Luščenje in analiza slovenskih in angleških definicij v Korpusu jezikovnih tehnologij

Senja Pollak

Abstract

The paper discusses various definition types from the perspective of automatic definition extraction from specialized text corpora. More precisely, we analyse definitions from the Slovene-English comparable corpus of language technologies. We present three different approaches for automatic definition extraction and focus on the analysis of various definition types. The traditional *genus et differentiae* definition type – in which the given term (*definiendum*) is defined by its hypernym (*genus*) and the differences specific for the definiendum compared to other members of the class of the *genus* – is far from being the only defining strategy encountered in our corpus. And even this definition category contains several subtypes. The other definition types that we address in this paper are relational definitions, functional definitions and extensional definitions. We focus on two aspects, on one hand we are especially interested in analysing the sentences from the perspective of automated extraction of definitions from a corpus, and on the other hand we comment on the differences between definitions in the Slovene and English part of the corpus. The differences can be on one hand attributed to Slovene and English grammatical categories that exist only in one of the two languages, such as the possibility of using case information (esp. nominative) in Slovene defining patterns, while in English definite and indefinite articles are frequent markers of defining sentences. On the other hand, we observe pragmatic differences, where the context must be taken into consideration. For example, Slovene terms are often followed by their English translation, since the field of language technologies has a longer tradition and more consolidated terminology in English. The presented methodology of automatic definition extraction can be beneficial for terminological and translation purposes.

Ključne besede: luščenje definicij, vrste definicij, jezikovne tehnologije, specializirani korpusi

1 UVOD

Za prevajalce so terminološki priročniki ključnega pomena. Ročna izdelava glosarjev in terminoloških slovarjev je časovno in finančno zahtevna investicija, zato so raziskovalci na področju jezikovnih tehnologij začeli preučevati možnosti samodejnega luščenja terminologije in definicij. V članku predstavimo metodologijo, s katero lahko uporabnik za poljubno področje interesa, za katerega ima na voljo specializirani korpus, pridobi terminološke in definicijske kandidate, ki so pomembni nosilci znanja tega področja.

Z definicijami se filozofi in leksikografi ukvarjajo že dolgo, vse od Platona in Aristotela 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 (za natančnejši pregled glej npr. Rey 2000). V leksikografski tradiciji se avtorji nanašajo predvsem na Aristotela in tradicijo logike z definicijami z nadpomenko in rodno razliko (tj. definicije tipa *genus et differentiae*). Definicijskih tipov pa je v resnici mnogo več.

Na področju jezikovnih tehnologij poskušamo z metodami procesiranja naravnega jezika s kombinacijo jezikoslovnega in računalniškega znanja iz korpusov avtomatsko izluščiti znanje, v našem primeru kandidate za definicije. Podobno nalogo so si že zastavili za različne druge jezike: metodo luščenja iz angleških korpusov predlagajo npr. Navigli in Velardi (2010), nizozemske definicije obravnava Westerhout (2010), francoske Malaisé et al. (2004), nemške Fahmi in Bouma (2006) itd. Luščenje podatkov iz slovanskih jezikov velja za težjo nalogo, saj so to morfološko bogati jeziki z relativno prostim besednim redom (Przepiórkowski 2007). Pristopi k luščenju iz poljščine, češčine in bolgarščine so predstavljeni v Przepiórkowski et al. (2007), za poljščino pa so nadaljevanje predstavili v Degórski et al. (2008).

Pričujoči članek uporabi metode luščenja definicij iz specializiranih korpusov za slovenščino in angleščino, ki smo jih že predstavili v Pollak et al. (2012) in Pollak (2014). V tem članku manj pozornosti namenimo samim metodam in eksperimentom, posvetimo pa se analizi definicij iz Korpusa jezikovnih tehnologij; posebno pozornost namenimo razlikam v slovenskih in angleških strukturah definicij.

Struktura članka je naslednja: v drugem poglavju predstavimo obravnavani korpus, v katerem smo zajeli članke s področja jezikovnih tehnologij, v tretjem poglavju na kratko predstavimo metode luščenja definicij, v četrtem, osrednjem poglavju pa se posvetimo analizi slovenskih in angleških definicij. Članek sklenemo z zaključki in načrti za nadaljnje delo.[1]

1 Poglavitni del predstavljene raziskave je bil izveden v okviru doktorske naloge pod mentorstvom dr. Špele Vintar, ki se ji najlepše zahvaljujem. Hvala tudi J. Sterle in Ž. Malovrh za pomoč pri gradnji korpusa in osnutka glosarja.

2 KORPUS JEZIKOVNIH TEHNOLOGIJ

Korpus jezikovnih tehnologij je zgrajen kot primerljivi korpus v slovenskem in angleškem jeziku. Korpus sestavljajo predvsem znanstvena besedila s področja jezikovnih tehnologij, nekaj pa je tudi bolj poljudnih tekstov. Korpus vsebuje 1.089.968 slovenskih in 1.073.470 angleških pojavnic. Največji del korpusa sestavljajo članki zbornikov konference *Jezikovne tehnologije*, poleg njih pa korpus vključuje tudi tematske članke drugih revij in konferenc. Dodali smo tudi diplomske, magistrske in doktorske naloge ter nekatera poglavja iz knjig ter nekatere članke iz Wikipedije.

