

Univerza v Ljubljani/University of Ljubjana  
Filozofska fakulteta/Faculty of Arts  
Oddelek za prevajalstvo/Department of Translation

Senja Pollak

**Polavtomatsko modeliranje področnega znanja iz večjezičnih korpusov**

**Semi-automatic Domain Modeling from Multilingual Corpora**

Doktorska disertacija/Doctoral dissertation

Mentorica/Supervisor:  
*Izr. prof. dr. Špela Vintar*  
University of Ljubljana, Faculty of Arts,  
Ljubljana, Slovenia

Študijski program: Prevodoslovje  
Study program: Translation Studies

Somentorica/Co-supervisor:  
*Prof.ssa. Paola Velardi,*  
University La Sapienza, Rome, Italy

Ljubljana, 2014



# Table of contents

<b>Acknowledgements .....</b>	<b>v</b>
<b>Povzetek .....</b>	<b>vii</b>
<b>Abstract .....</b>	<b>ix</b>
<b>1 Introduction .....</b>	<b>1</b>
1.1 Domain modeling .....	1
1.2 Research goals .....	1
1.3 Contributions to science .....	2
1.4 Structure of the thesis .....	3
<b>2 Background and related work.....</b>	<b>5</b>
2.1 Language and meaning: Linguistic perspective .....	5
2.1.1 Lexical meaning .....	5
2.1.2 Lexicography and terminography .....	7
2.1.3 Dictionaries and terminological collections.....	11
2.1.4 Definitions.....	12
2.1.5 Types of lexical definitions (defining strategies) .....	14
2.1.6 Lexicographic principles of meaningful definitions .....	22
2.2 Domain modeling: Computational perspective.....	24
2.2.1 (Semi-)automatic domain modeling: Extracting terms, definitions, semantic relations and ontology construction .....	24
2.2.2 Modeling the domain of language technologies .....	28
2.2.3 Web services and workflows.....	29
<b>3 Problem description, corpus presentation and initial domain modeling .....</b>	<b>31</b>
3.1 Problem description.....	31
3.2 Building the Language Technologies Corpus .....	33
3.2.1 Constructing the small LTC proceedings corpus .....	34
3.2.2 Constructing the main Language Technologies Corpus (LT corpus).....	35
3.3 Domain modeling through topic ontology construction .....	37
3.3.1 Modeling the LTC proceedings corpus .....	37
3.3.2 Modeling the Language Technologies corpus.....	42
3.4 Setting the stage for automatic definition extraction: Analyzing definitions in running text .....	44
3.4.1 Genus et differentia definition type.....	45
3.4.2 Defining by paraphrases, synonyms, sibling concepts or antonyms.....	51
3.4.3 Extensional definitions.....	52
3.4.4 Other types of definitions: defining by purpose or properties .....	53
<b>4 Methodology and background technologies.....</b>	<b>57</b>
4.1 Overview of the definition extraction methodology .....	57
4.2 Definition extraction evaluation methodology.....	59
4.3 Background technologies and resources .....	60

4.3.1	ToTrTaLe morphosyntactic tagger and lemmatiser .....	60
4.3.2	LUIZ terminology extractor .....	62
4.3.3	WordNet and sloWNet.....	63
4.3.4	CloudFlows workflow composition and execution environment .....	64
4.4	Evaluation of selected background technologies .....	64
4.4.1	ToTrTaLe evaluation.....	65
4.4.2	LUIZ evaluation.....	66
<b>5</b>	<b>Definition extraction from Slovene and English text corpora .....</b>	<b>71</b>
5.1	Extracting definitions from Slovene texts.....	71
5.1.1	Pattern-based definition extraction .....	72
5.1.2	Term-based definition extraction .....	85
5.1.3	SloWNet-based definition extraction .....	96
5.2	Extracting definitions from English texts .....	100
5.2.1	Pattern-based definition extraction .....	100
5.2.2	Term-based definition extraction .....	107
5.2.3	WordNet-based definition extraction.....	114
5.3	Results of Slovene and English definition extraction methods and their combinations .....	118
5.3.1	Combining different approaches on the Slovene subcorpus.....	118
5.3.2	Combining different approaches on the English subcorpus .....	120
5.3.3	Subjectivity of evaluation results.....	123
5.3.4	Analysis of different types of definition candidates .....	124
<b>6</b>	<b>Workflow implementation in CloudFlows .....</b>	<b>133</b>
6.1	Load corpus widget.....	134
6.2	ToTrTaLe widget.....	135
6.3	LUIZ widget.....	137
6.4	Definition extraction widgets.....	137
6.4.1	Pattern-based definition extraction widget .....	137
6.4.2	Term-based definition extraction widget .....	138
6.4.3	Wordnet-based definition extraction widget.....	138
6.5	Auxiliary widgets.....	138
6.5.1	Merge sentences widget.....	138
6.5.2	String to file widget .....	138
6.5.3	Term viewer widget .....	139
6.5.4	Sentence viewer widget .....	139
<b>7</b>	<b>Conclusions and further work .....</b>	<b>141</b>
<b>8</b>	<b>References .....</b>	<b>145</b>
<b>9</b>	<b>List of figures .....</b>	<b>161</b>
<b>10</b>	<b>List of tables.....</b>	<b>163</b>
<b>APPENDIX A: Term-based definition extraction experiments .....</b>		<b>165</b>
<b>Razširjeni povzetek .....</b>		<b>171</b>

## Acknowledgements

I would like to thank my supervisor Špela Vintar, for her enduring confidence, patience and friendly support during the entire duration of my doctoral studies. I would like to thank the co-supervisor Paola Velardi for welcoming and hosting me during my stay at the Sapienza University, as well as for her prompt reading of the final version of the thesis. The two remaining members of the doctoral committee, Darja Fišer and Dunja Mladenić, both gave very useful and detailed comments that improved the final version of the dissertation. It was Dunja who initially, even before I began the PhD studies, directed me into research.

I am grateful to my colleagues: Janez Kranjc, Nejc Trdin, Tomaž Erjavec, and especially Anže Vavpetič, who helped with the workflow implementation; Jasmina Smailović, who was of great help in the corpus preprocessing phase and initial topic domain model construction; Janja Sterle and Živa Malovrh who were helpful in the corpus and glossary construction phase, and Ana Zwitter Vitez and Damjan Popič who helped me with the evaluation of results and reviewing the text.

As a junior researcher, I was funded by the Slovene Research Agency (ARRS), and at this point my deep gratitude again goes to my supervisor Špela Vintar who selected me as her doctoral student. This led to my employment at the Department of Translation at the Faculty of Arts, which provided a friendly work environment, where I was also able to gain valuable teaching experience. During my stay at the Sapienza University in Rome, I was also financially supported also by the Italian Government. Finally, I would like to express my gratitude also to my current employer, the Jožef Stefan Institute, where in the last year I was able to gain new knowledge and complete my thesis in perfect working conditions.

Above all, the greatest thank you goes to my mother whose moral and scientific support was invaluable all along. Last but not least, there are many good friends and family members, without whom the time of writing and completing this thesis, with all the unpredictable events in the last few years, would have been much more difficult, even impossible: my profound gratitude goes to each and every one of you.



## Povzetek

Človeško znanje je dostopno v strokovnih besedilih, terminoloških slovarjih in enciklopedijah, v zadnjem času pa tudi v računalniku razumljivih predstavitev 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 eno- ali večjezično luščenje terminologije, definicij ali semantičnih relacij kot tudi (pol)avtomatske pristope h gradnji ontologij. Luščenji terminologije in definicij sta pomembna koraka modeliranja strokovnega znanja, vendar so razvite metode in orodja večinoma prilagojena za posamezne jezike, a le redko za manj razširjene jezike, kot je slovenščina. Zato je glavni doprinos doktorske disertacije, ki ponuja metodologijo za luščenje definicijskih stavkov iz korpusov v slovenskem in angleškem jeziku, prav luščenje definicij iz slovenskih nestrukturiranih besedil.

Predlagana metodologija temelji na treh različnih pristopih k luščenju definicijskih stavkov. 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). Razvito metodologijo, ki uporablja kombinacijo treh metod, smo preizkusili na resničnem problemu modeliranja področja jezikovnih tehnologij. V ta namen smo zgradili primerljivi slovensko-angleški korpus tega področja, ki vsebuje približno dva milijona pojavnic. Od približno 3400 izluščenih definicijskih kandidatov smo več kot 700 stavkov ocenili kot definicije. Rezultat tega dela je tudi pilotni *Slovarček jezikovnih tehnologij*.

Poleg predlagane metodologije je doprinos doktorske disertacije tudi kvalitativna analiza avtomatsko izluščenih definicijskih kandidatov. Poleg osnovne razvrstitve v dve kategoriji (stavek je ali ni definicija) smo končni nabor stavkov analizirali in označili tudi s podrobnejšimi kategorijami. V predlagani analizi so dodatne oznake ločene v kategorije, vezane na obliko definicije, vsebino definicije, definiendum, segmenatacijo ter označevanje.

Dodatni prispevek disertacije 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.

**Ključne besede:** luščenje definicij, spletni delotoki, modeliranje domene, specializirani korpusi, jezikovne tehnologije



## Abstract

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. Terminology and definition extraction are important domain modeling steps. The approach is proposed for Slovene and English, but is easily adaptable to other languages. 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.

An additional contribution of the dissertation is the qualitative analysis of automatically extracted definition candidates. This set of candidate definitions was analyzed and annotated with fine-grained categories added to the binary definition/non-definition tags. In the analysis, the tags are sorted into definition form, definition content, definiendum, segmentation, and annotation-related categories. One of the important results is the pilot Glossary of Human Language Technologies for Slovene.

An additional contribution is the proposed domain-modeling pipeline—from corpus uploading and preprocessing to inspecting the extracted term and definition candidates—implemented as an online publicly available workflow, easy to use for translation, linguistic or terminological tasks. The developed workflow components, including the ToTrTaLe corpus annotation tool, can be easily integrated in other natural language processing workflows.

**Key words:** definition extraction, online workflows, domain modeling, specialized corpora, language technologies



# 1 Introduction

Domain modeling and extracting domain-specific knowledge from texts have become proliferate areas of natural language processing and information extraction, involving tasks such as topic detection, terminology extraction, extraction of semantic relations, definition extraction, named entity recognition and other tasks aimed at harvesting meaningful items of knowledge. This dissertation focuses on the task of domain understanding through definition extraction from domain corpora, with an emphasis on developing methods and tools for definition extraction from Slovene texts. The introduction of the present thesis presents the topic of research, research goals, contributions to science and the structure of the thesis.

## 1.1 Domain modeling

Domain terms and definitions—as means for domain knowledge modeling—are normally collected in handmade monolingual or multilingual terminological dictionaries or glossaries. Taxonomies and ontologies (e.g., Gruber, 1993) have proven to be very adequate knowledge representation formalisms for expressing the relations between domain terms, while topic ontologies (Fortuna et al., 2007), expressing a taxonomy of domain topics, provide a different view on a domain (where a domain is represented as a set of documents).

Since a large amount of domain knowledge is represented in domain texts in an unstructured way, as well as in semi-structured encyclopedic resources such as Wikipedia and WordNet (Fellbaum, 1998), researchers in natural language processing, computational linguistics and text mining have started developing methods and tools for semi-automatic extraction of domain-specific knowledge from texts, involving tasks such as terminology extraction, definition extraction, semantic relations extraction or semi-automatic ontology construction. However, a vast majority of these methods and tools are language specific, often not developed for minor languages such as Slovene.

The potential of the proposed framework for domain modeling from multi-lingual text corpora will be demonstrated on a selected case study - the domain of language technologies. The application of the proposed methodology (using existing and newly developed automatic and semi-automatic knowledge extraction techniques) on a selected corpus, will result in a proof of concept glossary in the domain of human language technologies (HLT).

## 1.2 Research goals

Manual construction of glossaries and taxonomies, let alone the development of domain specific ontologies, represent a significant investment of effort and resources constructed for a new domain. Moreover, the need for constant upgrading/development/evolution of specialized domain models represents a threat that once the project is completed, it quickly becomes outdated. For this reason, the area of

natural language processing showed significant interest in automatic term and definition extraction as well as in semi-automatic taxonomy/ontology construction.

The main research question addressed in this thesis is how to automatically extract domain knowledge from unstructured domain corpora in Slovene and English in order to semi-automatically generate a domain knowledge model formed of a glossary of the selected domain, as a basis for further refinement by human experts. This research question is motivated by the idea that such a model has the potential to decrease the amount of human resources needed for modeling a new domain.

The dissertation focuses on the task of domain understanding through definition extraction from domain corpora, as definitions are an important mode of representation for specialized concepts. They play a crucial role in the process of establishing the conceptualization of a given domain as they help to delimit and differentiate concepts. They are an indispensable part of specialized dictionaries, thesauri and ontologies, and can help non-experts and translators to understand and correctly use specialized linguistic expressions which are the vehicles of knowledge transfer. The overall goal of this dissertation is to develop a methodology and a tool for semi-automatic domain modeling through definition extraction from domain corpora, focusing on Slovene. We also aim to implement the methodology in an easy to use workflow environment, without any computational knowledge needed to use it.

Another important aspect of this dissertation is the analysis of definitions in running text. Our focus is on in-depth analysis of automatically extracted definition candidates, in order to better understand the definition extraction task and related problems, defining strategies in academic writing and a variety of definition types.

In brief, given a corpus of domain texts, the main goals of the dissertation are the following:

- To develop an overall methodology for definition extraction as a process starting from a raw text corpus, annotating it automatically with morphosyntactic descriptors, followed by term extraction, definition candidate extraction, human selection and evaluation.
- To relate the lexicographic theory of definitions to the task of automatic definition extraction from running text.
- To provide a pilot Slovene glossary for the language technologies domain.
- To implement the proposed definition extraction methodology as a reusable workflow, show-cased for definition extraction from Slovene and English text corpora but easily adaptable to other languages.

### 1.3 Contributions to science

Main contributions of this dissertation are as follows:

- A definition extraction methodology, based on three different modules for extracting definition candidates: the pattern-based, the term-based and the wordnet-based<sup>1</sup> definition extraction module.

---

<sup>1</sup> As in Fišer (2009), we use small caps with the word *wordnet* to refer to the type of collections with literals, synsets, hypernyms, etc., whereas *WordNet* denotes the particular wordnet of English,

- Construction of a corpus of articles from the domain of language technologies in Slovene, and a comparable corpus in English. A part of the corpus is available through a concordancer at the following address: [http://nl.ijs.si/cuwi/sdjt\\_sl/](http://nl.ijs.si/cuwi/sdjt_sl/).
- Annotated set of more than 33,000 corpus sentences (labeled with definition/non-definition categories).
- Qualitative analysis and typology of over 3,400 definition candidates detected in the corpus under investigation (categorization by definition type, content, etc.).
- A pilot Glossary of language technologies,<sup>2</sup> consisting of approximately 500 definitions (available at: [http://kt.ijs.si/senja\\_pollak/jt\\_glosar/](http://kt.ijs.si/senja_pollak/jt_glosar/))
- Definition extraction workflow implementation of the proposed definition extraction methodology as an online workflow, available for public reuse (available at: <http://clowdflows.org/workflow/1380>).

Additional contributions are the workflow implementations of previously existing tools:

- ToTrTaLe workflow implementation of the ToTrTaLe preprocessing tool (Erjavec et al., 2010) for corpus preprocessing (tokenization, lemmatization and morphosyntactic annotation) as an online workflow, available for public reuse (available at: <http://clowdflows.org/workflow/228>).
- LUIZ term extraction web service and widget implementing the monolingual part of the LUIZ system (Vintar, 2010) was previously available online only as a demo for Slovene, while now it is fully functional for both languages, easily reusable in any new workflow.

Parts of this work have been published in the following papers: Initial definition extraction methodology was published in Fišer et al. (2010), where the approach was applied to a text corpus of a different genre (mainly textbooks) than the corpus used for definition extraction in this thesis (mainly scientific texts). Pre-final definition extraction methodology implemented in the workflow environment was published in Pollak et al. (2012a, 2012c). The ToTrTaLe workflow for corpus preprocessing in Slovene was published in Pollak et al. (2012b, 2012c), while the Language Technologies Corpus and initial domain models in terms of topic ontologies were presented in Smailović and Pollak (2011, 2012).

## 1.4 Structure of the thesis

The thesis is structured as follows. Following a brief introduction given in this chapter, Chapter 2 presents the background and the related work, where linguistic and natural language processing perspectives are provided. Chapter 3 defines the problem of domain modeling in terms of topic ontology construction, terminology and definition

---

developed at the Princeton University (Fellbaum, 1998; PWN, 2010), the first collection of this kind.

<sup>2</sup> The glossary includes also definitions of Živa Malovrh in Janja Sterle.

extraction, and is followed by the description of the process of corpus construction and the analysis of several definitions from the corpus. Chapter 4 introduces the definition extraction methodology, presents the evaluation measures and discusses the background technologies and their evaluation. Chapter 5 contains the core of the thesis: the pattern-based, term-based and wordnet-based approaches to definition extraction and their evaluation for each language. Section 5.1 presents the three methods and the results of definition extraction from the Slovene part of the corpus and Section 5.2 the three methods and the definition extraction results obtained on the English subcorpus. Section 5.3 summarizes the results, proposes different novel combinations of the three methods for each language and proposes a qualitative systematization of the results. In Chapter 6 the details of the definition extraction workflow implementation are discussed, while Chapter 7 concludes the thesis and gives directions for further work. Throughout the thesis numerous examples of extracted definition candidates are presented and analyzed, illustrating the difficulty of the task (corpus) and improving the understanding of the domain in terms of domain modeling. The thesis is supplemented with Appendix A, which describes the testing of different parameter settings for the term-based approach.

