning Word-Class Lattices for Definition and Hypernym Extraction. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*: 1318–1327. Upssala.

Parry, W. T., in Hacker, E. A. (1991): *Aristotelian Logic*. Albany, NY: State University, New York Press.

Pollak, S., Vavpetič, A., Kranjc, J., Lavrač, N., in Vintar, Š. (2012a): NLP workflow for on-line definition extraction from English and Slovene text corpora. *Proceedings of the 11th Conference on Natural Language Processing (KONVENS 2012)*: 53–60. Dunaj.

Pollak, S., Trdin, N., Vavpetič, A., in Erjavec, T. (2012b): NLP Web Services for Slovene and English: Morphosyntactic Tagging, Lemmatisation and Definition Extraction. *Informatica*, 36: 441–449.

Pollak, S. (2014): *Polavtomatsko modeliranje področnega znanja iz večjezičnih korpusov: Doktorska disertacija*. Ljubljana: Filozofska fakulteta.

Przepiórkowski, A. (2007): Slavonic Information Extraction and Partial Parsing. *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing organized in collocation with the 45th Annual Meeting of the Association of Computational Linguistic*: 1–10. Praga.

Przepiórkowski, A., Degórski, L., Spousta, M., Simov, K., Osenova, P., Lemnitzer, L., Kuboň, V., in Wójtowicz, B. (2007): Towards the

automatic extraction of definitions in Slavic. *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing organized in collocation with the 45th Annual Meeting of the Association of Computational Linguistic*: 43–50. Praga.

Randolph, J. J. (2008): *Online Kappa Calculator*. Dostopno prek: justusrandolph.net/kappa/ (2. september 2013).

Reiplinger, M., Schäfer, U., in Wolska, M. (2012): Extracting glossary sentences from scholarly articles: A comparative evaluation of pattern bootstrapping and deep analysis. *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*: 55–65. Jeju Island.

Rey, S. (2000): Defining definition. V J. C. Sager (ur.): *Essays on Definition*. Philadephia: John Benjamins Publishing.

Robinson, R. (1972): *Definitions*. Oxford: Oxford University Press.

Schmid, H. (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*: 44–49. Manchester.

Sclano, F., in Velardi, P. (2007): TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities. *Proceedings of the 9th Conference on Terminology and Artificial Intelligence (TIA 2007)*: 8–9.

SDL Multiterm. Dostopno prek: http://www.translationzone.com/products/sdl-multiterm/desktop/ (1. februar 2014).

Sinclair, J. (1987): *Collins COBUILD English language dictionary*. London, Glasgow: Collins ELT.

Storrer, A., in Wellinghoff, S. (2006): Automated detection and annotation of term definitions in German text corpora. *Proceedings of the 5h*

International Language Resources and Evaluation Conference (LREC 2006): 2373–2376. Genoa.

Svensen, B. (1993): *Practical Lexicography: Principles and Methods of Dictionary Making*. Oxford: Oxford University Press.

Termania. Dostopno prek: http://www.amebis.si/termania (1. februar 2014).

Uršič, M., in Markič, O. (1997): *Osnove logike*. Ljubljana: Filozofska fakulteta.

Velardi, P., Navigli, R., in D'Amadio, P. (2008): Mining the Web to Create Specialized Glossaries. *IEEE Intelligent Systems*, 23 (5): 18–25.

Vidovič Muha, A. (2000): *Slovensko leksikalno pomenoslovje: Govorica slovarja*. Ljubljana: Znanstveni inštitut Filozofske Fakultete.

Vintar, Š. (2010): Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16 (2): 141–158.

Walter, S., in Pinkal, M. (2006): Automatic extraction of definitions from German court decisions. *Proceedings of the ACL'06 Workshop on Information Extraction beyond the Document*: 20–26. Sydney.

Westerhout, E. (2010): *Definition extraction for glossary creation: a study on extracting definitions for semi-automatic glossary creation in Dutch: Dissertation*. Utrecht: Universiteit Utrecht.

WordNet. Dostopno prek: http://wordnet.princeton.edu (20. februar 2014).

Zgusta, L. (1971): *Manual of lexicography*. Berlin, New York: De Gruyter Mouton.

Žagar Karer, M. (2011): *Terminologija med slovarjem in besedilom*. Ljubljana: Založba ZRC, ZRC SAZU.

EXTRACTING DEFINITION CANDIDATES FROM SPECIALIZED CORPORA

Human knowledge is available in different forms, including domain texts, terminological dictionaries, encyclopaediae, and recently also in computer-understandable representations of domain knowledge, such as taxonomies and ontologies. Since manual domain modeling is costly and time-consuming, researchers in human language technologies have started developing methods and tools for semi-automatic extraction of domain-specific knowledge from unstructured texts, involving tasks, such as terminology extraction, definition extraction, semantic relations extraction, or semi-automatic ontology building.

This article presents a methodology for definition extraction from domain corpora, currently available for Slovene and English. Since most of the existing methods and tools are language specific and not developed for minor languages, the main contribution of the dissertation is the developed definition extraction methodology for Slovene. The proposed definition extraction methodology is based on three different approaches to extracting definition candidates. The first follows the traditional pattern-based approach, in which patterns are composed of lemmas and morphosyntactic descriptions; the second approach relies on pairs of domain terms extracted through automatic term extraction; the third approach exploits wordnet hypernym pairs. We propose an original combination of the three approaches.

The developed methodology was applied to a real-case problem of modeling the language technologies domain, for which we constructed a comparable Slovene-English corpus consisting of about two million tokens. We extracted more than 3,400 definition candidates, of which over 700 (approximately 480 for Slovene and 230 for English) were evaluated as definitions. The results are used as a basis for the *Language Technologies Glossary*.[17]

An additional contribution is the proposed domain-modeling pipeline—from corpus uploading and preprocessing to inspecting the extracted term and

[17] http://kt.ijs.si/senja_pollak/jt_glosar/

definition candidates—implemented as an online publicly available workflow, easy to use and suitable for translation, linguistic and especially terminological tasks. The developed workflow components can be easily integrated in other natural language processing workflows.

In future research, we will act in several directions. Since we have a modular workflow, an obvious step to take is to add or substitute the tools in current implementation with other state-of-the-art preprocessing tools. In addition, we foresee to continue the research in the following lines. First, machine learning methods will be used trying to improve the results, the methodology for automatic term-alignment from comparable corpora (Fišer et al. 2011) will be implemented in the workflow environment and tested on our corpus. Our major focus will be on performing new experiments on other corpora and explore the influence of text type on definition extraction, as well as in comparison of our approach with other systems. For the developed *Language Technologies Glossary,* once we finish with uploading and refining the results, we will think of including the collection into Termania[18] or I-Slovar[19] collections.

Keywords: definition extraction, online workflows, language technologies, natural language processing, automatization of terminographic processes

[18] http://www.amebis.si/termania
[19] http://www.islovar.org