Dokumente smo najprej poenotili v osnovno tekstovno obliko (kodiranje v UTF8), nato smo uporabili orodje ToTrTaLe (Erjavec 2011), s katerim smo besedilo razčlenili na stavke in besede, besedam pripisali leme ter oblikoskladenjske oznake, ki vsebujejo informacije o besedni vrsti in obliki (npr. sklon, oseba, število itd.).

3 LUŠČENJE DEFINICIJSKIH KANDIDATOV

Za avtomatsko luščenje definicij predlagamo tri pristope, ki jih med seboj lahko tudi kombiniramo. Metode so podrobno predstavljene v Pollak (2014).

3.1 Metode luščenja

Prva metoda uporablja *leksikoskladenjske vzorce.* Definiramo vzorce, ki jih sestavljajo leme, besedne oblike ter oblikoskladenjske oznake, kot so skloni samostalnikov, glagolske osebe itd., nato pa v korpusu izluščimo stavke, ki tem vzorcem ustrezajo. Eden najbolj osnovnih vzorcev je tako »samostalniška besedna zveza v imenovalniku + *je/sta/so* + sam. b. zv. v imenovalniku«. Vendar pa uporabljamo veliko širši nabor (ročno določenih) vzorcev.

Naslednja metoda luščenja kandidatov za definicije uporablja *zaznavo terminov* v stavkih. Prilagodili smo luščilnik terminov LUIZ (Vintar 2010), ki na podlagi oblikoskladenjskih vzorcev in izračuna terminološkosti na podlagi relativne frekvence kandidatov glede na referenčni korpus predlaga eno- in večbesedne terminološke izraze. Terminov oz. natančneje rečeno terminoloških kandidatov ročno ne pregledamo, tako da je termin v našem primeru definiran kot avtomatsko izluščena samostalniška besedna zveza z orodjem LUIZ. Minimalni pogoj je, da stavek vsebuje dva termina, dodatni poljubni pogoji pa so: glagol med dvema terminoma, samostalnik v imenovalniku (za slovenščino), termin na začetku stavka in podobno. Seveda niso vsi stavki, ki vsebujejo najmanj dva termina,

definicije, so pa to pogosto pomensko bogati konteksti (angl. *knowledge-rich contexts*, Meyer 2001).

S tretjo metodo luščimo stavke s *pojmi v hierarhičnem odnosu.* Za izbor stavkov, ki vsebujejo dva izraza, od katerih je eden direktna nadpomenka drugega, smo uporabili semantični leksikon WordNet (Fellbaum 1998) za angleščino ter sloWNet (Fišer in Sagot 2008) za slovenščino.

Zgornje tri metode lahko med sabo poljubno kombiniramo, pri vsaki poljubno nastavimo parametre in z njimi iščemo stavke, ki so izluščeni z vsaj eno od treh metod, stavke na presečišču dveh ali treh metod ali pa uporabimo dodatne kombinacije z različnimi nastavitvami parametrov posameznih metod.

Metodologijo za luščenje terminologije in definicij smo implementirali kot prosto dostopen delotok, dostopen na naslovu http://clowdflows.org/workflow/1380.

3.2 Rezultati luščenja

V pričujočem članku se posvečamo predvsem analizi definicij, manj pozornosti pa namenjamo eksperimentom in kvantitativnim rezultatom. Za podrobnejše rezultate posameznih metod in njihovih kombinacij glej Pollak (2014), na tem mestu pa povzamemo, da smo z unijo kandidatov, izluščenih z vsaj eno od treh metod, izluščili 6606 kandidatov za slovenske ter 4727 za angleške definicije, od katerih smo jih ocenili kot definicije 646 za slovenščino in 344 za angleščino. Veliko je tudi mejnih primerov. Z različnimi kombinacijami metod lahko dosežemo večjo natančnost (večji odstotek definicij med izluščenimi definicijskimi kandidati). S presekom vsaj dveh metod je natančnost nad 25 odstotkov, vendar smo tako izluščili le 129 slovenskih in 82 angleških definicij, s kompleksnejšimi kombinacijami pa smo izluščili 389 slovenskih in 230 angleških definicij z natančnostjo nad 20 %.

4 ANALIZA SLOVENSKIH IN ANGLEŠKIH DEFINICIJ

Ko govorimo o luščenju definicijskih kandidatov iz korpusov, se je treba zavedati, da definicije nimajo enotne strukture, temveč je definicijskih tipov mnogo. Sama definicija definicije ni nekaj univerzalnega in samoumevnega in definicijo sámo je smiselno definirati glede na končno aplikacijo. Pri avtomatizaciji iskanja definicijskih kontekstov iz korpusov, ki jih obdelujemo zato, da bi prevajalcem ali stroki pomagali s polavtomatsko izdelavo glosarjev,[2] se je pomembno zavedati, da so v

2 Glosar razumemo kot najpreprostejšo (začetno) obliko terminološkega slovarja, ki vsebuje termine in njihove definicije oz. razlage.

njih termini definirani v tekočem besedilu. Ne gre izključno za iskanje klasičnih slovarskih definicij, temveč za iskanje primernih kandidatov, ki jih lahko naknadno še ročno prečistimo in dopolnimo pred vključitvijo v zbirko.