## 2 Background and related work

The practice of creating dictionaries in the sense of recording and explaining the lexical inventory of a language or language pair is the central goal of lexicography and boasts a long tradition. Building terminological dictionaries and other terminological collections is a younger, but very dynamic activity due to the rapid emergence of new specialized fields. In recent years, semi-automatic approaches to the creation of dictionaries and glossaries, as well as the extraction of other relevant information from text corpora have been developed. In this chapter we discuss the linguistic (Section 2.1) and natural language processing (Section 2.2) perspectives on this topic.

### 2.1 Language and meaning: Linguistic perspective

Concrete or abstract realities, to which an utterance refers, give rise to certain ideas or mental images in the human mind. A particular group of such ‘things’ can give rise to ideas that resemble one another (form the same concept) and that are different from other ‘things’, since they have some distinctive features in common. In communicative acts, we do not list the distinctive features of the concepts, but give the concepts a linguistic representation by way of a name. To understand the relationship between the words (lexical units) and what they refer to, we analyze the notion of lexical meaning. The meaning of the words can be explained by means of definitions, which can be collected in dictionaries. The definitions can be categorized into different definition types, and in dictionary compilation there are some principles that should be respected in order to have meaningful definitions.

#### 2.1.1 Lexical meaning

*Lexical meaning* consists of several components, the *designation* or *denotation* (the ‘objective’, ‘real’ meaning), the *connotation* (the ‘subjective’, ‘emotive’ meaning), and (possibly) the *range of application*, the latter being related to the fact that every word’s applicability is limited by some of its properties, being related to its stylistic value, semantic connections or its grammatical category (Zgusta, 1971, p. 27, 42, 89; Svendsen, 1993, p. 118). Lexical meaning is not carried only by single words, but concerns also multiword lexical units (Zgusta, 1971, p. 154), while it links to the function of cognition as a reflection and reconstruction of experience (Geeraerts, 2010, p.11). When we use a word in a sentence, it is not the lexeme in a sentence, but a particular *instantiation* of that lexeme, and those instantiations are called *lexical units* (Murphy, 2010, p. 10).

*Denotation* can be understood as the relation between words and the extralinguistic world—the things or classes of things they denote—i.e. *denotatum*. However, this relation is neither simple nor direct; for example, the denotatum (or the reference) of two expressions may be the same, while their meaning may be different (Geeraerts, 2010, p. 78). Between the word (lexical unit) and the denotatum, there is *designatum*, which can be understood as the conception which stands between the reality and the

word. For the speakers of a language, the whole extralinguistic world is organized into designata (Zgusta, 1971, p. 27–32).

Word meanings are namely not substantial, but relational (also, defined by what they are not), and according to structuralists, we can differentiate between paradigmatic and syntagmatic relations. Paradigmatic relations amount to the fact, that they can fill the same position in a sentence or an expression (cf. Lyons, 1977), while the syntagmatic approach maintains, that the word is defined by the other words that accompany it in language use, or the totality of its uses (Paradis, 2012). Glanzberg (2011) for example argues, that usually, words get their meanings in part by associating with concepts, but only in conjunction with substantial input from language (i.e., they get linguistically modulated meanings).

A notion similar to designatum is (*scientific*) *concept*. The difference between them is, according to Zgusta (1971, p. 32), that the concept is the result of exact scientific work or at least logical thinking, and is usually exactly defined and rigorously used, while the designatum generally does not have these qualities. In a certain way a concept is therefore a special case of the designatum. There is no sharp line between the two notions—that of *designatum* and that of *concept*—and it often happens that a precise scientific concept is worked out on the basis of a designatum of a word which is then used both as a *word* of general use (expressing the designatum) and as a *term* (expressing the concept). For instance, if we have a word like *polyvinylchloride*, the designatum and the precise scientific concept coincide, but if we have the word *animal* and the term *animal*, there is a difference, because the term expressing the concept covers also entities, which would not necessarily be conceived as animals in a general use of the word (Zgusta, 1971, p. 32–33); take *corals* as an example.

The difference between *terms* and *words* does not correspond exactly to the difference between the general and scientific usage, but triggers practically all the spheres of the languages and concerns different degrees of preciseness. We can also see in different literature that the notion of *concept* often relates to both (*scientific*) *concept* as Zgusta uses it, as well as the less precise *denotatum* and as Zgusta notes (1971, p. 33), in the case of languages that have a long tradition of philological, philosophical and generally cultural work a great part of the designata indeed tend to approach the status of precise concepts. In the Saussurian tradition, the *concept* would be on the side of the *signifié*—content aspect of a sign—while the *word* or *term* is the counterpart of the expressional aspect—the *signifiant* (Saussure, 1997).

To sum up, in the field of designation, the relation of the words to the segments of the extralinguistic world, there are three main elements: the (form of the) *word* (or term) as the expression capable of being communicated to the hearer (or the reader, etc.), the *designatum* (or the concept) as the respective mental, conceptual content expressed in it, and the *denotatum* as the respective segment of the extralinguistic world (Zgusta, 1971, p. 33). Note however, that all the words do not have precisely the same type of lexical meaning; for purely designative words (lexical units) the denotatum is easier to conceive than for function words, pragmatic operators, deictic words, etc.

The *connotation* as the second component of lexical meaning can be understood as “all components of the lexical meaning that add some contrastive value to the basic, usually designative, function” (ibid., p. 38). Hjelmslev (1975) notes, that in the process of signification, connotation necessarily follows denotation as a second step. Examples of words with the same designation, but different connotation are *to decease*, *to die* and *to peg out*.

“Any stylistic property of a word, the fact that a word belongs to a certain style of the respective language, to a certain slang or a social dialect, or even to a geographical dialect (if the word is used in a non-dialectal context), or that it is either recently coined (a neologism) or on the contrary, obsolete carries additional semantic relevance, additional ‘information’ about the speaker, about his attitude to or evaluation of the subject, gives ‘color’ to the subject, conveys the information more powerfully, humorously, emotionally, ironically, is in consequence more expressive, and is, therefore, connotative” (Zgusta, 1971, p. 40).

The third component of lexical meaning is according to Zgusta (1971, p. 41) the *range of application* or in other words *selectional restrictions*. Briefly, it concerns words that have the same designation and connotation, but are differently used depending on the context, e.g., *stipend* and *salary* both refer to financial remuneration for work, but the first is mostly used in connection with a teacher or priest and whereas the latter is used in connection with an official. Also the aspect of a style whether the word belongs to the general or technical language is part of the meaning related to the range of application (Svensen, 1993, p. 118).

One of the properties of lexical meaning is its *generality*. This can be perceived in different dimensions. First, a given designative lexical unit can be used in reference to any member of the class that belongs to the designatum (i.e., word *flower* can be used in reference to any flower); second, designata are usually broad and frequently overlapping (many different things can be referred to as *flower*); last but not least, the polysemy adds considerably to the generality of lexical meaning (Zgusta, 1971, p. 47). However one should note that in terminological work, the generality is much more limited, even if Zgusta warns that “even technical terms are polysemous more frequently than one would think (e.g., *carburettor* (1) in a combustion engine (2) in an apparatus for manufacturing water gas)” (ibid., p. 61) and as will be discussed in Section 2.1.2 the term *terminology* itself is the best proof of polysemy.

We can also differentiate between general nouns that are used to express general concepts denoting a group of things with common distinctive features, and proper nouns which are used when individual concepts are referred to (Svensen, 1993, p. 115–116). In other words, common nouns are used to refer to *categories* of things, while proper nouns are used for *instances*.

In contrast to generalization, *concretization* is the result of the application of a lexical unit in an actual utterance. Therefore, whereas lexical meaning is general, signification is concrete. This concretization is the result of the contextualization of the relation between the concrete thing and the context (Zgusta, 1971, p. 47).

## 2.1.2 Lexicography and terminography

The totality of means of expression in a language can be divided into *general language* and *special language*. Even if between the two there is no distinct boundary, it can be said that *general language* defines the sum of the means of linguistic expression encountered by most speakers of a given language, whereas *special language* goes beyond the general vocabulary based on the socio-linguistic or the subject-related aspect. Consequently, two different categories of special language can be identified. *Group language* serves the purpose of strengthening the sense of belonging within a social group, whereas *technical language* arises as a consequence of constant development and specialization in the fields of science, technology, and sociology

(Svensen, 1993, p. 48–49).

In the context of terminology, *special language*, also called *language for special purposes*, was defined as “language used in a subject field and characterized by the use of specific linguistic means of expression”, where it is also specially noted that “the specific linguistic means of expression always include subject-specific terminology and phraseology and also may cover stylistic or syntactic features” (ISO 1087-1:2000a). We can see that in this sense the term *special language* corresponds to the technical language and does not cover group language in Svensen’s terminology. Another term frequently used as synonym is *specialized language*.

The discipline that deals with studying the meaning(s) of words and their structure in general language is called *lexicology*. Even if lexicology and lexicography are terms that are sometimes used as synonyms, *lexicography* has in fact a different notion. The most basic understanding of lexicography is related to compiling dictionaries, but according to Svensen (1993, p. 1) lexicography means more than that:

“lexicography is a branch of applied linguistics which consists in observing, collecting, selecting, and describing units from the stock of words and word combinations in one or more languages. /.../ Lexicography also includes the development and description of the theories and methods which are to be the basis of this activity.”

In contrast to lexicology, the science of terminology deals with special language, i.e., language from special subject field (domain). First, we investigate different meanings of the term *terminology*. It can refer to (at least<sup>3</sup>) three things: (a) the methods of collecting, disseminating and standardizing terms, (b) the theory explaining the relationships between concepts and terms, and (c) a vocabulary of a particular discipline (Pearson, 1998, p. 10). In ISO standards, *terminology* is used and defined as a “set of designations belonging to one special language” (ISO 1087-1:2000a) and therefore corresponds to the notion under (c), while (a) can be linked to the term *terminology science* defined as “science studying the structure, formation, development, usage and management of terminologies in various subject fields” (ISO 1087-1:2000a). Point (b) has been already discussed in the section above and we continue with this topic and mention the changes in the understanding of this dichotomy in the rest of this section, as well as closely related questions of differences between words and terms or the disciplines of lexicology and terminology.

As new concepts constantly appear, new linguistic expressions have to be coined. A new denotatum (either newly discovered or invented) results in a new designatum/concept (or they come into existence step by step together), and the new designatum/concept finds expression in a new lexical unit, word/term. In terminological work it often happens that one can readily describe a concept that has no name. For example, if a new product has been developed, and the concept, with its name, is to be incorporated in the technical terminology of the field, the terminologist’s first step is to clarify and describe the content of the concept, and only then to provide it with a suitable name (Svensen, 1993, p. 48–49, 116). However, this is only one—*onomasiological*—view, in which a concept is taken to be a prior and the name (term in our case) is found for it and is the basis of the traditional, classical approach to

---

<sup>3</sup> Humbly (1997, p. 13) mentions that Bergenholtz (1995) gives four and Bruno de Bessé (1994) five different meanings.

terminology. The opposite view is the *semasiological* perspective, which starts from terms and the work of a terminologists is to explain their meanings. This—semasiological perspective—is also a background principle for the contemporary, corpus-based terminography, although both *onomasiological* and *semasiological* principles coexist in terminographic work.

If lexicologists and lexicographers mainly focus on words or lexemes, the terminologists focus on words that became terms, i.e., the words that have acquired protected status when used in special subject domains or as called above subject fields (Pearson, 1998, p. 7).

The different understanding of *word* vs. *term* can have different notions within different theories. We have already mentioned in the section above how Zgusta relates it to different levels of preciseness and to the difference between designatum (less precise, expressed by words) and concepts (precise, expressed by terms). Wüster (1979 in Pearson, 1998, p. 10) sees the difference between the terminology and lexicology disciplines based on the fact that terms should be treated differently from general language words. In contrast to lexicology where the lexical unit is the usual starting point, terminology word starts from the concept and the concept should be considered in isolation from its label or term. Concepts are understood to exist independently of terms, since they are mental constructs to which we assign labels. Each concept is the product of a mental process whereby objects and phenomena in the real world are first of all perceived or postulated.

In contemporary approaches, the dichotomy ‘word-term’ is wiped-out. For Kageura (2002) terms are functional variants of words. Cabré (2003) claims that all terms are words by nature. Cabré (2003, p. 189) notes that “we recognize the terminological units from their meaning in a subject field, their internal structure and their lexical meaning”. Myking (2007, p. 86) says that the traditional terminology is concept-based and the new directions are lexeme-based. The difference is seen also in the form, since a term can also contain non-alphabetic signs. Next, we provide the definitions of basic notions as defined by ISO standards.

**Term:** Verbal designation of a general concept in a specific subject field (ISO 1087-1:2000a).

**Word:** Smallest linguistic unit conveying a specific meaning and capable of existing as a separate unit in a sentence (ISO 5127:2001).

**Designation:** Representation of a concept by a sign which denotes it. Note: In terminology work three types of designations are distinguished: symbols, appellations and terms (ISO 1087-1:2000a).

**Concept:** Unit of knowledge created by a unique combination of characteristics (ISO 1087-1:2000a).

Finally, we discuss the term *terminography*. In ISO standards (ISO 1087-1:2000b) terminography refers to the part of terminology work concerned with the recording and presentation of terminological data, similar definition was coined by Marie Claude L'Homme (2004, p. 15) who defines terminography as the acquisition, compilation and management of terms:

“la terminographie regroupe les diverses activités d’acquisition de compilation et de gestion des termes.”

As noted by S. E. Wright (2011) about the ISO definition “many native-speakers of English object to the term ‘terminography’, but it is widely used in Canada”.

Terminography and *terminological work* can be also used as synonyms and in ISO standards *terminological work* is defined as work concerned with the systematic collection, description, processing and presentation of concepts and their designations (ISO 1087-1:2000b).

In analogy to lexicography, concerned with collecting and describing the basic units of general language, i.e., lexemes (or words), as well as building general language dictionaries and reflecting the theoretical aspect of this process, terminography can be described as science dealing with concepts and their namings, i.e., terms, with the aim of building terminological dictionaries (or other terminological manuals, e.g., term banks, glossaries, etc.). On one hand *lexicography* can also deal with theoretical aspect of dictionary building without actually constructing the dictionaries (cf. Svensen, 1993, p. 1). The question remains whether *terminography* can also exist aside from actual building of terminological collections. In the majority of works, the terminology covers the theoretical part and terminography concerns only actual development of terminological collections. For example, Vintar (2008, p. 5) states that the final aim of any terminographic work is the construction of a terminological collection, being an extensive terminological dictionary or small personal glossary. Similarly, Cabré (1999) suggests that terminography is *terminology in practice*, while Baker and Saldanha (2009, p. 288) mention also an alternative naming *applied terminology*. Similar to this distinction, for some authors theoretical lexicography means lexicology and the practical part lexicography. Since terminography deals with special language, the term *specialized lexicography* is sometimes used.

The growth of electronic resources and tools has substantially influenced the traditional dictionary building processes, where the term *electronic lexicography* is used to refer to the design, use and application of electronic dictionaries (Granger, 2012, p. 2). The integration of computer technology into dictionaries can vary from simply making the paper dictionary content available through the electronic medium, up to taking full advantage from its electronic form (Fuertes-Olivera and Bergenholtz, 2010, p. 1; Granger, 2012, p. 2).

Granger (2013, p. 2–11) highlights six most significant innovations of electronic lexicography in comparison to the traditional methods: *corpus integration* meaning the inclusion of authentic texts in the dictionaries, *more and better data* since there are no more space limitations and one has the possibility to add multimedia data, *efficiency of access* (quick search and different possibility of database organization), *customization* meaning that the content can be adapted to the user's needs, *hybridization* denoting that the limits between different types of language resources—e.g., dictionaries, encyclopedias, term banks, lexical databases, translation tools—are breaking down, and *user input* since collaborative or community-based input is integrated. The principles of Slovenian e-lexicography are discussed by Krek et al. (2013) in the proposal for a new Slovene dictionary.

The merging of lexicography with computer technology and constant growth of corpora has enabled the development of (semi-)automatic processes for term extraction and alignment between different languages, and more recently, also definition extraction, which is the main topic of this dissertation. Automatic approaches will be discussed in Section 2.2. Note, however, that as it can be seen from this section, the distinctions lexicology vs. lexicography and terminology vs. terminography are far from being unanimous and are also often used interchangeably, illustrating also the difficulty of the terminology and definition extraction tasks addressed.

### 2.1.3 Dictionaries and terminological collections

A *dictionary* is a document which contains a list of lexical units and relevant information about them. It is composed of short dictionary entries, arranged in a conventional, usually alphabetical order (Svensen, 1993, p. 2; De Bessé et al., 1997, p. 129). However, the organization of dictionary entries has become much more flexible and dynamic in the era of electronic dictionaries.

Traditionally dictionaries were printed books, but today one can say that they “are most familiar in their printed form; however, increasing numbers of dictionaries exist also in electronic forms which are independent of any particular printed form” (TEI P5, 2013, p. 261). The future of dictionary making lies in electronic dictionaries and many specialists predict the disappearance of paper dictionaries in the near future (Granger, 2012, p. 2).

Traditionally, the difference between a dictionary and an *encyclopedia* is that the dictionary gives information about individual units of the language, whereas encyclopedia communicates the knowledge about the world. In other words, linguistic dictionaries are primarily concerned with language, i.e., focus on explaining the meaning of words/lexical units of language and their linguistic properties, while encyclopedic dictionaries deal with explaining the meaning of phenomena, i.e., the denotata of the lexical units/words (Svensen, 1993, p. 2; Zgusta, 1971, p. 198).

If the lexical items are structured according to semantic relations (synonyms, hypernyms, etc.), we talk about *thesauri* (De Bessé et al., 1997, p. 154).