Pojem, ki ga definiramo, se imenuje *definiendum*, del definicije, ki definira njegov pomen, je *definiens,* oba dela pa sta lahko povezana z *zglobom* (angl. *hinge*). Glavna razlika glede načina definiranja definienda je že pri Aristotelu postavljena med *intenzionalnimi* in *ekstenzionalnimi* definicijami. 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 definienda, kar pomeni, da navajajo vse možne oz. najbolj tipične realizacije definiranega pojma (gre torej za naštevanje vseh ali tipičnih pripadajočih elementov razreda) (Copi in Cohen 2009, Svensen 1993, Zgusta 1971, Geeraerts 2003). V slovenščini so različne tipe definicij obravnavali npr. Žagar (2011), Krek (2004), Kosem (2006) ter Gantar in Krek (2009).

V nadaljevanju obravnavamo različne tipe intenzionalnih in ekstenzionalnih definicij, ki jih najdemo v slovenskem in angleškem delu *Korpusa jezikovnih tehnologij.* Nekateri primeri izhajajo iz študije, ki smo jo izvedli za definiranje leksikoskladenjskih vzorcev, drugi pa iz analize izluščenih definicij s pomočjo treh navedenih metod. Kategorizacija na posamezne podtipe pa ni pri vseh avtorjih enaka (glej npr. Borsodi 1967, Robinson 1972, Jackson 2002, Westerhout 2010, Kosem 2006, Geeraerts 2003), tako da naredimo svoj podizbor kategorij, glede na prevladujoče tipe v korpusu.

4.1 Definicije tipa *genus et differentiae*

Najbolj znan tip intenzionalnih definicij (in tudi definicij nasploh) je analitični tip definicij z obliko *genus et differentiae*. *Definiendum* je v njih definiran z nadpomenko oz. najbližjim rodom (*genus proximum*) in vrstnimi razlikami (*differentiae specificae*) oz. vsaj eno bistveno značilnostjo, ki definiendum (oz. razred definienda) ločuje od ostalih pripadnikov rodu (npr. Svensen 1993, Béjoint 2000). V leksikografiji ni nujno, da nadpomenka predstavlja prav najbližji rod, temveč je lahko tudi bolj oddaljen pojem v taksonomiji. V nadaljevanju definicije ločimo glede na zglob, tj. glagol, s katerim povezujemo oba dela definicije.

4.1.1 Definicije genus et differentiae z glagolom biti

Splošna struktura definicije te kategorije je »sam. bes. zv. *je* sam. bes. zv. ...«. Samostalniške besedne zveze so lahko vse od samega samostalnika (npr. *termin*), pa

vse do bolj kompleksnih struktur, npr. pridevnik + samostalnik + predlog + pridevnik + samostalnik ter veliko vmesnih variant (npr. *sistem za luščenje definicij*).

V angleščini definicije pogosto uporabljajo določni ali nedoločni člen *a/an* ali *the*, kar je v veliko pomoč pri avtomatskem luščenju na podlagi vzorcev. Pomožni glagol *be* pa je lahko v ednini ali množini. Primera angleških definicij iz korpusa podamo spodaj.

i. *A syntactic parser is a tool that gives the structural composition of a sentence in the form of a tree.*
ii. »*Resource-poor*« *languages are languages for which few digital resources exist; and thus, languages whose computerization poses unique challenges.*

Na prvem primeru lahko podrobneje razložimo splošno strukturo definicij tipa *genus* et *differentiae.* Definiendu *syntactic parser* sledi zglob *is*, definiens pa sestavljajo nadpomenka ter razlikovalne značilnosti. Nadpomenka ni najbližji rod, temveč bolj splošna beseda *tool*, razlikovalne značilnosti (rodne razlike), ki ločijo *skladenjski razčlenjevalnik* od drugih orodij, pa so podane v nadaljevanju kot funkcijska definicija (glej poglavje 4.3).

V slovenščini se glagol *biti* lahko pojavlja v ednini, dvojini (redko) ali množini, pri definiranju vzorcev pa nam pomaga tudi informacija o sklonu, saj sta obe samostalniški besedni zvezi v imenovalniku. Primer definicije tega tipa je:

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

Lombardov efekt je termin v imenovalniku, sledi glagol *biti* v 3. osebi ednine, genus oz. nadpomenka je beseda *pojav* (tudi v imenovalniku), sledi pa del, odvisnik, ki razlikuje *lombardov efekt* od drugih pojavov (rodna razlika).

V slovenščini, ki ima terminologijo obravnavanega področja manj ustaljeno, se za terminom pogosto pojavlja angleška različica termina, ki je velikokrat eksplicitno uvedena s kratico, npr:

iv. *Branje ustnic (angl. lipreading) je vizualna percepcija govora, ki temelji izključno na opazovanju artikulatornih gibov (ustnic) govorca brez poslušanja.*
v. *Avtomatsko oblikoslovno označevanje (part-of-speech tagging oz. word-class syntactic tagging) je postopek, pri katerem se vsaki besedi, ki se v besedilu pojavi, pripiše oblikoslovna oznaka.*

V obeh jezikih pa so lahko po terminu naštete tudi kratice, sinonimi ali variacije termina.

vi. *In corpus linguistics, part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part*

of speech, based on both its definition, as well as its context—i.e. relationship with adjacent and related words in a phrase, sentence, or paragraph.

Genus je včasih uveden z bolj splošnimi besedami, kot so *vrsta, beseda, izraz, tip*, kar velja za slovenščino (glej stavek vii) in angleščino (npr. stavek viii).

vii. *Besedna poravnava (Word Alignment) je izraz za tehnologijo statističnega pridobivanja leksikonov prevodnih ustreznic iz vzporednih korpusov.*

viii. *Translation memories are a type of computer-assisted translation (CAT) tool, one that enjoys huge popularity among professional translators, as they ensure that no sentence ever needs to be translated twice.*

V slovenščini pred nadpomenko občasno najdemo tudi kazalni zaimek:

ix. *Osnova je tisti del besede, ki ima predmetni pomen, končnica pa tisti, ki zaznamuje slovnične lastnosti besede.*

x. *Ujemanje je »/t/ista vrsta slovnične, skladenjske vezi med besedami samostalniške besedne zveze oz. med stavčnimi členi, ko se odvisna beseda (ali stavčni člen) v sklonu, številu, osebi ali tudi spolu ravna po svojem nadrejenem delu /…/«.*

Definiendum se ne nahaja vedno na začetku stavka. Pogosto je domet definicije določen s področjem ali avtorji:

xi. *According to ISO 9126 software standards ([EAGLES96]) usability is a quality characteristic that is composed of three subcategories: • understandability • learnability • operability /.../*

V angleščini uvodni del manj vpliva na samo strukturo definicije, v slovenščini pa se besedni red bistveno spremeni. V primerih, kakršen je (xii), glagolu neposredno sledita definiendum in genus, kar je za avtomatsko procesiranje lahko težavno.

xii. *Kot podatkovne strukture so semantične mreže usmerjeni grafi, v katerih so pojmi predstavljeni s točkami oz. vozlišči, razmerja pa s puščicami oz. povezavami med njimi.*

xiii. *Tako v filozofiji kot v računalništvu je ontologija po definiciji predstavitev entitet, idej in dogodkov, skupaj z njihovimi lastnostmi in medsebojnimi razmerji – glede na izbrani sistem kategorij.*

V obeh jezikih se v znanstvenih besedilih pogosto pojavlja tudi referenca:

xiv. *Accordingly, a topic ontology (Fortuna, 2007) is a hierarchical organization of documents' topics and their sub-topics.*

Poseben podtip definicij *genus et differentiae* so *razvrstitvene definicije,* ki podajajo le nadpomenko, ne pa vrstnih razlik (glej primer xv). Taka definicija ni popolna, je pa vseeno uporabna kot osnutek prave definicije.

xv. *Vsebinsko rudarjenje je veja spletnega rudarjenja.*

4.1.2 Definicije genus et differentiae z drugimi glagoli

Pri definiranju se uporablja tudi veliko drugih definicijskih glagolov. V slovenščini so to npr. *definirati, imenovati (se), opredeliti, predstavljati,* v angleščini pa na primer *define, denote, mean, refer.* Podobno kot v prejšnji kategoriji tudi tu pri tej kategoriji veljajo razne variacije (kratice, reference, omejitev veljave definicije in podobno).

Spodnji primer angleške definicije (xvi) uporablja glagol *denote*, definicija ima določeno polje veljave (*in computer science*), vendar pa osnovna struktura (*definiendum,* ki ga zglob povezuje z nadpomenko in ostalim delom *definiensa*) ostaja enaka. Podobno velja za definicijo v primeru (xvii), ki uporablja glagol *refer to.*

xvi. *In computer science the term ontology denotes a formal representation of a set of concepts of a domain and the relationships among these concepts.*

xvii. *Document summarization refers to the task of creating document surrogates that are smaller in size but retain various characteristics of the original document, depending on the intended use.*

V slovenščini se v definicijah uporabljajo tudi povratni glagoli, kot sta *imenovati se* (glej primer xviii) ali *nanašati se (na)* (primer xix). Oba primera odsevata relativno prost besedni red, ki kaže tudi na to, da morajo biti vzorci za luščenje dovolj prožni, da ne opredelijo celotne strukture stavka, temveč le dele, pomembne za prepoznavo definicij.