*Terminological dictionaries*, also called *technical dictionaries*, are collections of terminological entries presenting information related to concepts or designations from one or more specific subject fields (ISO 1087-1:2000a). In contrast to general dictionaries dealing with general vocabularies, they cover specialized domain vocabularies and are more focused on defining and naming concepts than on the linguistic side, such as pronunciation and inflection of the included lexemes (Svensen, 1993, p. 3, 21).

A *glossary* can have two different meanings. Either it is defined as a “terminological dictionary which contains a list of designations from a subject field, together with equivalents in one or more languages” (ISO 1087:2000a), or—as used also in this thesis—a glossary can refer to a (unilingual) list of terms and their definitions (or other explanations of their meaning) in a particular subject field (De Bessé et al., 1997, p. 134).

If the collection of terms is structured according to the conceptual relationships established for a subject field, it is a *terminological thesaurus* (De Bessé et al., 1997, p. 154).

If a terminological collection is in a computer-processable form, it is called a *terminological database* or *termbase*, defined as “database containing terminological data” (ISO 1087-1:2000b) or in the previous ISO version as “structured sets of terminological records in an information processing system” (ISO 1087:1990). A collection of terminological databases including the organizational framework for recording, processing and disseminating data is called—in the later withdrawn ISO 1087-2:2000 standard—a *term bank*.

With the era of the 21<sup>st</sup> century, an important change was observed with more and more dictionaries and other collections in electronic format (Granger and Paquot, 2012). The limits between the above mentioned categories of lexical resources are blurred. Very broadly *electronic dictionaries* can be defined as “primarily human-oriented

collections of structured electronic data that give information about the form, meaning, and use of words and are stored in a range of devices (PC, Internet, mobile devices). *Computer-oriented lexicons* are, on the other hand, lexical tools that are primarily designed for use in natural language processing applications (Granger, 2013, p. 2) and often the resources can be used by humans and computers (cf. WordNet (Fellbaum, 1998)).

Recently, much effort has been invested in building modern language resources for Slovene. *Slovene lexical database* (Gantar and Krek, 2011) can be used as the basis for lexicographic purposes—as described in the proposal for a new dictionary of Slovene language (Krek et al., 2013)—as well as an enhancement of natural language processing tools for Slovene. The database provides different levels of lexico-grammatical information, spanning from simple morphological data to syntactic and collocational data and corpus examples. Another lexical resource, sloWNet (Fišer, 2009), is (since we use it in our methodology) presented more in detail in Section 4.3.3, and is a resource of high value for various language technology applications providing the information about word senses, their hypernyms, other relations, translations and definitions. Termania,<sup>4</sup> on the other hand, is a web portal, combining many different mono- or multilingual, general and terminological dictionaries that can be submitted and searched through by any user.

#### 2.1.4 Definitions

In the following three subsections we discuss different views on definitions as found in the related lexicographic and philosophical literature. Different categorization that we discuss in this chapter summarize others' work, but will be referred to in our analyses in Section 3.4, as well as throughout Chapter 5.

The meaning of dictionary entry words and word combinations is specified by *definitions* in monolingual dictionaries and by means of *equivalents in the other language* in bilingual dictionaries (Svensen, 1993, p. 6). We use definitions to define the meaning. Definitions are definitions of symbols and not of objects/things, because only symbols have the meanings that definitions may explain. For example, we can define the word *chair* because it has meaning, but not a chair itself, since an actual chair is not a symbol that has meaning (on the other hand, we can sit on a chair or describe it, but we cannot sit on a symbol/word *chair* (Copi and Cohen, 2009, p. 88).

One of the fundamental tenets of traditional lexicography is that the meanings of all lexical items can be expressed by means of a paraphrase in the same language, the definition (Béjoint, 2000, p. 195). Béjoint (ibid.) referring to (Dubois and Dubois, 1971, p. 85) also identifies the presupposition that there are always at least two ways of expressing something, without changing the meaning, as *semantic universal*.

Zgusta (1971, p. 252) claims, focusing on the general dictionary building, that the basic instruments for the description of lexical meanings are the *lexicographic definition*, the *location in the system of synonyms*, the *exemplification* and the *glosses*. We focus on definitions, as well as synonyms as alternative method of defining a concept, while setting aside the purpose of examples and/or glosses in dictionaries, the glosses being defined by Zgusta (1971, p. 270) “as any descriptive or explanatory note within the entry”, where also labels indicating the connotation, style, etc. are in his

---

<sup>4</sup> <http://www.amebis.si/termania> (Last accessed: February 1, 2014)

opinion a species of glosses.

A *definition* is a characterization of the meaning of the (sense of the) lexeme (Jackson, 2002, p. 93). It is “a representation of a concept by a descriptive statement which serves to differentiate it from related concepts” (ISO 12620:2009).

The concept to be defined is called a *definiendum* and corresponds to the headword in the context of dictionary building. And the definition defining the meaning of the concept (*definiendum*) is called *definiens*. In fact the *definiens* is not the meaning of the *definiendum*, but it is—as the *definiendum* itself—a symbol, or group of symbols, that has the same meaning as the *definiendum* (Copi and Cohen, 2009, p. 88). In monolingual dictionaries the two parts are usually separated. However, in some of the second language learner dictionaries (cf. COBUILD dictionary projects (Sinclair, 1987a)) as well as from the point of view of automatic definition extraction, the entire sentences, containing the *definiendum* and the *definiens* are considered and the linking element between the two parts is in this context called a *hinge* (most commonly a verb) (Sinclair, 1987b; Hanks, 1987; Pearson, 1996; Barnbrook and Sinclair, 1994; Krek, 2004; Kosem, 2006).

Definitions as found in dictionaries, are only one definition category. In philosophy, several other categories are identified, depending on their function (cf. Copi and Cohen, 2009, p. 88; Parry and Hacker, 1991, p. 89–97).

In *lexical* (or *real*) *definitions* a term being defined has already some established use and therefore the definition reports the *definiendum*'s (prior and independent) meaning. These definitions are *true* or *false* (depending on whether they do or do not accurately report common usage—conventional meaning). An example of true lexical definition is defining a word *bird* as *any warm-blooded vertebrate with feathers* (Copi and Cohen, 2009, p. 89–90).

*Stipulative* (or *nominative*) *definitions* are the definitions that are not factually *true* or *false*, but are the ones in which a new (or already existing) term is assigned specific meaning by definition and did not have (that) meaning before. It is a “proposal /.../ to use a *definiendum* to mean what is meant by the *definiens*” (Copi and Cohen, 2009, p. 89, see also Robinson, 1962) and if a term already exists it might be in contradiction with its lexical definition. For instance, the number equal to a billion trillions ( $10^{21}$ ) has been named a *zeta* by stipulation.

*Precising definitions* are used to eliminate ambiguity or vagueness of terms. An example provided by Copi and Cohen (2009, p. 92) is the vagueness of the term *horsepower* that initiated a precising definition (*the power needed to raise a weight of 550 pounds by one foot in one second*). In contrast to stipulative definitions, the *definiendum* of precising definitions is not a new term, the *definiendum* should be assigned a more precise meaning, but respecting the established usage.

Copi and Cohen (2009, p. 94–95, p. 116) list also *theoretical definitions* that serve as comprehensive compressed summaries of some theory (their aim is to encapsulate the understanding of some intellectual sphere) and *persuasive definitions* which are used to influence the conduct of others (e.g., commonly used in political argumentation).

In the next two subsections we explain different defining strategies and the types of lexical definitions, as well as the principles for well-formed definitions, as stated in the lexicographic literature.

### 2.1.5 Types of lexical definitions (defining strategies)

In this section we focus on lexical definitions and different defining strategies. We list the most important strategies and categories as found in the related literature. When analyzing the definitions in our corpus (c.f., Section 3.4), we refer to a selected (simplified) subset of these categories.

Svensen (1993, p. 117) distinguishes between *true definitions*, *paraphrases* (also including synonyms and near synonyms), *combined definitions* (hybrids of the two types mentioned before) and definitions by *describing the use* of the defined term. We can note that the term *true definitions* has itself a connotation of better defining a concept than a paraphrase or synonym or other defining strategy. The ‘true’ definitions can define the concept by specifying its *intension* or its *extension*. The intension denotes the *content* of the concept, which can be defined as the combination of the distinctive features which the concept comprises, while the extension denotes the *range* of the concept, which can be defined as the combination of all the separate elements or classes which the concept comprises (Svensen, 1993, p. 120–121). To illustrate the difference, Svensen (1993, p. 121) provides elements that should be specified by each method to define e.g., a *motor vehicle*. The intension should be specified as ‘vehicle + engine-driven + steerable + mainly for use on roads or tracks’, while the extension could be specified as ‘car or motor cycle or moped or van or bus or truck’. The *extensional meaning* of a term<sup>5</sup> is the collection of the objects that constitutes the *extension* of the term (Copi and Cohen, 2009, p. 96). All the objects within the extension of a given term have some common attributes or characteristics that lead to the same term to denote them. The *intension* of the term is the set of attributes shared by all and only those objects to which a general term refers. The *intentional meaning* supposes some criterion for deciding whether a given object falls in the extension of that term. Every general term has both an intensional and extensional meaning, where the term’s intension determines its extension; terms may have different intensions and the same extension (e.g., *living person* and *living person with a spinal column* have different intension, the latter being greater than the first, but the same extension; the extension of a term can also be empty), but terms with different extensions cannot possibly have the same intension (ibid., p. 96–98). To sum up, the basic difference can be made by defining strategies that approach the term by focusing on the class of *objects* to which the term refers (extensional definitions) and the others focusing on the *attributes* that determine the class denoted by the term (intentional definitions) (ibid., p. 98–99). Next, we examine different principles and types on these two main defining strategies.

#### Intentional definitions

Intension of a term means the attributes shared by all the objects denoted by the term, and shared only by those objects—or in other words—all the attributes shared by all and only the members of the class designated by that term (Copi and Cohen, 2009, p. 102, 116).

Copi and Cohen (2009, p. 102) distinguish three different senses of intension: *subjective intension* (the set of all attributes the speaker believes to be possessed by

---

<sup>5</sup> Note that in the following sections, for simplification purposes, we do not make a distinction between words and terms (and designata and concepts). We use the term *term* in a wider sense, not necessarily in the terminological sense related to a specific subject field.

objects denoted by the word), *objective intension* (the factual total set of characteristics shared by all the objects in the term’s extension) and *conventional intension*. The latter is used for definitions and refers to a stable meaning of a term based on the implicit agreement between users to have the same criterion for deciding about any object whether it is part of the term’s extension, but does not presuppose the omniscience (ibid.).

In related literature, we found six different subtypes of intentional definitions that are analyzed below. The main and most common definition type for dictionary building is *definition by genus and differentiae*, where the meaning of a term is *analyzed* and the term is defined by superordinate concept (class)—*genus*—and the differences of the species denoted by the term from the members of all other species of the *genus*; the second is *synonymous and paraphrases definition type* where another word (or paraphrase) has the same meaning as the word being defined. We extensively discuss these two types, since they are the most important for lexicographic and terminographic work. We also mention *relational definitions* in which terms are defined by relation (other than synonyms) to other terms, *operational definitions*, which state that a term is applied correctly to a given case if the performance of specified operations in that case yields a specific result, *functional definitions* defining a term by explaining its use and *typifying definitions* defining a term by means of its typical properties.

### ***Genus-differentia definition type (Analytical definitions)***

The most common form of lexicographic definition is the ‘analytical one-phrase definition’, which consists of the *genus proximum* (superordinate concept) next to the definiendum—or just after the hinge in a full sentence definition—together with *differentia specifica*, i.e., at least one distinctive feature typical of the *definiendum* (Svensen, 1993, p. 122). It is called *analytical*, because the definition does not only provide the meaning of an unknown concept, but it also analyzes its definiens into constituent features (Geeraerts, 2003, p. 89). The analyzability can be understood in terms of classes. Any class of things having members may have its membership divided into subclasses. The class whose membership is divided into subclasses is called *genus* and the various subclasses are its *species* (Copi and Cohen, 2009, p. 105). The *definiendum*’s superordinate concept—*genus*—specifies the class containing the definiendum as one element, while the distinctive features—*differentiae*—specify in which ways the definiendum differs from other elements in the same class (Svensen, 1993, p. 122).

We provide two *genus-differentia* definitions in which we add the notation of different parts of definiens, namely *genus proximum* and *differentia specifica*. The most typical order is that of Example (a), where *genus* precedes the *differentia*. However, *differentia* can be also specified before the *genus*, as in Example (b) below.

- a) **pedestrian:** person who goes or travels on foot (Zgusta, 1971, p. 254)
- A horizontal line is drawn under the words 'person who goes or travels on foot'. A bracket below the line spans from 'person' to 'who', labeled 'genus proximum'. Another bracket below the line spans from 'goes' to 'foot', labeled 'differentia specifica'.
- b) **square:** equal-sided rectangle (Svensen, 1993, p. 122)
- A horizontal line is drawn under the words 'equal-sided rectangle'. A bracket below the line spans from 'equal-sided' to 'rectangle', labeled 'differentia specifica'. Another bracket below the line spans from 'rectangle' to the end of the line, labeled 'genus proximum'.

It is also possible that the two methods are used, as in Example (c).

- c) **horse:** a solid-hoofed plant-eating domesticated mammal with a flowing mane and tail, used for riding, racing, and to carry and pull loads (Jackson, 2002, p. 94).

In the last example the *definiendum* (*horse*) is related to its *genus*<sup>6</sup> (*mammal*) and given a number of *differentiae* (*solid-hoofed, plant-eating, domesticated, with a flowing mane and tail, used for riding, etc.*) which are typical features of a *horse* compared to other *mammals* (see Jackson, 2002, p. 94). Even if it is more common that the *differentia* come after the *genus* part, the *genus* can already be restricted by some specific elements (*differentiae*), such as the first three properties in the above-mentioned example.

Svensen (1993, p. 124) warns that since the content of a sign and not the expression is to be defined, if possible a definition should not use expressions such as *name of...* or *objects, such as...* with exception of definitions of e.g. function words. Note that this position is not fully aligned with the one of Copi and Cohen (2009) that definitions are always definitions of symbols (this discussion is above the scope of this thesis).

The *genus* should be neither too general nor too specific (Ayto, 1983 in Kosem, 2006), but this also depends on the final application. A definition of a concept in a terminological dictionary is different than in general dictionaries: the terminological dictionaries have more detailed definitions (Svensen, 1993, p. 3, 22). A difference between terminological and general language dictionaries is in Svensen's (1993, p. 122–123) opinion that in terminological work, the definition should include as many distinctive features as are needed to demarcate the concept from every other member of the class, whilst in general-language dictionaries, this rule is not applied to the same extent and only “enough distinctive features should be mentioned to represent the content of the sign with accuracy sufficient for the purposes of the dictionary”. Zgusta similarly notes when signaling the differences between the *logical definition* and the *lexicographic definition*, saying that “whereas the logical definition must unequivocally identify the defined object (the *definiendum*) in such a way that it is both put in a definite contrast against everything else that is definable and positively and unequivocally characterized as a member of the closest class, the lexicographic definition enumerates only the most important semantic features of the defined lexical unit, which suffice to differentiate it from other units” (Zgusta, 1971, p. 252–253). Zgusta also claims that the lexicographer should respect that the (lexicographic) “definition should be sufficiently specific, but not overspecific”, where the “indication of semantic features is based on what appears to be relevant to the general speaker of the language in question, not on properties that can be perceived only by a scientific study” (Zgusta, 1971, p. 254). This again differentiates the lexicographic definition for the purposes of general dictionary building, compared to the terminological perspective, where specialists (or translators in need of exact translations) and not a general speaker are the addressed audience. However, when technical terms are defined in general dictionaries, it is often difficult to satisfy the scientific correctness and general intelligibility (Zgusta, 1971, p. 255).

Svensen (1993, p.123) therefore notes that for general language dictionaries, it is often enough to provide only *genus proximum* (which does not need to be a direct superordinate concept) and possibly—but not necessarily—one or two distinctive

---

<sup>6</sup> Note that it is not a *genus proximum*.

features. An example he provides (ibid.) is defining *canasta* as “a card game” or *calcium* as “a chemical element”, but also notes that in these cases it is obligatory to use the indefinite article or expression such as *a kind of* or *a type of*, in order to prevent the interpretation of the definition as a paraphrase (e.g., not every card game is called *canasta*). This definition, providing only the *genus* (the referent’s class) but not the *differentia*, can be therefore considered as a special subtype of analytical definition. It is called *classificatory* (Borsodi, 1967, in Westerhout, 2010) or *exclusive genus* (Sierra et al., 2006, p. 230) definition type.

*Quantitative* and *qualitative definitions* can be considered as special subtypes of analytical definitions, since their specificities concern more the *differentia* part than the general *genus-differentia* structure. These categories were introduced by Borsodi (1967) and are summarized in Westerhout (2010, p. 37). *Quantitative definitions* describe the dimensions (size, weight, length, age...) of the definiendum (e.g., *A mountain is a peak that rises over 2,000 feet (609,6m)*), while *qualitative definitions* state the qualities, characteristics, or properties of the definiendum.

A special subcategorization was made by Nakamoto (1998) in the context of language learners monolingual dictionaries. He distinguishes between two groups of lexicographic definitions, based on two different ‘perspectives’. He analyzed four British dictionaries of English as a foreign language. *Referent-based definitions* (RBSs) define the definiendum from the perspective of the entity to which they refer, while *anthropocentric definitions* (ACDs) are written from the perspective of a person. To illustrate the two types, two examples of dictionary definitions of *watch* are provided by Nakamoto (1998, p. 205):

- d) a small clock to be worn, esp. on the wrist, or carried
- e) a small clock that you wear on your wrist or carry in your pocket

Even if both definitions are *analytical definitions* consisting of the *genus proximum* (*clock*) and the *differentiae specificae* (what differentiates a *watch* from other types of *clocks*), Nakamoto (1998) identifies the most important difference in the perspective (cf. the use of second person pronoun *you, your*). The use of informal pronoun *you* was introduced systematically, along with full sentence definitions, in Sinclair’s (1987a) COBUILD dictionary (Nakamoto, 1998, p. 211).

Even if the *analytic (genus-differentia)* type of definition is the most common type in dictionary construction and is considered even as the most “prestigious” type, Béjoint (2000, p. 199) questions this presupposition and proposes that the relative efficacy of different types should be compared depending on different user groups and different word categories.

### ***Paraphrases and synonyms***

A second major type of definition consists of a *paraphrase*, i.e., “a brief rewriting of a name” (Svensen, 1993, p. 118). In this definition type, we can find a synonym, a collection of synonyms or a synonymous phrase (paraphrase). Jackson (2002, p. 95) and Zgusta (1971, p. 261) state that smaller dictionaries with limited space use this defining method more frequently and that it is especially used for abstract words. It depends on different authors whether paraphrases on one side and synonyms and near-synonyms on the other are considered as one or two different definition types.

Complete synonyms are the words that are equivalent in their *denotative* and *connotative meaning*, as well as their *range of applications*, i.e., the three aspects of the

lexical meaning elaborated in Section 2.1.1 above. Complete synonyms are more usual in technical terminology, but outside the technical language *near-synonyms* are much more frequent, meaning that words are denotatively equivalent, but have different connotations and/or belong to different style level. *Near-synonyms*, or *synonyms in the broader sense of the word*, express great similarity instead of absolute identity.

Svensen (1993, p. 131) and Zgusta (1971, p. 260) mention that a quite usual case is to find a hybrid form of a definition, consisting of a *genus-differentia* or paraphrase definition, followed by one or more synonyms. However, Svensen (1993, p. 132) thinks that a better combination is using the paraphrase and a (near-)synonym, whilst the combination of *genus-differentia* definition and (near-)synonym can be misleading and that if used it should be labeled as such (e.g. using abbreviation *syn.*). Defining by synonyms and near synonyms—also called *synthetic definitions* in contrast to *analytical definitions* that are analyzing the meaning of a term—has been judged differently by different authors. Zgusta (1971, p. 261) claims that “if handled with due care, this method can yield good results”.

### ***Relational definitions***

*Relational definitions* refer to other than synonym-based synthetic definitions in which the definiendum is defined in relation to other terms. Borsodi (1967, in Westerhout, 2010), distinguishes between *antonymic definitions* that refer to the help of referents of the opposite nature (e.g., *Bad is the opposite of good* (Westerhout, 2010, p. 38)) and *meronymic definitions*<sup>7</sup> that explain the word by situating it between two other terms—“a simple definition by the enumeration of words which refer to any thing or any document which is between, or which is mediatory of, the extremes represented by synonyms and antonyms of the word being defined” (Borsodi, 1967, p. 27) (e.g., *The present is the moment in time between past and future* (ibid.)).

### ***Operational definitions***

Term’s intention can also be explained operationally—by tying the definiendum to some specific test. The test should be a public and repeatable operation (using specific processes or validation sets), i.e., a prescribed procedure that has an observable result. It is mainly used for distances, durations, etc. (Copi and Cohen, 2009, p. 103; Parry and Hacker, 1991, p. 106). An example is a sentence like: *An acid is a substance that turns blue litmus red*. This definition type was mainly related to the fields of physics or psychology (an example is identifying intelligence with the score in IQ tests) and is strongly criticized by some authors (e.g., Swartz, 2010).

### ***Functional definitions* (use/purpose/function)**

Jackson (2002, p. 95) identifies the type where the definition explains the *use* to which a (sense of the) word/term is put, especially when defining function words with no reference outside the language. Since we define the term by its purpose/function, we call it a *functional definition*. Jackson’s (ibid.) example is: *and (conjunction) used to connect words of the same part of speech, clauses or sentences*. Functional definitions

---

<sup>7</sup> In our opinion the chosen name of this definition type is somewhat confusing, since it is not used the classical ‘part/member-of’ meaning of meronymy.

can be also found as a subtype of analytical definitions, in which the *genus* is mentioned and the purpose/use is described in the *differentia* part. For instance, E. Westerhout (2010, p. 38) gives as an example *Gnuplot is a program for drawing graphs*, but calls it an operational definition, which is confusing with regard to the above-mentioned operational definitions as usually used in the literature. Also Copi and Cohen (2009, p. 107) mention that a lexical definition “should state the conventional intension of the term being defined”, which is not always the intrinsic characteristic of the things denoted by the term. The use of shape, or material as specific difference of a class is therefore according to Copi and Cohen usually an “inferior way to construct a definition”. For instance, for a *shoe* it is not essential that it is made of e.g., leather, but the use, i.e., being an outer covering for the foot. In functional definition the use/purpose/function can concern the *differentia* part (and thus be sort of a subcategory of the *genus-differentia* definition type) or have an autonomous structure.

### ***Typifying definitions***

Typical properties of the referent are usually used in combination with one of the previous techniques, especially with analytical definitions or paraphrases. Since they normally contain the *genus*, they can also be considered as a subtype of analytical definitions, where the *differentia* part mentions the typical properties. Jackson (2002, p. 95) provides an example for *measles* defined as *an infectious viral disease causing fever and a red rash, typically occurring in childhood*.

“Typifying definitions are structured similarly to analytical ones. They contain a *genus proximum* and one or more characteristic features. Nonetheless, instead of focusing on additional inherent facts, the definition gives more information on what is typical of the referent” (Heuberger, 2000, p.16 in May, 2005, p. 74).

If the most typical characteristic is definiendum’s use, we can refer back to functional definitions, interpreted in that way as a subtype of the typifying definitions (May, 2005).

The last two defining strategies, i.e., *functional definitions* defining by mentioning the function/use of the definiendum and *typifying definitions* defining by the most typical characteristics of the definiendum are very often employed in the *full sentence definitions* introduced in COBUILD dictionaries (Sinclair, 1987a). The advantage of full sentence definitions is also discussed in Gantar and Krek (2009).

### **Extensional definitions**

In contrast to its intention, the extension of the definiendum can be used for defining its meaning. The extension basically means the objects denoted by a term. The most obvious way is to identify all the objects denoted by a term. However, it is not always possible and/or useful to list all the objects (Copi and Cohen, 2009, p. 99). We can therefore say that in extensional definitions the meaning of a term can be provided by means of listing *some* or *all* of the objects named by a term (Parry and Hacker, 1991, p. 113). Instead of referring to the content, they refer to the range of the concept (Svensen, 1993, p. 123). Svensen (ibid.) states that this type of definition is less usual in general-language dictionaries, and occurs mainly in terminographic work.

Extensional definitions can be categorized into different types, based on the situation in which they occur. For *ostensive definitions* the examples indicated are perceptually present to the audience, which is not the case for *citational definitions* (Parry and

Hacker, 1991, p. 113–114). Copi and Cohen (2009, p. 116) make a slightly different categorization: they distinguish between three types of extensive definitions: *definition by example*, in which we list, or give examples of the objects denoted by the term; *ostensive definitions*, in which we point to, or indicate by gesture the extension of the term being defined; and *semi-ostensive definitions*, in which the pointing or gesture is accompanied by a descriptive phrase whose meaning is assumed to be known (Copi and Cohen, 2009, p. 116).

We list five types of extensional definitions, where the basic distinction is taken from (Parry and Hacker, 1991, p. 113–114) with the categories of *citational* and *ostensive definitions*. We describe also the *partitive concept definitions*, *definition by paradigm example* and *contextual definitions*. The limitation of extensional definition type is with expressions that have no *observable* or *known denotata* or *no denotata at all*; those cannot be extensionally defined (Parry and Hacker, 1991, p. 115).

### ***Citational definitions***

“A citational definition is an extensional definition, in which some or all of the objects named by the definiendum are indicated verbally or represented by pictures, drawings, etc., but these objects are not perceptually present to the person to whom the definition is intended” (Parry and Hacker, 1991, p. 114). One should note that this understanding is not accepted by all the authors, since the use of extralinguistic elements such as drawings and pictures in a dictionary is for some authors always an example of ostensive definition (see below) (cf. Zgusta, 1971, p. 255).

An example of extensional defining strategy provided by the authors (Parry and Hacker, 1991, p. 113–114) is defining *West-Germanic languages* by giving positive examples of such languages (*English, German, Dutch, etc.*). Such a definition, does not explicitly state the property of a language in order to be *West-Germanic*, but gives the names of such languages to exemplify these properties. An optional strategy used with extensional definitions is to provide also negative examples, which are close to the category being defined but do not fall in it (in the example of *West-Germanic languages*, for instance *Danish* or *Icelandic* can be negative examples, since they are *Germanic* but not *West-Germanic* languages). If the negative examples do not share at least some properties of the defined category, they are not useful negative examples, e.g., *Chinese* as negative example does not tell much about *West-Germanic languages*, since the only common point is that it is a *language* (Parry and Hacker, 1991, p. 113). Citational definitions are most frequently simply referred to as extensional definitions, but also as definitions by example or exemplifying definitions. A special case of extensional (citational) definitions, in which all the examples of a finite set are enumerated, are sometimes called *enumerative definitions* (cf. Wikipedia, 2013).

The critics of this defining method say that its limitations are that it is not (always) possible to give a collection of cases to determine the exact meaning of a term and that one cannot be sure that the common element extracted is the right one (Lewis, 1929 in Westerhout, 2010, p. 39), since one cannot be sure which property or set of properties is being referred to (Parry and Hacker, 1991, p. 115). For instance, two terms with different intention can have the same extension (e.g., *equilateral triangle* and *equiangular triangle*). Moreover, not all types of term can be defined by this method (Robinson, 1972 in Westerhout, 2010, p. 39) and the condition is that the denotata are mutually known, otherwise the examples cited are of no use.

### ***Ostensive definitions***

“An ostensive definition is an extensional definition in which some or all of the objects denoted by the definiendum term are actually produced, presented, or shown to the audience” (Parry and Hacker, 1991, p. 114). Ostensive methods do not use only words to explain unknown concepts but define the object by extralinguistic strategies, such as indicating, pointing the object or using drawings or by demonstrative expressions or as Zgusta (1971, p. 256) claims “instead of differentiating semantic features, the ostensive definition indicates an example or some examples from extralinguistic world”. As already mentioned, Zgusta (ibid.) states that “the extreme case of an ostensive definition is a picture of the denotatum. Such a picture is an absolutely extralinguistic element within a dictionary”.

For example, to define *turquoise blue*, an object of that color can be pointed at or verbally indicated in sentences such as *That lamp shade is turquoise blue* (Parry and Hacker, 1991, p. 114). When a descriptive phrase is added to the definiens, the definition type is by some authors called *quasi-ostensive definition* (Copi and Cohen, 2009, p. 101). An example is *the word ‘desk’ means this article of furniture*, accompanied by the appropriate gesture, but such additions suppose prior understanding of the phrase *article of furniture*.

However, if no denotatum of the word is perceptually present, if words, although perfectly meaningful do not denote anything at all, an ostensive definition is impossible (Parry and Hacker, 1991, p. 115; Copi and Cohen, 2009, p. 101). For lexicographic and terminological work this definition type is less relevant<sup>8</sup>, but it is a very frequent defining strategy in everyday life, in children’s language acquisition or second language learning. *Ostensive definitions* are also sometimes called *demonstrative definitions*.

### ***Definition by paradigm example***

This definition type can be either ostensive or citational and is “a non-equivalential extensional definition using as example an object or objects intended to illustrate clearly and non-controversially the conventional intension of the definiendum term” (Parry and Hacker, 1991, p. 114). An example provided by the authors is when a person names Leonardo da Vinci or Rembrandt as paradigm examples of term *artist*, but adds that it is often followed by some form of conceptual definition (Parry and Hacker, 1991, p. 114–115). As we understand the notion, instead of enumerating all or at least a ‘sufficient’ number of the elements in the extension of the definiendum, one (or several) good representative of the class is chosen.

### ***Partitive concept definitions***

Svensen (1993, p. 122) mentions that the combination of separate parts belonging to a whole is sometimes ‘rather incorrectly’ also referred to as *extension*, but does not provide a separate category for this defining strategy. In Westerhout (2010) referring to Borsodi (1967) this type, in which the parts of the definiendum are listed, is called *anatomic definitions* (e.g., *A table consists of rows and columns*), but she lists this type under *analytical, intentional* definition type. An example when definiendum is an individual and not a general concept is to define *Benelux* by *Belgium, the Netherlands, and Luxemburg* (Svensen, 1993, p. 124). In the case when the definiendum is a general

---

<sup>8</sup> Except for pictures in dictionaries, if considered as ostensive definitions.

collective concept, and consequently the concepts included in the definiens are all of the same kind, no listing is necessary and the definition has usually the following form (ibid.), *quintet: group of five musicians*.

### **Contextual definitions**

Westerhout (2010, p. 39), referring to Gergonne (1818) and Borsodi (1967), explains the *contextual* (also called *illustrative* or *implicative*) *definition type* as definitions in which a term is *used* instead of *mentioned* and in which there is no distinction between definiendum and definiens, as there is no phrase equivalent to the term provided. A sentence implies/illustrates what something means. An example provided by Robinson (1954, p. 106) is *A square has two diagonals, and each of them divides the square into two right-angled isosceles triangles*. This defining strategy is widely used in COBUILD dictionaries (cf. Sinclair, 1987a; Gantar and Krek, 2009). In contextual definitions, again, definiendum's function or properties can be highlighted, so they can be recategorized in other definition types.

## **2.1.6 Lexicographic principles of meaningful definitions**

In monolingual dictionaries, the same language is being described and used for describing. And the lexicography has identified several principles in order to have meaningful definitions (Jackson, 2002, p. 2–3; Atkins and Rundell, 2008, p. 412). The principles and the problems related to them as stated in lexicography literature are enumerated below.

- A word should be defined in terms *simpler* than itself (Zgusta, 1971, p. 257). In other words, as Béjoint (2000, p. 195) says: “the definition works /.../ not only because the two sides (word and definition) have the same meaning, but also because the right-hand side (the definition or *definiens*) is more easily understandable than the left-hand side (the word, or *definiendum*). One perspective of implementing this principle is that words must be defined by more frequent words (Weinreich, 1962, p. 37 in Béjoint, 2000, p. 201). However Béjoint (ibid.) warns that this is not possible if the word to be defined is very frequent and that additionally more frequent words are more polysemous. Jackson also comments that this rule is not always possible with ‘simple’ words (Jackson, 2002, p. 93).

Similarly, Svensen (1993, p. 118) comments on paraphrases that they “should consist only of words that can be expected to be better known to the users than the headword or phrase they are intended to explain” and should be “more understandable than the headword” (Svensen, 1993, p. 119), as well as that “general-language words and phrases must not have technical paraphrases” (Svensen, 1993, p. 119). A special case of applying this rule is to have a systematically chosen range of (simple) words to be used in lexeme definitions: this is called a *defining vocabulary*. In terminographic work, this postulate is less applicable, since the use of technical vocabulary is important and needed. A technical-language definition of a technical term is often considerably more detailed and complex than a general-language definition (Svensen, 1993, p. 134).

- The previous principle can be contextualized with a more general, pragmatic principle. “The language used should be appropriate to the linguistic skills, and

presumed technical knowledge, of the user” (Atkins and Rundell, 2008, p. 412).

- A definition should be *substitutable* for the item being defined and therefore the head noun of the definition phrase should belong to the same word class as the defined lexeme (Zgusta 1971, p. 258). This point is discussed by Béjoint (2000, p. 205), mentioning that the rule cannot be respected in several cases, for example for function words. Therefore, a closely related principle is that different forms of definitions are appropriate to different types of words (Jackson, 2002, p. 94). Concerning paraphrases, Svensen (1993, p. 118) claims that “[a]s far as possible, a paraphrase should have a syntactic form such that it can be substituted for the headword or phrase in a passage of text without yielding an artificial-looking result”. This is an important point and Zgusta (1971, p. 258), quoting Weinreich, goes even further when saying that “a claim of interchangeability between the term and its definition ... is preposterous for natural language”.
- *Circularity* should be *avoided* (Svensen, 1993, p. 126; Jackson, 2002, p. 93). The most unacceptable type of circularity is when a definition uses the word being defined, for example defining a *triangle* as *polygon in the form of a triangle* (Svensen, 1993, p. 126). On the other hand, a very frequent—and according to Svensen (ibid.) perfectly acceptable—way is to use a part of the headword in the definition, since in many cases the name of the *genus proximum* can be the same as a part of the headword (e.g., *blacksmith: smith who works with iron*). Circularity can also go beyond a single definition, i.e., also two or more lexemes should not be defined in terms of each other (Jackson, 2002, p. 93). An example of circularity in definition of two lexemes is (Svensen, 1993, p. 126): *north: point of horizon to left of person facing east; left: direction of north when one is facing east*. The main point is that the user “shouldn’t have to consult another definition to understand the one s/he is looking up” (Atkins and Rundell, 2008, p. 412).

The circularity also depends on definition type. For example, relational definitions are always exposed to circularity issues. Also the synonymous defining strategy is likely to create circularity (Jackson 2002, p. 94). In partitive concept relationship definitions, the circularity can have the following form: *cell: part of a battery; battery: group of cells*.

Béjoint (2000, p. 203) says that the simple types of circularity can be avoided, but that circularity with more transition steps is “unavoidable and probably does not cause any practical problems”. Svensen (1993, p. 127) goes even further and says that in the context of general-language lexicography “it may even happen that circularity within a certain group of definitions in a system is necessary for the system to operate”.

- *Closed dictionary*, where closed means that all the lexical items in the microstructure should be also the elements of the macrostructure (Béjoint, 2000, p. 200). This point is related to the circularity restriction mentioned above, and postulates that the words used in the definitions should be defined in their own entries. Béjoint questions this point by example of small metalinguistic words, which would have to be defined by this principle, but would not qualify for inclusion in the (e.g., terminological) dictionary by any other criteria. And as Béjoint (2000, p. 201) says, the more ‘scientific’ a dictionary is in its definitions, the more difficult it is for a lexicographer to ‘close it’.