xviii. *Znanstvenokritične izdaje se v literarnih vedah imenujejo tiste edicije, v katerih so besedila pregledana, prepisana, rekonstruirana, komentirana in naposled objavljena po načelih tekstne kritike ali ekdotike kot pomožne literarnovedne discipline.*

xix. *V skladu z jezikoslovno tradicijo se nanaša pojem simbolične prozodije na govorne značilnosti, ki se ne nanašajo na en sam fonetični segment, glas, temveč na večje enote, ki vključujejo več fonetičnih segmentov, kot so besede, fraze, stavki ali celo večji odseki govorjenega besedila.*

V korpusu so pogoste tudi trpne oblike definicij. V angleščini je v nekaj primerih vršilec dejanja izražen in uveden s strukturo *by* (*defined by* v primeru xx), v slovenščini pa se v teh primerih pogosteje uporablja tvornik (primer xxi).

xx. *Translation memory (TM) is defined by the Expert Advisory Group on Language Engineering Standards (EAGLES) Evaluation Working Group's document on the evaluation of natural language processing systems as »a multilingual text archive containing (segmented, aligned, parsed and classified) multilingual texts, allowing storage and retrieval of aligned multilingual text segments /.../«.*

xxi. *Mona Baker v enciklopediji prevodoslovja [2] definira lokalizacijo kot proces ustvarjanja ali prilagoditve določenega izdelka specifičnemu okolju, na primer jezikovnemu okolju, kulturnemu kontekstu, dogovorom ali pogojem ciljnega trga.*

Pogosto pa v trpnih strukturah vršilec dejanja ni izražen (primer v xxii). Avtor definicije oz. polje veljavnosti je lahko tudi izraženo s prislovnim določilom ali podano kot referenca v oklepaju (glej xxiii).

xxii. *Eager methods are defined as methods which seek to create an abstract, generalised, model of the data.*

xxiii. *V klasični teoriji (Katz in Fodor 1963) so pomeni besed predstavljeni kot množice potrebnih in zadostnih pogojev, ki zajemajo pojmovno vsebino, izraženo z besedami.*

Strukture, za katere ne moremo reči, da so izključno definicijski glagoli, so pa pri definiranju pogosto uporabljene, so *biti znan pod imenom, veljati za, nanašati se, govoriti o*.

xxiv. *Tradicionalno velja ontologija za vejo filozofije in je bila dolgo znana pod imenom metafizika, ukvarja se z vprašanji t.i. entitet, ki obstajajo ali veljajo za obstoječe; kako se te entitete združujejo v večje razrede; kako so znotraj njih razdeljene hierarhično v smislu podobnosti in razlik.*

xxv. *Kadar gre za dvoumnost, pri kateri so različni pomeni besede med seboj povezani, govorimo o polisemiji ali večpomenskosti (npr. miška, ki je lahko del računalnika ali glodavec).*

4.1.3 Definicije genus et differentiae v odvisnih stavkih

Povezava med *definiendom* in *definiensom* v nekaterih primerih ni vzpostavljena z definicijskim glagolom, temveč brez njega, npr. z vrinjenim stavkom (glej primer xxvi), pristavkom (primer xxvii), v nekaterih primerih pa je podana v oklepajih (glej primer xxviii). Definicij tega tipa nismo izluščili z uporabo vzorcev, temveč z metodama luščenja s termini in s semantičnim leksikonom.

xxvi. *Pri korpusih usvajanja tujega jezika sta pomembna ciljni jezik, torej jezik, »ki se ga nekdo uči z namenom, da bi ga obvladal bodisi kot svoj prvi, drugi ali tuji jezik« (Pirih Svetina 2005), in izhodiščni oziroma prvi jezik, »iz katerega se nekdo uči vse druge ali tuje jezike« (navedeno delo).*

xxvii. *V središču vsake semantične zbirke, pa tudi pričujoče raziskave, so leksikalne enote oziroma leksemi, osnovni gradniki pomena v jeziku.*

xxviii. *Pri tem pristopu naletimo na posebnosti, ki jih lahko razdelimo v dve skupini: leksikalne vrzeli (pojem, ki je v nekem jeziku izražen z leksikalno enoto, je v drugem mogoče izraziti samo s prosto kombinacijo besed) in denotacijske razlike (v ciljnem jeziku obstaja prevodna ustreznica pojma izvornega jezika, vendar je nekoliko splošnejša ali nekoliko bolj specifična).*

4.2 Relacijske definicije

Analitične definicije tipa *genus et differentia* niso edini način definiranja. Zelo pogosto pri definiranju uporabljamo sinonime pa tudi antonime ali sorodne termine. Kadar je podan le sinonim, to ni polna definicija, vendar je tovrstna definicija kljub temu zanimiva alternativa analitični definiciji. Sinonimi oz. sorodni pojmi pa morajo biti splošno razumljeni ali pa bolj natančno definirani znotraj iste zbirke. Pogosto so relacijske definicije podane znotraj daljših, nedefinicijskih stavkov ali kot dodatek k drugim tipom definicij.

4.2.1 Sinonimi

Sinonimi so uvedeni z glagoli (primer xxix), navedeni v oklepajih ali uvedeni s tipičnimi izrazi, kot sta *also known* ali *also called* (primer xxx). Drugi primer je zelo pogost, saj sinonim oz. alternativno poimenovanje pojma služi le kot dodatek bolj polni definiciji. V slovenščini se bolj pogosto kot slovenski sinonimi uporabljajo angleški prevodi.

xxix. *Enopojavnice v korpusnem jezikoslovju imenujemo tudi hapax legomena in predstavljajo posebej zanimivo področje raziskovanja..*

xxx. *In other words, translation memory (also known as sentence memory) consists of a database that stores source and target language pairs of text segments that can be retrieved for use with present texts and texts to be translated in the future.*

4.2.2 Definicije s sorodnimi koncepti

Sorodni koncepti se pogosto pojavljajo v kombinaciji z drugimi tipi definicij. Westerhout (2010) po Borsodi (1967) imenuje to kategorijo *definicije z analogijo* in jih razume kot podtip definicij s sinonimi. Sorodni koncept mora biti manj specifičen od definienda. V spodnjih primerih vidimo, da sorodnemu konceptu sledi tudi druga definicija. V primeru (xxxii) pa so *klasični slovarji* podani kot sorodni termin, vendar je rodna razlika podana kot funkcijska definicija:

xxxi. *Wikislovar je sorodni projekt Wikipedije in je prost večjezični slovar z definicijami, izvorom besed, naglaševanjem in navedki.*

xxxii. *Za razliko od klasičnih slovarjev semantične zbirke pomen besede definirajo glede na to, kako je ta povezan s pomeni drugih besed.*

V angleščini smo opazili relativno pogost izraz *is a variant of*:

xxxiii. *Local beam search is a variant of the hill-climbing algorithm.*

xxxiv. *Text mining (Feldman and Sanger, 2007) is a variant of data mining in which models and patterns are extracted from unstructured natural language text.*

4.2.3 Antonimi

Spodnji primer namesto sinonima poda inverzno relacijo, vendar je tovrstna definicija problematična.

xxxv. *Najpogostejša relacija je hipernimija, s tem pa tudi njena inverzna relacija hiponimija.*

Težava, ki se pojavlja ob relacijskih definicijah (tudi v zgoraj navedenem primeru), je problem krožnosti, imenovane tudi *circulus in definiendo.* To je pojav, v katerem se znotraj iste definicije en termin definira z drugim. Četudi se tej pogosti leksikografski napaki zlahka izognemo na ravni povedi, pa problem pogosto ostaja na ravni zbirke.

4.3 Funkcijske definicije

Skoraj najpogosteje v našem korpusu opazimo *funkcijske definicije,* v katerih je pojem definiran z navedbo tipične uporabe, namena oz. funkcije definienda. Ta strategija se lahko uporablja znotraj kategorije definicij *genus et differentiae,* v katerih je termin definiran z nadpomenko in je njegova uporaba navedena kot del *differentiae* ali pa je tovrstna vrsta definiranja samostojna kot v spodaj navedenih korpusnih primerih:

xxxvi. *Leksikalna semantika se ukvarja s pomenom besed in proučuje različne vidike besednega pomena, ki se realizirajo v tipični (pa tudi netipični) rabi v slovnično ustreznih kontekstih.*

xxxvii. *A trie (also known as retrieval tree or prefix tree) provides a compact representation of strings with shared prefixes, which is exactly what is needed.*

xxxviii. *Translation memory helps the translation process by recognising previously translated texts: the system »keeps« sentences that have been previously translated, with their corresponding translation.*

Funkcijske definicije smo v glavnem izluščili z metodo z zaznavo terminov, manjši delež funkcijskih definicij pa lahko prepoznamo tudi z vzorci, kot sta »*naloga X je...*« oz. »*role of X is...*«:

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

xl. *The role of the language model is to provide the decoder with the possible phone sequences, along with their corresponding probabilities.*

Podobna kategorija kot funkcijske definicije je definiranje s tipičnimi lastnostmi definienda:

xli. *Za protipomenke je značilno, da imajo skupnih večino element* (sic!) *pomena, s to razliko, da zavzemajo skrajne vrednosti neke dimenzije (npr. vroče ↔ mrzlo).*

V slovenščini so v stavkih, ki vključujejo definicije, večkrat uporabljeni modalni glagoli oz. prislovi (*morati, je mogoče, lahko*, *verjetno*), kar je po prvi analizi v angleškem delu korpusa izredno redko. Izražanje modalnosti bi vsekakor zahtevalo podrobnejšo analizo.

xlii. *Programi za oblikoslovno označevanje morajo poljubnim besednim oblikam določiti možne oznake, nato pa izmed teh oznak izbrati pravo glede na kontekst, v katerem se besedna oblika pojavi.*

xliii. *Sintetizator govora lahko pretvori poljubno slovensko besedilo v razumljiv računalniški govor.*

Zadnja zgornja definicija je sicer preozka, saj lahko sintetizator govora v resnici pretvori katerokoli besedilo in ne le slovenskega. Tovrstni primeri so mejni primeri, saj nujno potrebujejo ročne popravke.