- In the *genus-differentia* definition type section, we have already discussed that a definition should state the *essential attributes* of the species, as well as that it should be neither too specific nor too broad.
- Other style related rules are that *ambiguous, obscure* or *figurative language must not be used* in a definition, as well as that a definition should *not be negative* when it can be affirmative (Copi and Cohen, 2009, p. 108–109). However, there are some examples, e.g., *bald* can be defined only negatively (as *the state of not having hair on one's head*). Another stylistic advice is that the language should “conform as far as possible to ‘normal’ prose” (cf. Atkins and Rundell, 2008, p. 412, contextualized in Gantar and Krek, 2009).

In the context of Slovene lexicography and analysis of dictionary definitions, several authors should be mentioned. Gantar and Krek (2009, p. 155) discuss different defining strategies by comparing classical dictionary definitions with full sentence definitions in the context of building the Slovene Lexical Database. The definitions of the reference dictionary of Slovene language (SSKJ) have been discussed and critically analyzed by Kosem (2006), Rozman (2010), Müller (2008), and Gantar and Krek (2009).

## 2.2 Domain modeling: Computational perspective

The dissertation addresses (semi-)automatic domain modeling. In this section we provide an overview of different approaches to automatic term and definition extraction, as well as means to taxonomy and ontology construction. We also present how the domains related to the language technologies domain were modeled in related research.

### 2.2.1 (Semi-)automatic domain modeling: Extracting terms, definitions, semantic relations and ontology construction

The dissertation deals with domain modeling by means of knowledge extraction, and in particular definition extraction from domain corpora. The basic units of knowledge in the domain text are terms. Automatic terminology extraction has been implemented for various languages (e.g., for English by Sclano and Velardi (2007), Ahmad et al. (2007), Frantzi and Ananiadou (1999), Kozakov et al. (2004)), and for Slovene by Vintar (2002, 2010). For bilingual term extraction, commercial (SDL MultiTerm, Similis (Planas, 2005)) and non-commercial (e.g., Gaussier, 1998; Kupiec, 1993; Itagaki et al., 2007; Lefever et al., 2009; Macken et al., 2013; Vintar, 2010) systems have been developed.

A more complex knowledge extraction task, which is also the main focus of this thesis, is the definition extraction task. Definition extraction approaches have been developed for several languages: for English (Navigli and Velardi, 2010; Borg et al., 2010), Dutch (Westerhout, 2010), French (Malaisé et al., 2004), German (Fahmi and Bouma, 2006; Storrer and Wellinghoff, 2006; Walter, 2008), Chinese (Zhang and Jiang, 2009), Portuguese (Del Gaudio and Branco, 2007; Del Gaudio et al., 2013), Romanian (Iftene et al., 2007), Polish (Degórski et al., 2008a, 2008b) as well as for other Slavic languages such as Czech and Bulgarian (Przepiórkowski et al., 2007). For Slovene, we have started to develop the methodology in Fišer et al. (2010) and Pollak et al. (2012a).

Besides definitions, (semi-)automatic extraction of other types of *knowledge-rich contexts* (Meyer, 1994) is of great importance, especially for terminographic purposes. The definitions usually contain *hypernymy* relations, while the extraction of other knowledge-rich contexts is based on semantic relations such as *meronymy* (part-whole),

*attribute* relations (how something looks like), *function/purpose*, *synonymy*, *antonymy* or *causal* relations (Meyer, 2001; L'Homme and Marchman, 2006).

Techniques for extracting definitions and semantically related concepts from large specialized corpora or web resources are either based on (manually crafted) rules or on machine learning, whereby recent studies often combine both.

Rule-based approaches are based on pattern matching using mainly syntactic and lexical features, but also in some cases paralinguistic and/or layout information. The patterns can be manually crafted, where Hearst's method (1992) for extraction of hyponym relations from large corpora, based on a set of lexico-syntactic patterns, is the main reference. Synonyms and hypernyms have been addressed in Malaisé et al. (2004). Other studies include patterns expressing meronymy (Berland and Charniak, 1999; Meyer, 2001; Girju et al., 2003), cause-effect (Marshmann et al., 2002; Feliu, 2004; Meyer et al., 1999; Garcia, 1997) or function patterns (Meyer et al., 1999).

Several authors use technical texts, where more structured knowledge is observed. Muresan and Klavans (2002) propose the DEFINDER rule-based system to extract definitions from technical medical articles and aim at automatic dictionary construction. They use cue-phrases, such as "is a term for", "is defined as", etc.) and text markers (e.g., "-“ or ”()") in combination with a finite state grammar (using part-of-speech rules and noun phrase chunking). They also perform grammar analysis based on statistical parsing to identify linguistic phenomena in definition writing (e.g., appositions, relative clauses, and anaphoras). Storrer and Wellinghoff (2006) based their patterns on 19 verbs that typically appear in definitions and used their valency frames for definition extraction. Using simple pattern-based methods on unstructured text, such as the internet, performs worse. Therefore some additional filtering methods are required, as proposed by Velardi et al. (2008) in the context of glossary building. They use patterns (e.g., "refers to", "defines", "is a") to extract candidates from the web and then filter them with a domain filter (based on available domain terms) and a stylistic filter in order to obtain the definitions with a '*genus et differentia*' structure from a specified domain.

Distinguishing different definitions types is proposed in Westerhout and Monachesi (2007) and was used in the European project LT4eL (2008) to build glossaries for e-learning in eight languages. Walter and Pinkal (2006) applied definition extraction to a legal domain with an interest in ontology building. They performed different experiments based on 33 rules (based on connectors) and identified 18 best-performing rules. Del Gaudio and Branco (2007) distinguish between *is*, *verb* and *punctuation* type. In Pollak et al. (2012a) and in this thesis we propose a combination of different methods for selecting definition candidates, i.e., patterns, term extraction and WordNet-based hypernyms and propose a web service implementation for the system.

The second line of relevant work is based on fully automatic methods, often using machine learning (ML). ML techniques are often used in combination with pattern recognition approaches or in more recent work as the main learning approach. Compared to pattern-based approaches, ML techniques require more training data and have to deal with often unbalanced datasets. The manually crafted rules can be considerably improved by ML techniques. For extracting definitions from medical articles, Fahmi and Bouma (2006) used a rule-based approach using cue-phrases and improved the performance of the method by using Naive Bayes, maximum entropy and SVM. As part of their feature set, they included sentence positioning, which is corpus specific. Standard ML classifiers were applied also to Polish texts (Degórski et al., 2008a) and Slovene texts (Fišer et al., 2010). Westerhout (2009, 2010) reports a

combination of grammar and ML techniques used in different experiments for different definition types, extracted from Dutch texts.

The problem of unbalanced datasets, typical for definition extraction tasks, was approached by Kobylński and Przepiórkowski (2008), who used Balanced Random Forest classifier to extract definitions from Polish texts, and by Del Gaudio et al. (2013) using different algorithms on English, Dutch and Portuguese corpora.

A fully automatic method is proposed by Borg (2009) and Borg et al. (2009, 2010), who use genetic algorithms to learn distinguishing features of definitions and non-definitions and weight the individual features.

Cui et al. (2007) propose a method for definitional question answering based on the use of probabilistic lexico-semantic patterns (i.e., soft patterns) that generalize over lexico-syntactic ‘hard’ patterns. Soft patterns allow a partial matching by calculating a generative degree of match probability between the test instance and the set of training instances. Navigli and Velardi (2010) propose automatically learnt Word Class Lattices (WCLs), a generalization of word lattices that they use to model textual definitions, where lattices are learned from an annotated dataset of definitions from Wikipedia. Their method is applied to the task of definition and hypernym extraction from corpora, as well as from the web. The method has been adapted to French and Italian by automatically constructing the training sets from Wikipedia (Faralli and Navigli, 2013). Reiplinger et al. (2012) also show how definitinal patterns can be extracted partly automatically, by bootstrapping initial seed of patterns.

The majority of studies focus on definitions and hypernymy concept relation extraction, but also several other types of relations have been explored. Navigli and Velardi (2010) and Snow et al. (2004) proposed methods for automatic hypernymy extraction. For meronyms, Girju et al. (2006) used machine learning techniques and WordNet. Ittoo and Bouma (2009) improved meronymy extraction precision by disambiguating polysemous meronymy patterns using modified distributional hypothesis (Harris, 1968). Yang and Callan (2009) presented a metric-based framework for the task of automatic taxonomy induction and consider hypernymy and meronymy relations. Their framework incrementally clusters terms based on *ontology metric*, using very heterogeneous sets of features (contextual features, co-occurrence, syntactic dependency, lexico-syntactic patterns...). Besides hypernymy and meronymy, Pantel and Pennacchiotti (2006) used generic patterns, refined using the web, for other types of relations, such as reaction, succession, production.

Beyond definitions and single semantic relations, recent research has addressed the hierarchical organization of knowledge, i.e., (semi-)automatic taxonomy (a hierarchy of is-a relations) and ontology induction, using databases, textual data or the web (Buitelaar et al., 2005; Biemann, 2005; Gómez-Pérez and Manzano-Macho, 2003; Maedche and Staab, 2009). One research direction is to consider ontology learning as a classification/clustering task, relying on the hypothesis of distributional similarity (Harris, 1954), where similar contexts define similar concepts (Cohen and Widdows, 2009; Pado and Lapata, 2007). Such approaches are able to discover relations, which do not explicitly appear in the text, but are less accurate and often unable to provide labels for the discovered semantic classes. Others rely on syntactic information as the first step of taxonomy construction.

Ontology-related tasks can be classified depending on whether the aim is to enlarge an existing, hand-crafted ontology (e.g., WordNet (Fellbaum, 1998) or the Open

Directory Project,<sup>9</sup> a task known as ontology extension and population, or build one from scratch. Snow et al. (2006) proposed a probabilistic model, combining evidence from multiple classifiers using syntactic or contextual information for the incremental construction of taxonomies. Yang and Callan (2009) proposed an incremental clustering approach based on calculating the semantic distance of each term pair in a taxonomy. Kozareva and Hovy (2010) took an initial given set of root concepts and basic level terms and used Hearst-like lexico-syntactic patterns to recursively harvest new terms from the web. To induce taxonomic relations between intermediate concepts they then searched the web again with surface patterns. Finally they applied graph-based methods for building the acyclic graph. In Navigli et al. (2011) and Velardi et al. (2013) the authors propose a well-performing method for inducing lexical taxonomies from scratch using a domain corpus or the web. Their first step is automatic term, definition and hypernym extraction (cf. Navigli and Velardi, 2010) which produces a cyclic, possibly disconnected graph, followed by using their new algorithm for inducing a taxonomy from the graph.

From the methodological perspective, our definition extraction approach is the most similar to the traditional pattern-based approach introduced by Hearst (1992), but our originality is in its combination with other methods, applying more shallow criteria (using the automatically recognized domain terms and wordnet terms with their hypernyms). Unlike the majority of authors, we aim at extracting definitions in their broader sense, since we consider focusing only on the genus-differentia definition type too limiting, especially when dealing with languages with less resources available and when extracting definitions from highly specialized running text. Since our main contribution is definition extraction from Slovene, we encounter similar problems as the authors working on other Slavic languages (e.g., Degórski et al., 2008; Przepiórkowski et al., 2007) since we deal with a morphologically rich, relatively free word order, determinerless language. Our results are comparable to the results of definition extraction on other Slavic languages. Content-wise our research is the most similar to Reiplinger et al. (2012), since she models the computational linguistics domain (cf. Section 2.2.2 below). As we do, she limits her search to scientific articles and we can see that in this setting even for English the results are far from very well performing systems using the web (e.g., Navigli and Velardi, 2010). Very frequently, the authors limit the search to a selected predefined list of terms, while we prefer the perspective as in Westerhout (2010), where the search is open to all the definitions. She is also one of the few authors that, in line with our research, lays great stress on the qualitative linguistic analysis and has the aim of final glossary construction. A big comparative advantage is that we implement our method as a freely available, online workflow, needing no previous installation or background knowledge to run the system on a new corpus, or thanks to its modularity combine and compare it with other approaches. Alternatively, we could try to train best performing existing systems (e.g., Navigli et al., 2011) on Slovene corpora, which will be considered in further work. The authors themselves (Faralli and Navigli, 2013) propose a solution for the automatic acquisition of reliable training sets for new languages from Wikipedia first sentences. We can also imagine adding information from wordnets and using parallel corpora. However, the performance of their lattice-base method has not yet been tested for any Slavic language.

---

<sup>9</sup> <http://www.dmoz.org/> (Last accessed: December 1, 2013)

## 2.2.2 Modeling the domain of language technologies

For the present dissertation, we have chosen the domain of language technologies, as a target modeling domain. There has been some previous work addressing closely related domains.

The ACL Anthology (Kan and Bird, 2013) is a digital archive comprising today over 24,000 documents from conferences and journals in computational linguistics. A subset of this anthology was used for constructing the ACL ARC reference corpus (Bird et al., 2008). Based on the ACL Anthology, Radev et al. (2009, 2013) created the ACL Anthology Network (AAN), a manually curated networked database of citations, collaborations, and summaries in the field of computational linguistics. These resources were used for various studies.

For topic discovery, to our knowledge, three studies have been performed. Hall et al. (2008) used topic models to analyze the topic changes in the domain (trend analysis) using the unsupervised Latent Dirichlet Allocation (LDA) approach developed by Blei et al. (2003) for inducing topic clusters. They also introduced topic entropy for measuring the diversity of ideas, for measuring the difference in time, as well as between different computational linguistics conferences. Paul and Girju (2009) addressed interdisciplinary topic modeling (also using LDA), where computational linguistics topics were discovered and aligned with topics in border fields of linguistics and education. Moreover, they dealt with topic changes over time, and were interested in discovering how different languages are represented in the field. Anderson et al. (2012) used the ACL ARC and the AAN corpora from which they automatically generated topics—again using the LDA—which were later manually labeled and the papers (and their authors) were attributed to these topics. The epochs in computational linguistics and the flow of authors between the areas were then analyzed.

Radev et al. (2009, 2013) proposed the citation analysis, more precisely statistics about the paper citation (e.g., the most cited authors in the field), author citation, and author collaboration networks. In Radev and Abu-Jbara (2012) the citation analysis was also used for identifying the trends in computational linguistics, as well as for summarization purposes, finding controversial arguments, paraphrase extraction and other tasks.

The most similar to our interest is the work of Reiplinger et al. (2012) extracting glossary of computational linguistics domain in English. The authors use the lexico-syntactic patterns and the deep syntactic analysis approaches to extract the candidates for glossary definitions.

Several other experiments were performed and presented in the workshop proceedings, edited by Banchs (2012). For example, Vogel and Jurafsky (2012) annotated the AAN corpus with authors' gender and analyze the differences in topics chosen by male and female researchers, where the topic models were again constructed using the LDA, while text reuse and authenticity analysis on the ACL domain was performed by Gupta and Rosso (2012).

Related fields of computer science or artificial intelligence were modeled also in terms of definition extraction and taxonomy induction. In LT4eL project (2008), the ICT and e-learning domain corpus was considered for definition extraction in several languages (e.g., Westerhout, 2010; Del Gaudio et al., 2013), while the artificial intelligence domain (in English) was modeled in Navigli et al. (2011) and Velardi et al. (2013).

This thesis addresses the language technologies domain, modeling it by means of

definition extraction (which is the main part of the thesis) as well as by briefly analyzing it through an approach to semi-automatic topic ontology construction (cf. Smailović and Pollak, 2011, 2012). None of the above mentioned authors neither modeled the Slovene domain nor proposed the topic ontology view.

### 2.2.3 Web services and workflows

This section presents the underlying principles of workflow composition and execution, as the basics of the technology used for implementing the NLP definition extraction workflow presented in Section 6. To the best of our knowledge, a workflow-based approach using a web service implementation of term and definition extraction modules has not yet been proposed.

Data mining environments, which allow for workflow composition and execution, implemented using a visual programming paradigm, include Weka (Witten et al., 2011), Orange (Demšar et al., 2004), KNIME (Berthold et al., 2008) and Rapid-Miner (Mierswa et al., 2006). The most important common feature is the implementation of a workflow canvas, where workflows can be constructed using simple drag, drop and connect operations on the available components. This feature makes the platforms suitable also for non-experts due to the representation of complex procedures as sequences of simple processing steps (workflow components named *widgets*).

In order to allow distributed processing, a service-oriented architecture has been employed in platforms such as Orange4WS (Podpečan et al., 2012) and Taverna (Hull et al., 2006). Utilization of web services as processing components enables parallelization, remote execution, and high availability by default. A service-oriented architecture supports not only distributed processing but also distributed workflow development.

Sharing workflows is an appealing feature of the *myExperiment* website of Taverna (Hull et al., 2006). It allows users to publicly upload their workflows so that they become available to a wider audience and a link may be published in a research paper. However, the users who wish to view or execute these workflows are still required to install specific software in which the workflows were designed.

The ClowdFlows platform (Kranjc et al., 2012), used for constructing the definition extraction workflow in this thesis, implements the described features with a distinct advantage. ClowdFlows is browser-based, requires no installation and can be run on any device with an internet connection, using any modern web browser. ClowdFlows is implemented as a cloud-based application that takes the processing load from the client's machine and moves it to remote servers where experiments can be run with or without user supervision. Moreover, the constructed workflows can be shared and directly executed, improving over the facility enabled in *myExperiment*.



### 3 Problem description, corpus presentation and initial domain modeling

Extracting domain-specific knowledge from texts is a challenging research task, addressed by numerous researchers in the areas of natural language processing (NLP), information extraction and text mining. In this chapter we first define the problem of domain modeling in terms of topic ontology construction, terminology and definition extraction. Subsequently, the construction of the corpus is described, followed by the discussion on different definition types encountered in a subset of our corpus.

#### 3.1 Problem description

This dissertation addresses the problem of domain modeling from multilingual corpora with the aim of improving domain understanding. The main challenge addressed in the dissertation and the main motivation for this research is to develop a definition extraction methodology and a tool for extracting a set of candidate definition sentences from Slovene text corpora. The secondary goal is to adapt this methodology and the developed tool also for definition extraction from English texts. In both cases, the focus is also on the linguistic analysis of the different kinds of definitions occurring in the corpus. As an auxiliary goal, we also address a different modeling approach, which enables initial domain structuring and understanding through topic ontology construction from documents constituting the given corpus. The language technologies domain is used as a case study in this dissertation.

The object of analysis are specialized texts from a certain domain, i.e., scientific texts including conference papers, journal articles, dissertations, etc., where the domain of interest is the language technologies domain. The basic units of knowledge in specialized texts are concepts, which are designated by either definitions or *terms*. Domain terminology therefore represents a first domain modeling step. While terminology acquisition has formerly represented a tedious manual task in the process of terminological dictionaries construction, research in automatic term extraction in the last decade has enabled automatic or semi-automatic terminology extraction for different languages, including Slovene (Vintar, 2002; 2010).

A more complex domain modeling task, which is the main focus of this thesis, is the *definition extraction* task. Definitions of specialized concepts/terms are an important source of knowledge and an invaluable part of dictionaries, thesauri, ontologies and lexica, therefore many approaches for their extraction have been proposed by NLP researchers, with very good results achieved especially for English. However, for Slavic languages the results are less favorable, which can be attributed especially to rich inflection and free word order (cf. Przepiórkowski, 2007). The first experiments in Slovene definition extraction have been achieved in our own research (Fišer et al., 2010; Pollak et al., 2012a). While Navigli and Velardi (2010), Navigli et al. (2011) and Velardi et al. (2013) for example, extract definitions not only from domain corpora but also by searching for definitions on the Web, our research addresses domain modeling for a

given domain corpus, assuming the lack of additional information for languages with poorer web resources.

With the exception of the work by E. Westerhout (2010), the concept of definition itself is rarely discussed in detail or given enough attention in the interpretation of results. A popular way to circumvent the fuzziness of the “definition of definition” is to label all non-ideal candidates as *defining* or *knowledge-rich contexts* to be validated by the user. In this thesis, we devote a great deal of attention to the analysis of the extracted definition candidates and discuss many borderline cases.

Our work is mainly focused on Slovene, a Slavic language with a very complex morphology and less fixed word order, hence the approaches developed for English and other Germanic languages, based on very large—often web-crawled—text corpora, may not be easy to adapt. In general, definition extraction systems for Slavic languages perform much worse than comparable English systems, as reported by e.g., Przepiórkowski et al. (2007), Degórski et al. (2008a, 2008b), Kobyliński and Przepiórkowski (2008). One of the reasons is that many Slavic languages, including Slovene,<sup>10</sup> lack appropriate preprocessing tools, such as good lemmatization and part-of-speech tagging systems, parsers and chunkers, needed for the implementation of well-performing definition extraction methods. Another obstacle is the fact that very large domain corpora are rarely readily available.

The presented work follows our work reported in Fišer et al. (2010), in which we have reported on the methodology and the experiments with definition extraction from a Slovene popular science corpus (consisting mostly of textbook texts). In that work, in addition to definition candidate extraction, we used a classifier trained on Wikipedia definitions to help distinguishing between good and bad definition candidates, where the first sentences in Wikipedia typically follow the Aristotelian *per genus et differentiam* structure (“X is Y which/that...”), in which a term to be defined (*definiendum* X) is defined using its hypernym (*genus* Y) and the difference (*differentia*, introduced by which/that ...) that distinguishes X from other instances of class Y. When analyzing these results we already observed that the main reason for the mismatch between the classifier’s accuracy on Wikipedia definitions versus those extracted from textbooks was the fact that, in authentic running texts of various specialized genres, definitions run an entire gamut of different forms.

In this thesis—unlike in most related work where definitions are extracted from textbooks, manuals, Wikipedia articles or large web collections—the corpus consists of limited amount of scientific articles and theses (in language technologies domain), where concepts are more complex and knowledge is encoded in linguistically more intricate structures. In this setting, the assumption that definitions follow exclusively the Aristotelian *per genus et differentiam* structure is unrealistic, and other approaches to definition extraction need to be developed. Moreover, the domain is not chosen as a proof of concept only, but the extracted definitions are indeed proposed as a basis for a Slovene HLT glossary or terminological dictionary construction. Moreover, another focus of this thesis is to explore a variety of definition types appearing in running text.

An important distinguishing feature of this thesis is the implementation of the developed methodology in a novel, browser-based workflow construction and execution environment ClowdFlows (Kranjc et al., 2012). On the one hand, this approach enables

---

<sup>10</sup> Some of the tools for Slovene were elaborated very recently, but were not yet available of the time of conception of the work presented in this thesis (e.g., Grčar et al., 2012; Dobrovoljc et al., 2012).

the reuse of the developed workflow for definition extraction purposes from any new corpus, as well as the reuse of individual workflow components in the construction of new natural language processing workflows.

### 3.2 Building the Language Technologies Corpus

For English, the task of definition extraction is often performed on large corpora, in some cases gathered from the Web. For highly specialized domains and for languages other than English, the Web may not provide an ideal source for the corpus, especially not for the purpose of terminology extraction where a certain level of representativeness and domain coverage is crucial.

Within this dissertation we created a corpus of specialized texts. We decided to consider the domain of language technologies for several reasons. Firstly, the domain is not yet modeled in terms of Slovene terminology and definitions (the state of the art for English was discussed in Chapter 2). Secondly, the language technologies domain is an example of a highly specialized domain where researchers write more in English than in their mother tongue; this situation is very typical for Slovene specialized language also in other domains. Moreover, the language technologies domain was selected, as for the evaluation of the results the basic comprehension of domain vocabulary and extracted definitions is needed.

For covering the domain of language technologies we proceeded in the following way. The most straightforward decision was to consider the papers published in the Proceedings of the biennial Language Technologies conference *Jezikovne tehnologije*, organized in Slovenia since 1998. Seven consecutive conference proceedings (1998–2010) were included. The articles in the proceedings are in Slovene or in English. In the rest of the thesis we refer to this part of the corpus as the *Language Technologies Conference proceedings corpus (LTC proceedings corpus)*, or simply the ‘small corpus’.

To improve vocabulary coverage we added other text types from the same domain, including Bachelor’s, Master’s and PhD theses, as well as several book chapters and Wikipedia articles on the subject. Also these documents are in Slovene or in English. The extended corpus including the small corpus is hereafter referred to as the *Language Technologies Corpus (LT corpus)* or simply the ‘main corpus’ or ‘the corpus’. The Slovene part of the corpus is representative for the domain of language technologies as explored in Slovenia, while the English subcorpus was built as a comparable corpus to the Slovene part in terms of size, text types and topics (it includes also several articles of Slovene authors writing in English). The total size of the main corpus—the LT corpus—is 44,749 sentences (903,189 word tokens; 1,089,968 including the punctuation) for Slovene, and 43,018 sentences (909,606 word tokens; 1,073,470 including the punctuation) for English.

<b>LT corpus</b>	<b>Slovene</b>	<b>English</b>	<b>TOTAL</b>
Sentences	44,749	43,018	87,767
Tokens (including punctuation)	903,189	909,606	1,812,795
Tokens (without punctuation)	1,089,968	1,073,470	2,163,438

Table 1. Counts of the Language Technologies Corpus in term of sentences and word tokens.

We have not yet released the corpora for public use, since we had not collected the authors’ rights and we have therefore used the collected documents only for personal use and for scientific purposes. However, the Slovene part of the LTC proceedings

corpus, after a detailed preprocessing, was included in the concordancer by the editor of the proceedings and is available at [nl.ijs.si:3003/cuwi/sdjt\\_sl](http://nl.ijs.si:3003/cuwi/sdjt_sl). In further work, we might consider getting the necessary authors' rights to release the entire corpus for public use.

### 3.2.1 Constructing the small LTC proceedings corpus

The small—LTC proceedings—corpus consists of articles published in the language technologies conference *Jezikovne tehnologije*, which has been organized every two years since 1998, as a conference in the scope of the Information Society Multiconference.<sup>11</sup>

The size of the small corpus is 545,641 word tokens (640,095 tokens including the punctuation marks). In more detail, the Slovene part has 330,698 tokens including the punctuation and 279,508 if we count only word tokens (including the digits, abbreviations, etc.). The English part contains 309,397 tokens including the punctuation and 266,133 tokens including only words, abbreviations and digits.

The proceedings are available online (<http://www.sdjt.si/konference.html>). As the articles in Slovene and English are available as PDF documents, we had to transform them into an appropriate raw text format for further processing.<sup>12</sup> PDF to text conversion was performed using PDFBox<sup>13</sup> or Nitro PDF reader<sup>14</sup> and in some cases where these tools did not produce good results, the “Extract PDF text” functionality available in Mac Automator was used. In general, Nitro PDF reader was better for extracting text from older articles and PDFBox for extracting text from newer ones. For a few articles, where none of the tools performed well enough, we contacted the authors to get articles in the Word format. The text files were transformed to UTF-8 encoding. There were still some errors, especially Slovene characters č, š, ž in some documents and PDF specific errors, such as *f* and word splitting at the end of lines that were corrected using the find and replace function.

The corpus was then annotated with several XML tags. Using mainly Perl scripts (but also some manual intervention) we annotated specific parts of the corpus, such as title, abstract, references at the end of the article, tables, authors, footnotes, etc. and added the language identifier tag to Slovene and English articles or parts of articles, respectively. Based on the tagged corpus, we discarded parts of the corpus that we judged to be able to produce noise for our task, such as lists of references at the end of the articles, authors and their institutions, tables (but we kept the table and figure captions), footnote and page numbers, etc. The parts of tables, examples and footnotes were reinserted at the end of each document, in that way not breaking the original sentences in the main text.

Finally two, subcorpora—one containing articles in Slovene and the other in English—were created, in which each article was assigned a unique ID, containing the information about the source (JT- for the proceedings of *Jezikovne Tehnologije*), followed by the year of the publication, the article number, the language as well as the information about the text type: long article (Lart), short abstract (Sabs), as well as title

---

<sup>11</sup> <http://is.ijs.si>

<sup>12</sup> I wish to thank Jasmina Smailović for her help in corpus preprocessing in the frame of the work presented in (Smailović and Pollak, 2011).

<sup>13</sup> <http://www.codeproject.com/KB/string/pdf2text.aspx>

<sup>14</sup> <http://www.nitropdf.com/>

translations in the other language or quoted text in the other language (Stit and Scit, respectively).

To provide the corpus statistics in terms of the number of articles (or short parts of articles) per year, consider Table 2 and Table 3. One can see that there are 109 Slovene articles (Lart) and 81 English ones and each subcorpus can contain also smaller parts of articles, such as abstracts, translated titles and quotations.

Type/Year	1998	2000	2002	2004	2006	2008	2010	Total
Article (Lart)	17	12	28	15	15	12	10	109
Abstract (Sabs)	1	0	0	1	37	0	4	43
Tran. title (Stit)	0	0	0	0	30	0	1	31
Quotation (Scit)	0	0	0	0	0	0	0	0
<b>Total</b>	18	12	28	16	82	12	15	183

Table 2. Counts (# of articles) of the Slovene small corpus covering the Proceedings of the Language Technologies conference. For each year the information about the number of articles and other included units is provided.

Type/Year	1998	2000	2002	2004	2006	2008	2010	Total
Article (Lart)	4	5	9	8	37	11	7	81
Abstract (Sabs)	17	11	4	11	15	12	9	79
Tran. title (Stit)	0	0	0	1	15	0	2	18
Quotation (Scit)	0	0	1	3	0	0	0	4
<b>Total</b>	21	16	14	23	67	23	18	182

Table 3. Counts (# of articles) of the English small corpus covering the Proceedings of the Language Technologies conference. For each year the information about the number of articles and other included units is provided.

### 3.2.2 Constructing the main Language Technologies Corpus (LT corpus)

Since the small LTC proceedings corpus is rather limited in size and text types, we decided to extend it with other types of texts from the same domain, especially with several Bachelor's, Master's, PhD theses, and book chapters. For the choice of articles we<sup>15</sup> proceeded in the following way: besides the LTC proceedings corpus, which is the most representative collection of articles on the topic in Slovene, we first included the *Jezik in slovstvo* journal special issue on language technologies, and next proceeded by searching by key words the national library archive and online collections or contacting authors by mail. The English part of the main corpus was built as a comparable corpus to the Slovene part, also covering articles written by Slovene authors in English.

On the extended corpus, i.e., the Language Technologies (LT) Corpus—that includes the LTC proceedings corpus and is further in the thesis referred to also simply as the corpus—much less preprocessing work was performed than on the more structured small LTC proceedings corpus. We did not use the automatic scripts that we used for cleaning the small corpus, since the main corpus is much more diverse and does not have easily identifiable patterns (sections are not uniformly numbered, abstracts in English that were in the LTC proceedings corpus always one paragraph long can be

<sup>15</sup> This part of the corpus was collected with the help of students Živa Malovrh and Janja Sterle.

much longer, etc.). This corpus was mainly manually processed, but except for deleting the sections in other languages (abstracts, etc.), the corpus still has a lot of noise, which was later also observed as a problem in the definition extraction task.

The total size of the large corpus is 44,749 sentences (903,189 word tokens; 1,089,968 including the punctuation) for Slovene, and 43,018 sentences (909,606 word tokens; 1,073,470 including the punctuation) for English. From these counts one can see that English sentences are on average longer than the Slovenian ones, which can also influence the performance of definition extraction methods. We provide the corpus statistics in terms of text types in pie charts of Figure 1 and Figure 2.

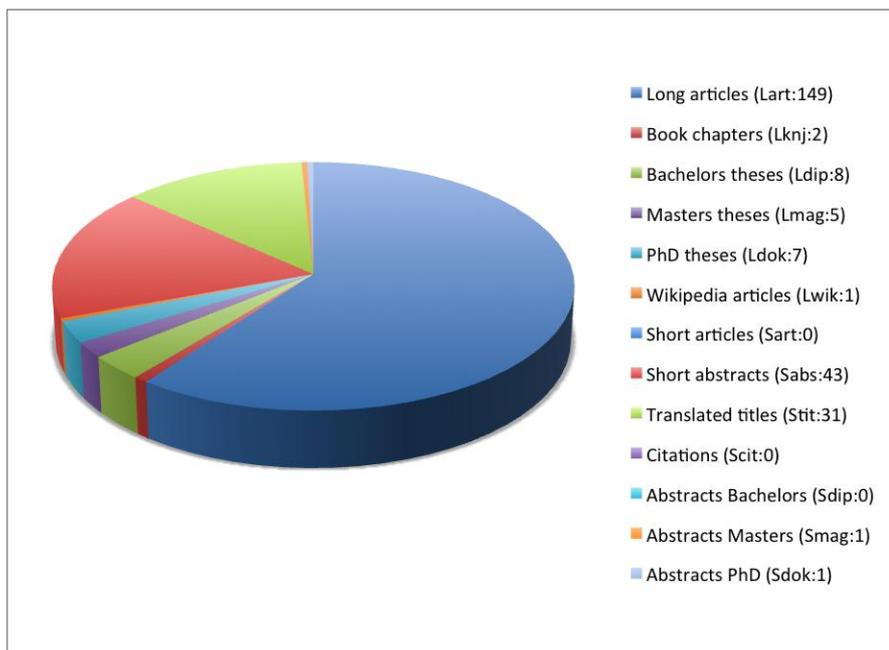


Figure 1. Slovene part of the main Language technologies corpus by text type.

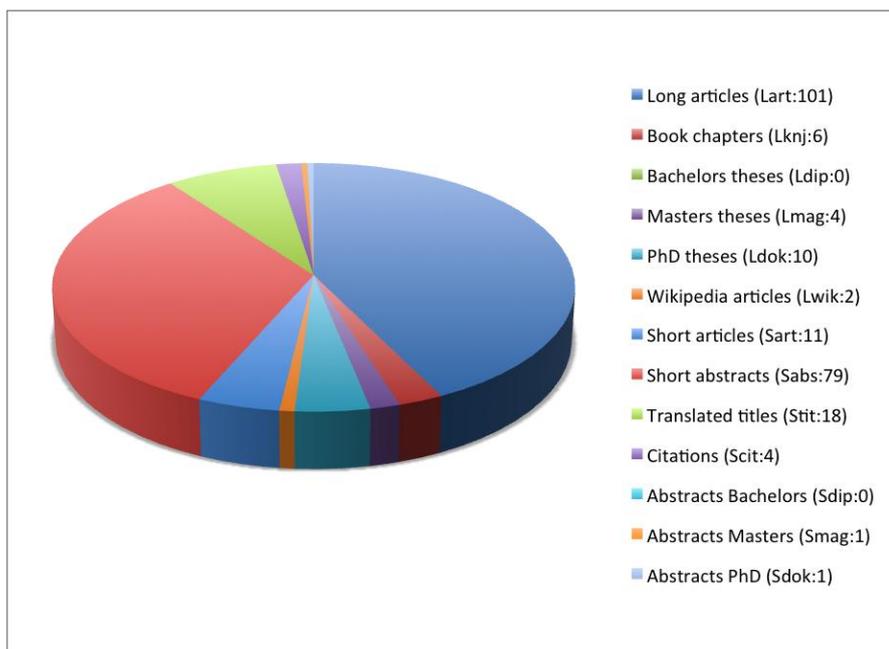


Figure 2. English part of the main Language technologies corpus by text type.