4.4 Ekstenzionalne definicije

Vsi do sedaj omenjeni tipi definicij sodijo med intenzionalne definicije, ki se osredotočajo na bistvene lastnosti, s katerimi je pojem definiran. Druga strategija za definiranje pojmov je z *ekstenzijo* oz. *obsegom* (množico stvari, na katere se pojem nanaša). Naštejemo lahko vse predstavnike definiranega razreda (*naštevalne definicije*) ali pa le najbolj reprezentativne. V spodnjem primeru so naštete različne *taksonomske relacije*, nekatere pa so tudi ponazorjene s primeri:

xliv. *Poleg nad- in podpomenskosti sta taksonomski razmerji tudi meronimija in holonimija, ki izražata odnos med delom in celoto (npr. volan ↔ avto), med glagoli pa troponimija, ki povezuje glagole glede na način izvajanja nekega dejanja (npr. govoriti ↔ šepetati) (Fellbaum 2002).*

Podobno so v spodnjem stavku v oklepaju navedeni različni tipi jezikovnih virov:

xlv. *Language resources (written and spoken corpora, lexicons, parsers, annotation tools, etc) are essential for the development of language technologies and for the training of students.*

5 ZAKLJUČKI IN NADALJNJE DELO

Glavni cilj prispevka je predstavitev razvite metodologije za luščenje definicijskih kandidatov ter analiza in primerjava rezultatov luščenja iz slovenskega in

angleškega dela *Korpusa jezikovnih tehnologij*. Metodologija je implementirana kot prosto dostopno orodje, ki uporabniku omogoča (pol)avtomatsko izluščiti področno znanje v obliki definicij iz nestrukturiranih besedil. Metoda je uporabna kot pomemben korak pri gradnji prevajalskih (terminoloških) priročnikov.

V analizi obravnavamo različne vrste definicijskih struktur, ki smo jih zaznali v korpusu. Najosnovnejša oblika definicije je analitična definicija z nadpomenko in rodnimi razlikami (definicija tipa *genus et differentiae).* Na podlagi korpusa pokažemo, da je v tekočem znanstvenem besedilu že ta kategorija zelo raznolika. S korpusnimi primeri ponazorimo tudi druge vrste definicij, kot so *sintetične* oz. *relacijske definicije,* ki izbrani termin definirajo preko sinonimov, sorodnih konceptov ali z drugimi semantičnimi relacijami, zelo pogoste pa so *funkcijske definicije*, v katerih je termin definiran z njegovo tipično funkcijo oz. rabo. Za razliko od intenzionalnih pa *ekstenzionalne definicije* uporabljajo drugo strategijo, saj pojem definirajo tako, da naštejejo vse ali tipične pripadnike razreda, ki ga pojem opisuje. Tudi ta definicijski tip ponazorimo s primeri iz korpusa.

Razlike med slovenskimi in angleškimi definicijami in vzorci za njihovo avtomatsko luščenje delno izhajajo iz razlike med jezikoma, delno pa gre za razlike na pragmatski ravni. V prvo kategorijo tako uvrščamo značilnosti, da se pri osnovnem vzorcu »sam. bes. zv. *je* sam. bes. zv.« v slovenščini pojavlja glagol *biti* v tretji osebi ednine, dvojine in množine, v angleščini pa seveda ni posebne oblike za dvojino. Pri določitvi vzorcev si v angleščini lahko pomagamo s členi (*a/an* in *the*), v slovenščini pa je pogosto, da sta obe samostalniški besedni zvezi v imenovalniku. V slovenščini se pred genusom občasno uporablja kazalni zaimek *tisti*. Naštejemo tudi številne primere z drugimi glagoli, ki so bolj ali manj specifični za definicije (*definirati, opredeliti, opisati, nanašati se, pomeniti, imenovati, poimenovati, govoriti o*) oz. v angleščini *define, refer to, denote* ipd. Razlike zaznamo v trpnih strukturah, kjer je v angleščini lahko avtor definicije podan kot vršilec uveden s predlogom *by*, česar v slovenkem delu nismo opazili. Določitev definicijskih vzorcev je v slovenščini težja zaradi relativno prostega besednega reda. Razlike, ki izhajajo iz širšega konteksta, pa so te, da slovenske definicije pogosto v oklepajih navajajo bolj uveljavljeno angleško terminologijo (na drugih področjih pa npr. latinsko ali grško). V angleškem delu korpusa na izbranem področju ne najdemo navedbe termina v drugih jezikih, pogosta pa so alternativna poimenovanja in sinonimi. Poleg tega smo v slovenskih definicijskih stavkih opazili tudi bolj pogosto izraženo modalnost.

Analizirani primeri so iz jezikoslovnega vidika zanimivi kot analiza definicijskih struktur in tipov, poleg tega pa so uporabni tudi za nadaljnje izboljšanje metodologije avtomatskega luščenja znanja na podlagi vzorcev ter za razumevanje možnosti in omejitev avtomatskih metod. V nadaljevanju bomo metodologijo preizkusili na drugih korpusih, analizirali vpliv funkcijskih zvrsti (definicije v

znanstvenih besedilih v primerjavi s poljudnoznanstvenimi definicijami), metodologijo pa bomo prilagodili tudi za francoščino.