### 3.3 Domain modeling through topic ontology construction

After providing the main information about the corpus in term of its size and text types, we want to get a quick overview about the topics it covers. In order to support initial domain/corpus understanding, we use a simple modeling approach using document clustering. This approach enables initial domain structuring through so-called *topic ontology* construction from documents constituting the given corpus, for which we use the OntoGen topic ontology construction tool (Fortuna et al., 2006a; 2007).

While an *ontology* is a “formal, explicit specification of a shared conceptualization” (Gruber, 1993), represented as a set of domain concepts and the relationships between these concepts, which can be used for domain modeling and reasoning about the domain entities, a *topic ontology* (Fortuna et al., 2006a) is a set of domain topics or concepts—formed of related documents—represented by the most characteristic topic keywords and related by the subconcept-of relationship.

OntoGen (Fortuna et al., 2006a; 2007) is a semi-automatic and data-driven ontology topic ontology construction tool, available at <http://ontogen.ijs.si/>. Semi-automatic means that the system is an interactive tool that aids the user during the topic ontology construction process. Data-driven means that most of the aid provided by the system is based on the underlying text data (document corpus) provided by the user. The system combines text mining techniques (*k*-means document clustering) with an efficient user interface to reduce both the time spent and complexity of manual ontology construction for the user. The result of user-guided document clustering is a topic ontology, i.e., a hierarchical organization of documents’ topics and their sub-topics.

In OntoGen, the hierarchical decomposition of a given set of documents into document subsets is performed semi-automatically by *k*-means clustering. In the *k*-means clustering method, parameter *k* is defined by the user at each step of the multi-layer hierarchical ontology construction process, and each subdomain is described by the main topics (i.e., the *n* most frequent keywords) describing the document cluster.

In this research, we used OntoGen separately on the Slovene and English parts of the corpus, for both the small LTC proceedings and the main Language Technologies Corpus (see also Smailović and Pollak, 2011; 2012).

The motivation for building a topic ontology is as follows. A topic ontology provides an initial idea of the corpus coverage in terms of topics, and in contrast to manually built ontologies, it represents the corpus content and not (or better said to a lesser extent) our perception of it. Moreover, in this type of ontology, the concepts are grounded with documents, meaning that in future work (not yet implemented in this thesis), we can foresee that the extracted glossary could be organized hierarchically, where the terms and their definitions could be attributed to topics and sub-topics (in a thesaurus-like structure).

#### 3.3.1 Modeling the LTC proceedings corpus

When building a topic ontology automatically, by only suggesting to OntoGen the number of clusters *k* at each node of the hierarchy, the result for the English articles can be seen in Figure 3. For every node, we tried different *k*-values and chose the one that splits the document set in the best way according to the user’s understanding of the domain.

As one can see from Figure 3, names of the concepts/topics are not intuitive, and in some cases it is hard to understand the concept that they represent. This happens since

for concept naming OntoGen selects the first three most frequent words from the automatically constructed keywords list. For example, if the concept is described by the following keywords: *slovenian, translation, vowel, speakers, synthesis, speech, corpus, tagging etc.*, OntoGen will name this concept “*slovenian, translation, vowel*”.

A better way of naming the concepts is by involving the user who can quickly find an appropriate concept name after observing all the topic keywords. Using this approach, the previous topic could be called *Speech technologies*. All the concepts in the English and Slovene topic ontologies shown in Figure 4 and Figure 5 thus manually renamed based on the automatically extracted topic keywords.

Moreover, we observed that several topics/concepts were not present in the topic ontology. For the terms often occurring in the keyword lists of different concepts, but not being one of the three main topics keywords, we decided to use the semi-supervised method for adding topics. It is based on the Support Vector Machine (SVM)<sup>16</sup> *active learning* method of OntoGen. For the English corpus, we entered queries for *Speech recognition* and *Speech translation* concepts and answered some automatically proposed questions like “Would you classify document number 41 as an article on the topic of Speech recognition?” which enabled the system to label the instances. After the concept node was constructed, it was added to the ontology as a sub-concept of the selected concept, in our case, as a sub-concept of the *Speech technologies* concept. Similarly, we performed active learning also on the Slovene corpus. We entered queries for *Prevajanje govora (Speech translation)* and based on our confirmation or rejection of automatically proposed articles to be attributed to this topic (active learning), OntoGen learnt which articles should be attributed to the topic and the new subconcept was added to the ontology.

After manually renaming the concepts, using active learning for adding concepts, and manually moving some documents from one concept to another, we got an improved topic ontology. The resulting English topic ontology is shown in Figure 4. This ontology is more intuitive and understandable than the one shown in Figure 3.

One can see from Figure 4 that the Language Technologies Corpus is divided into *Computational linguistics* and *Speech technologies* as its core topics. This is also the general division of the field of language technology (e.g., in Wikipedia, language technology is defined as follows: “*Language technology* is often called human language technology (HLT) or natural language processing (NLP) and consists of *computational linguistics* (or CL) and *speech technology* as its core but includes also many application oriented aspects of them.”). Thus, OntoGen has split the root concept correctly, we just had to change the sub-concepts’ names. More manual work—supervised learning and manually moving some documents from one concept to another—needed to be done in further concept splitting at the lower nodes of the hierarchy.

The Slovene topic ontology (after renaming the concepts, using active supervised learning and by manually moving some documents from one concept to another) is shown in Figure 5. One can see from the figure that the Slovene topic ontology is simpler than the English one. For the Slovene topic ontology we had to do more manual work—moving some documents from one concept to another. Besides the differences in the corpus content itself, it can be partly due also to different preprocessing for English

---

<sup>16</sup> SVM is a supervised learning method which constructs a separating hyperplane in a high-dimensional space of features (words), that has the largest distance to the nearest data points (documents) of different classes. See details of use of SVM for active learning in OntoGen in Fortuna et al. (2006b).

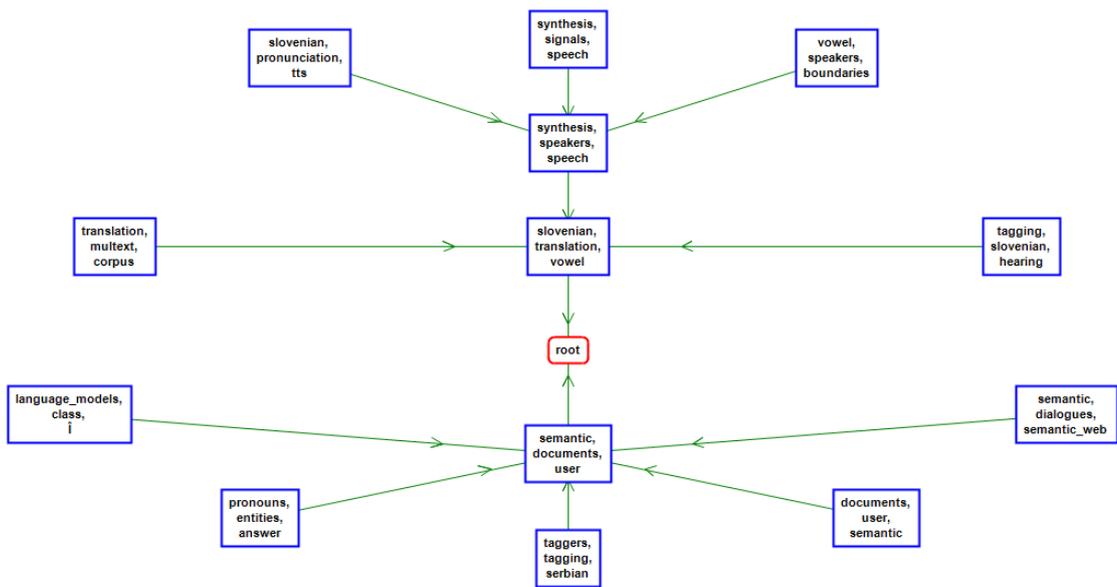


Figure 3. English topic ontology (LTC proceedings corpus) without cleaning the documents and without renaming concepts.

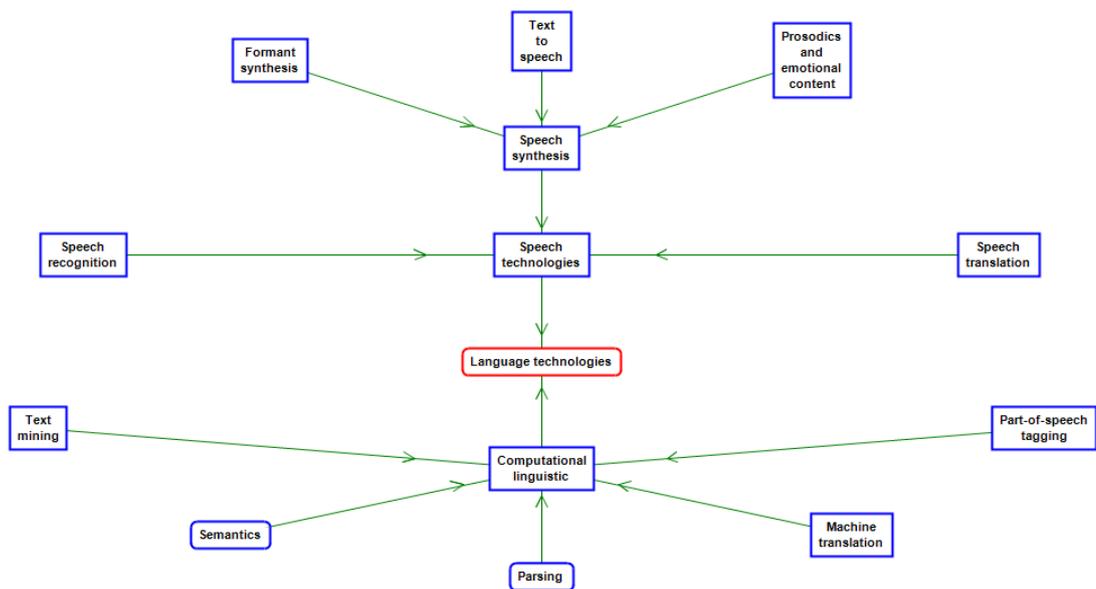


Figure 4. English topic ontology (LTC proceedings corpus) after manually renaming the concepts, using active learning for adding concepts and manually moving some documents from one concept to another.

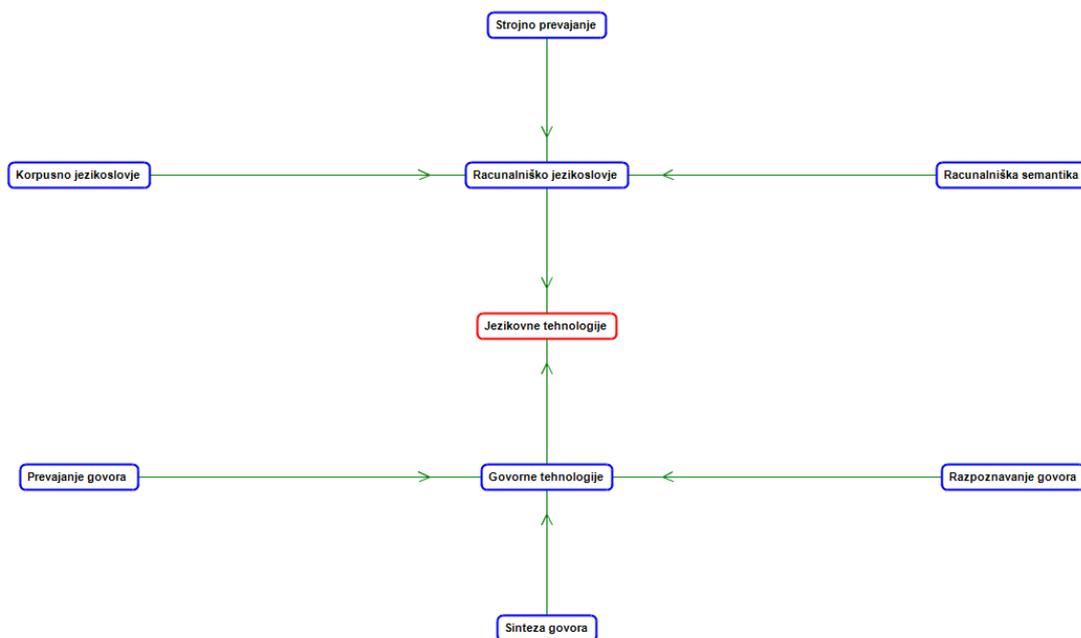


Figure 5. Slovene topic ontology (LTC proceedings corpus) after manually renaming the concepts, using active learning for adding concepts and manually moving some documents from one concept to another.

and Slovene. Because OntoGen does not provide stemming for Slovene, we lemmatized the input text documents in the data preprocessing step. As for the stop-word removal, the lists are included in the OntoGen and there might be some differences as well.

Interestingly, one third of the Slovene articles belong to the topic *Korpusno jezikoslovje* (*Corpus linguistics*). Note that the *Corpus linguistics* category in our ontology comprises also the compilation of corpora and development of tools for corpus processing (such as PoS taggers) and is not used in its strict sense, but in the sense of *Corpus construction and use*.

Concept visualization in the form of a *Document Atlas* (Fortuna et al., 2006c) is another functionality of OntoGen. It is based on using dimensionality reduction for document visualization by first extracting main concepts from documents and then using this information to position documents on a two dimensional plane via multidimensional scaling. Documents are presented as crosses on a map and the density is shown as a texture in the background of the map (the lighter the color, the higher the density). The most common keywords are shown for the areas around the map and therefore the same keyword can occur more than once.

Document visualization for English articles can be seen in Figure 6. Two main concepts are marked with green and orange dashed lines. In the upper left corner of concept visualization one can notice some non-standard characters. These show OntoGen's encoding problems probably due to the Slovene characters which are present in authors' names and references or some special characters from the tables.

Concept visualization for Slovene articles can be seen in Figure 7. The visualization is similar to the visualization for English articles, i.e., the articles are divided into two major topics (*Computational linguistics* and *Speech technologies*) and the *Computational linguistics* topic contains much more articles than the *Speech technologies* topic.

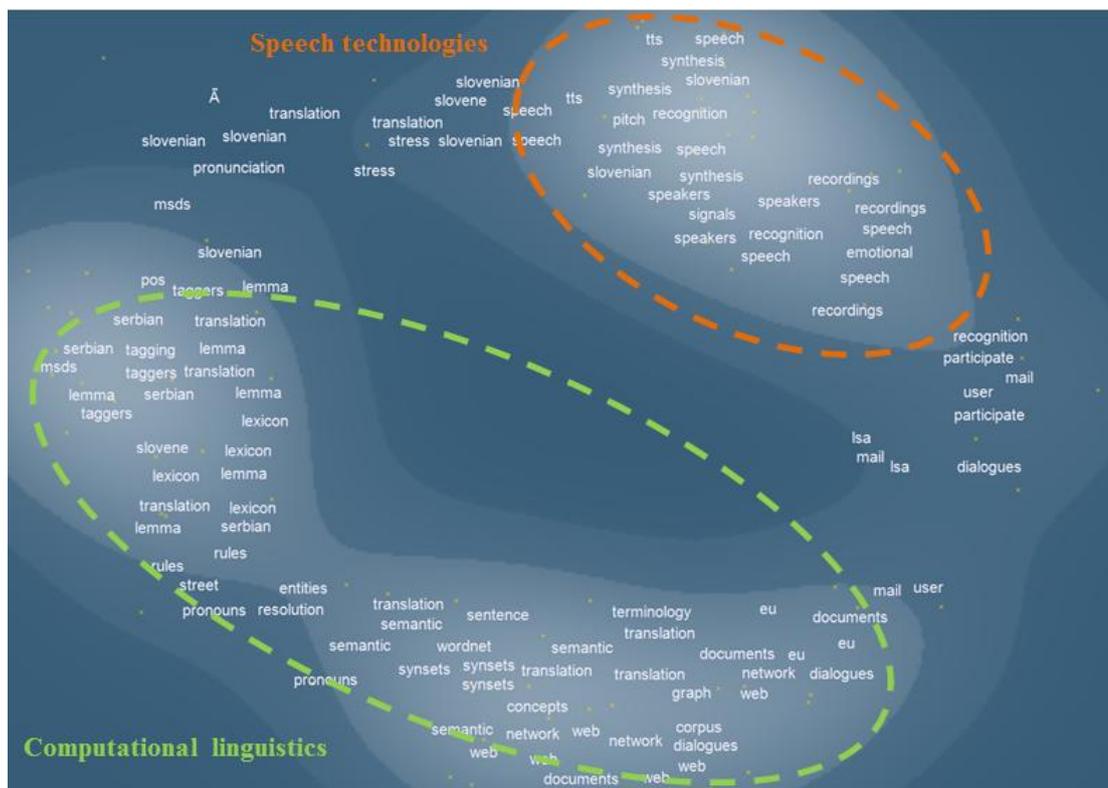


Figure 6. Concept visualization of English text documents (LTC proceedings corpus). The automatic splitting into two main topics, which we label computational linguistics and speech technologies, can be observed.