Literatura

Béjoint, Henri, 2000: *Modern Lexicography: An Introduction*. Oxford: Oxford University Press.

Borsodi, Ralph, 1967: *The definition of definition. Boston: Porter Sargent Publisher.*

Copi, Irving. M. in Carl Cohen, 2009: *Introduction to Logic* (13th ed.). Upper Saddle River: Pearson/Prentice Hall.

Degórski, Lukasz, Łukasz Kobyliński in Adam Przepiórkowski, 2008: Definition extraction: Improving balanced random forests. Ganzha, Maria, Marcin Paprzycki in Tomasz Pełech-Pilichowski (ur.): *Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT 2008): Computational Linguistics – Applications (CLA'08)*: Wisła: PTI. 353–357.

Erjavec, Tomaž, 2011: Automatic Linguistic Annotation of Historical Language: ToTrTaLe and XIX Century Slovene. Zervanou, Kalliopi in Piroska Lendvai (ur.): *Proceedings of the ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (ACL 2011).* Portland: ACL. 33–38.

Fahmi, Ismail in Gosse Bouma, 2006: Learning to Identify Definitions Using Syntactic Features. Basili, Roberto in Alessandro Moschitti (ur.): *Proceedings of the EACL Workshop on Learning Structured Information in Natural Language Applications.* Trento: ACL. 64–71.

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

Fišer, Darja in Benoît Sagot, 2008: Combining Multiple Resources to Build Reliable Wordnets. Sojka, Petr, Aleš Horák, Ivan Kopeček in Karel Pala (ur.): *Proceedings of the 11th Text, Speech and Dialogue International Conference (TSD 2008).* LNCS 5246. Berlin/Heidelberg: Springer-Verlag. 61–68.

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

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

Jackson, Howard, 2002: *Lexicography: An Introduction.* London: Routledge.

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

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

Malaisé, Véronique, Pierre Zweigenbaum in Bruno Bachimont, 2004: Detecting semantic relations between terms in definitions. Ananadiou, Sophia in Pierre Zweigenbaum (ur.): *Proceedings of the: 3rd International Workshop on Computational Terminology CompuTerm 2004 at COLING 2004*. Geneva. 55–62.

Meyer, Ingrid, 2001: Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. Bourigault, Didier, Christian Jacquemin in Marie-Claude L'Homme (ur.): *Recent Advances in Computational Terminology.* Amsterdam/Philadelphia: John Benjamins Publishing. 279–302.

Navigli, Roberto in Paola Velardi, 2010: Learning Word-Class Lattices for Definition and Hypernym Extraction. Hajič, Jan, Sandra Carberry, Stephen Clark in Joakim Nivre (ur.): Proceedings *of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*. Stroudsburg: ACL. 1318–1327.

Pollak, Senja, Anže Vavpetič, Janez Kranjc, Nada Lavrač in Špela Vintar, 2012: NLP Workflow for On-line Definition Extraction from English and Slovene Text Corpora. Jancsary, Jeremy (ur.): *Proceedings of the 11th Conference on Natural Language Processing (KONVENS 2012)*. Vienna: ÖGAI. 53–60.

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

Przepiórkowski, Adam, 2007: Slavonic Information Extraction and Partial Parsing. Piskorski, Jakub, Hristo Tanev, Bruno Pouliquen in Ralf Steinberger (ur.): *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing* organized in collocation with the 45th Annual Meeting of the Association of Computational Linguistic. Stroudsburg: ACL. 1–10.

Przepiórkowski, Adam, Lukasz Degórski, Miroslav Spousta, Kiril Simov, Petya Osenova, Lothar Lemnitzer, Vladislav Kuboň in Beata Wójtowicz, 2007: Towards the Automatic Extraction of Definitions in Slavic. Piskorski, Jakub, Hristo Tanev, Bruno Pouliquen in Ralf Steinberger (ur.): *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing* organized in collocation with the 45th Annual Meeting of the Association of Computational Linguistic. Stroudsburg: ACL. 43–50.

Rey, Alain, 2000: Defining Definition. Sager, Juan C. (ur.): *Essays on Definition.* Amsterdam/Philadelphia: John Benjamins Publishing. 1–14.

Robinson, Richard, 1972: *Definitions.* Oxford: Oxford University Press.

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

Vintar, Špela, 2010: Bilingual Term Recognition Revisited. The Bag-of-Equivalents Term alignment approach and its evaluation. *Terminology* 16/2. 141–158.

Westerhout, Eline, 2010: *Definition Extraction for Glossary Creation: A Study on Extracting Definitions for Semi-automatic Glossary Creation in Dutch.* Utrecht: Lot Dissertation Series (252).

Zgusta, Ladislav, 1971: *Manual of lexicography.* Berlin/New York: De Gruyter Mouton.

Žagar, Mojca, 2011: *Terminologija med slovarjem in besedilom: analiza elektrotehniške terminologije*. Ljubljana: Založba ZRC, ZRC SAZU.