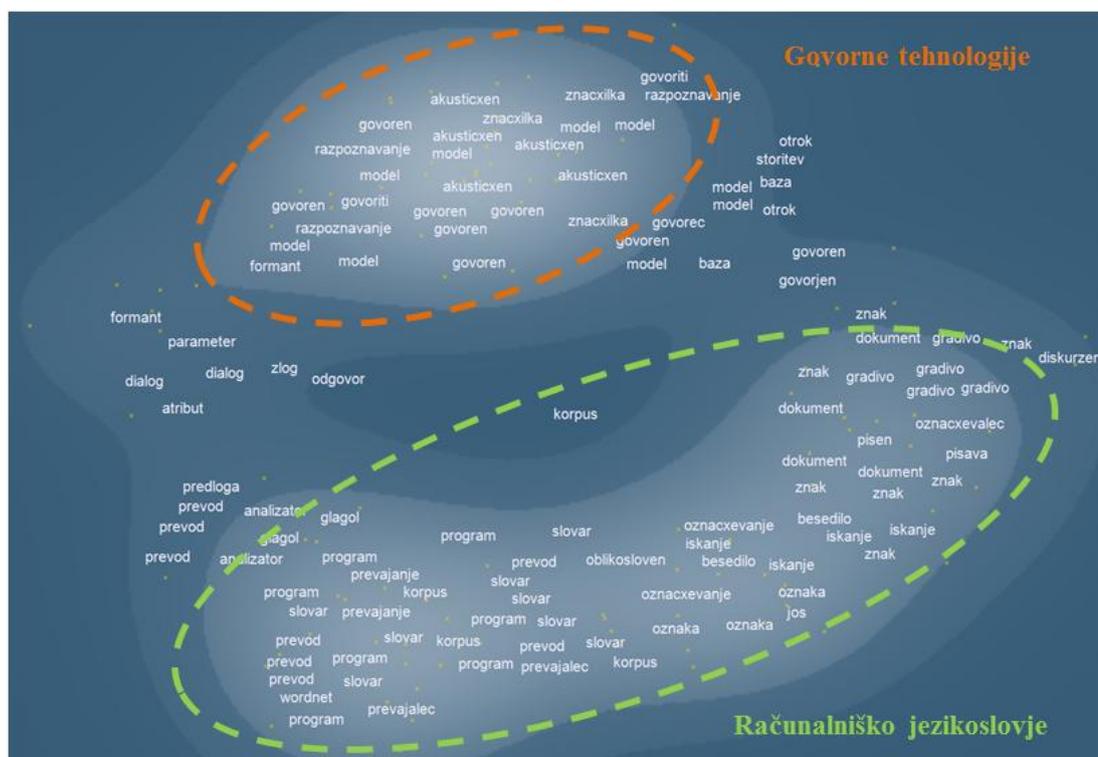


Figure 7. Concept visualization from Slovene text documents (LTC proceedings corpus). The automatic splitting into two main topics that we label Računalniško jezikoslovje and Govorne tehnologije can be observed.

### 3.3.2 Modeling the Language Technologies corpus

the main corpus, we used all the methods described above. We manually moved several documents from one category to the other, renamed the concept expressed by keywords and used the active learning option. There was much more manual work needed than with the small corpus, since the categories were less clear. We also used the document atlas as help in semi-automatic ontology construction. We attribute this to the fact that the main corpus is much more heterogeneous, the length of the documents varies from very short abstracts or even sentences from translated titles to entire doctoral dissertations. The created topic ontologies of Slovene and English language technologies domains are shown in Figure 8 and Figure 9.

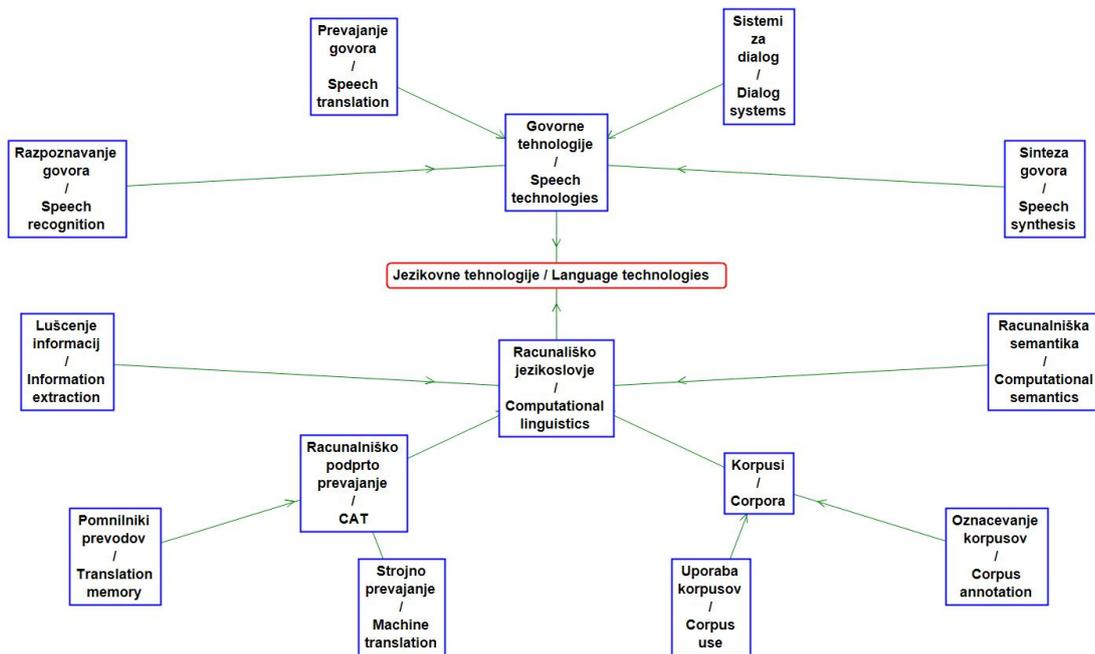


Figure 8. Topic ontology constructed from the entire Slovene LT corpus.

A precise evaluation of the ontology coverage is a very hard task. A test that we will consider in further work is to compare an ontology created with OntoGen to a manually created ontology from scratch by two equally competent experts. This type of experiment could tell whether we gain time and if OntoGen finds all the concepts present in the corpus which the expert finds. Note that OntoGen was itself evaluated in the experiments led by its developers (Fortuna et al., 2007) and proved to be good help in the human ontology construction process.

At this place we propose an evaluation of the estimated coverage. Since we do not have a gold standard ontology for this specific corpus, we were therefore only able to approximately evaluate the coverage of the research area of language technologies separately for individual topics, as illustrated on the following sub-topic. In Wikipedia, the concept of *Speech technologies* is divided into 6 subfields (*Speech synthesis*, *Speech recognition*, *Speaker recognition*, *Speaker verification*, *Speech compression*, *Multimodal interaction*). In the main English corpus we can see that two (*Speech synthesis*, *Speech recognition*) out of these six concepts are covered by the ontology, moreover there is the topic of *Dialog systems* which can be understood as partly

covering the topic of *Multimodal interaction*. The remaining uncovered topics are therefore *Speaker recognition*, *Speaker verification* and *Speech compression*. *Speaker recognition* is indeed a missing concept, since it occurs 9 times in the main corpus; it seems that the two related sub-domains (*Speech recognition* and *Speaker recognition* were in fact merged into a common category). On the other hand, *Speaker verification* (even if the concept itself is highly related to *Speaker recognition*) and *Speech compression* do not figure in the corpus more than once, meaning that the constructed ontology adequately reflects the nature of the corpus. In contrast, the concept of *Speech translation* that was present in the small corpus ontology, does not exist in the main corpus ontology and is therefore indeed missing in the ontology. Another concept present in the small corpus and not in the main corpus is *Text mining*, which is related to *Language technologies* but is not its direct sub-concept. However, the lesson learnt is to take into consideration all the concepts that were identified in the smaller corpus and inspect them as candidates for active learning on the larger corpus.

As already claimed, the ontologies we presented are topic ontologies, where instances are documents, classes are topics and subtopics, and relations are hierarchical topic/sub-topic relations and not any other type of relations. This is a different approach than e.g., the one of Navigli et al. (2011), where taxonomies are created from sentences in the documents, searching for terms and their hypernyms and connecting them to a taxonomy.

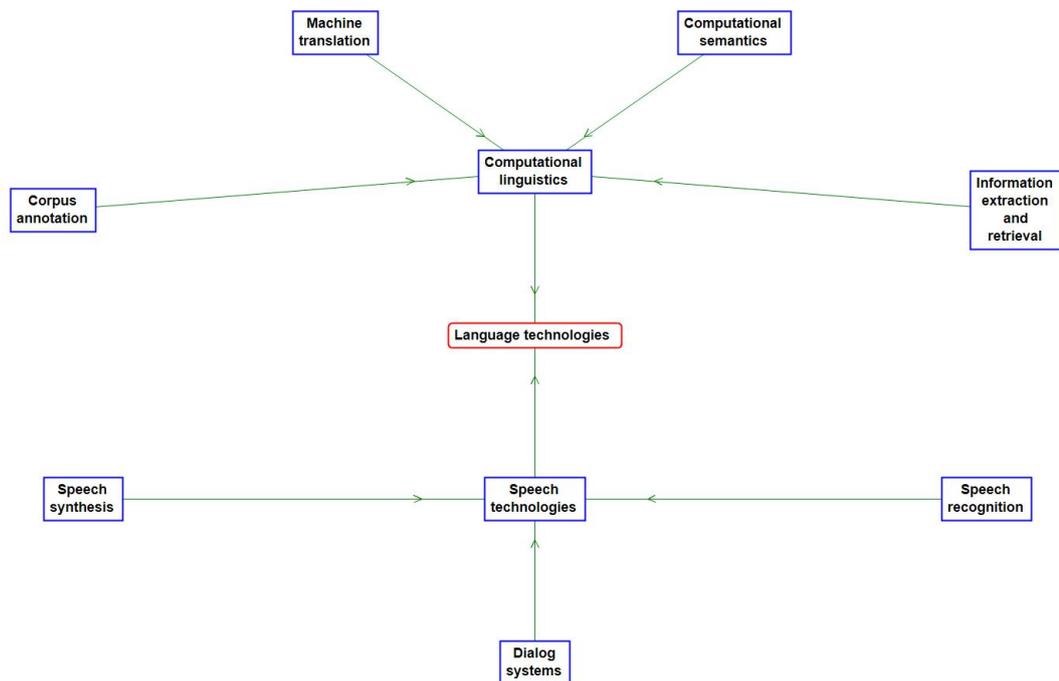


Figure 9. Topic ontology constructed from the entire English LT corpus.

In this section we obtained the initial insight into the language technologies domain topics and subtopics. In the core part of this thesis—focusing on definition extraction—we thus expect to extract the definitions covering these topics. Alternatively, this topic ontology construction phase could be used to determine the missing subtopics and enlarge the corpus adequately. In further work, we may also consider using the selected categories for hierarchically organizing the final glossary.

### 3.4 Setting the stage for automatic definition extraction: Analyzing definitions in running text

As defined in Section 3.1, the main task in this dissertation is the extraction of candidate definition sentences from Slovene and English unstructured text, applied to the domain of language technologies, focusing on the linguistic analysis of different kinds of definitions figuring in the corpus. We explicitly refer to unstructured texts, in order to highlight the differences with the tasks where the knowledge is extracted from (semi-)structured resources, such as (online) dictionaries, encyclopedias or Wikipedia. The main difference is that when extracting definitions from unstructured text, we focus on full sentence definitions, where the definitions are often embedded in longer sentences. Moreover, unless we deal with special collections of articles, one cannot use the position information (or at least not in a straightforward way), which is a very important factor when searching for definitions in Wikipedia articles, in which in most cases the first sentence is a definition. As linguistic analysis is of paramount importance for this thesis, this section discusses the types of definitions encountered in scientific articles in the language technologies domain, showing the complexity of the task addressed; even defining what is a definition itself is a non-trivial problem.

As discussed in Chapter 2, there is little agreement even among linguists, terminologists and domain experts of what constitutes a valid definition of a concept, and even less on the distinction between a definition and an explanation on the one hand, and a definition and a defining context on the other. Definitions can cover in the strictest view only sentences with a *genus* and *differentia* structure (i.e. *analytical definitions*), while in other views many other definitional types are accepted, such as *extensional*, *functional*, *relational* or *typifying definitions* to name but a few. In this section, we analyze a subsample of definition sentences to see how the categories discussed in Section 2.1.5 are applicable to our corpus.

The most straightforward way for defining a term is, as already discussed by Aristotle,<sup>17</sup> to provide the hypernym of the concept (*genus*) and the *differentia* that distinguishes the species from other instances from the same *genus*. This definition type assumes that the term to be defined (*definiendum*) is always a species and not an individual.

However, for (semi-)automatic definition extraction, limitation to this type of definitions could be insufficient. If we consider very specialized domains, especially in languages other than English (which is *lingua franca* for a large amount of scientific production), we often have quite limited resources for a specific domain. Also our corpus contains a limited collection of articles from a specific domain, using mainly highly specialized scientific discourse characterized by the fact that a high level of prior knowledge is presupposed, basic terms are considered common knowledge, additional information is provided through references to related work, and not through definitions and explanations. Therefore, one can expect that the terms will be defined in the text with many other defining strategies than only by using the above-mentioned *genus-differentia* definition type. Even if other types of defining a concept are less clear-cut cases and one has more difficulty to decide whether the sentence is a definition or not, considering other definition types is necessary for the extraction of definitions from

---

<sup>17</sup> Aristotle's *Metaphysics*. Stanford Encyclopedia of Philosophy, <http://plato.stanford.edu/entries/aristotle-metaphysics/>, (Last accessed: February 3, 2013).

completely unstructured resources. Therefore, in this thesis, we decided to use the broader interpretation of definition than the one defining the term by its genus and differentia, and accepted a large variety of definition types. We did not use any a priori condition of formal structure of a definition sentence, and agreed with the interpretation that “as there is no formal terminological status for meaning, there is also no formal designation, meaning or definition of definition (Dolezal, 1992, p. 2; p. 103). We simply evaluated as positive candidates the sentences that could be used (in their original or possibly manually edited form) for building a *glossary*, i.e., a list of terms particular to a field of knowledge (extracted from the corpus) with their definitions.

To get an initial insight into the kinds of definitions that can be found in a running text, we manually annotated a sample of our corpus with “definition” vs. “non-definition” tags. In this section we analyze and discuss different definition types, as well as the borderline examples.

Below we enumerate and provide examples for different definition types into which we manually categorized the selected set of definitions from the Slovene subcorpus, and discuss their specificities. All the examples are authentic sentences from the corpus. For each of them we provide an English translation. Note that in some cases the translation does not entirely reflect the discussed issue (e.g., when the word order of a Slovene sentence is discussed) and therefore we add a comment or a literal translation in the footnote.

The examples in this section illustrate the difficulty of the task of definition extraction from the given specialized corpus addressed in this thesis, as the definitions in this corpus are substantially more complicated than the ones in simpler corpora such as textbooks or semi-structured resources including Wikipedia.

### 3.4.1 Genus et differentia definition type

#### Genus et differentia: verb *be*

This category comprises formal definitions with the *genus et differentia* structure (“X is Y”). In Slovene, the definiendum (X) and the *genus* term (Y) are noun phrases in the nominative case (see Example (i)). The determinants (*a*, *the*), which can be good indicators for English structures (“a/the X is a/the Y, which...”) are not used in Slovene. Verb *be* can be realized as third person singular, dual or plural form. The dual verb form is specific to Slovene and very few other languages and is used to refer to two persons only. Third person dual form of verb *be* is *sta* and the plural form is *so*.

- i. *Lombardov efekt je pojav, pri katerem govorec poveča glasnost govora ob povečanju glasnosti šuma ozadja.*  
*[The Lombard effect is a phenomenon, where a speaker increases the intensity of the speech when the level of background noise raises.]*<sup>18</sup>

This sentence has a typical definition structure: the term to be defined, i.e., *definiendum*, is a noun phrase (*Lombardov efekt*, where *Lombardov* is an adjective and *efekt* is a noun

---

<sup>18</sup> Note that as in Slovene the determinants *a* and *the* are not used, a literal translation of the beginning of the sentence would thus be *Lombard effect is phenomenon /.../*. In order to facilitate reading of the translated sentences, we have opted for non-literal translations, when appropriate.

in the nominative case), followed by a third person present tense of copula verb *biti* [*be*] (*je* [*is*]) and a *genus*, also a noun in the nominative case *pojavn* [*phenomenon*]. It is followed by the *differentia* part, i.e., the part distinguishing the definiendum from other instances in the *genus* category, in a relative clause.

Other variations of this definition type are:

- Definiendum is followed by the term's English translation (Latin or Greek translations are less common in our corpus). Translation of a term can be explicitly introduced by abbreviation *angl.* or *ang.* [*Eng.*] as in Example (ii) or just using parentheses (see Example (iii)). Sometimes, a Slovene synonymous expression for a term is provided either between parentheses or separated by a conjunction *ali* [*or*] (cf. iv).
  - ii. *Branje ustnic (angl. lipreading) je vizualna percepcija govora, ki temelji izključno na opazovanju artikulatornih gibov (ustnic) govorca brez poslušanja.*  
*[Lipreading (Engl. lipreading) is a visual perception of speech that is based exclusively on the observation of articulatory movements (of the lips) of the speaker without listening to him.]*
  - iii. *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.*  
*[Part-of-speech tagging (part-of-speech tagging or word-class syntactic tagging) is a process where each word occurring in a text is assigned a part-of-speech tag.]*
  - iv. *Leksikalne ontologije ali semantične mreže so sistemi kategorij med besednimi in pojmovnimi sistemi (med urejenimi glosarskimi, slovarskimi, tezaverskimi, enciklopedičnimi in terminološkimi zbirkami), ki obe vrsti sistemov povezujejo in se naslanjajo predvsem na taksonomsko organizirano zgradbo pojmov nekega področja.*  
*[Lexical ontologies or semantic networks are systems of categories between word and conceptual systems (between ordered glossary, dictionary, thesaurus, encyclopedia or terminology collections) that connect both types of systems and are based especially on a taxonomic organization of concepts of a given domain.]*
- The *genus* noun phrase can also contain general words, such as *expression*, *word*, *kind* or *type* (see Example (v)). Therefore, a term instead of being followed by a copula and a hypernym noun phrase, the copula is followed by an introducing expression before the real hypernym. We can put in this category also the word *part*, whereby the “part\_of” relation is a meronymy and not a hypernymy relation.
  - v. *Besedna poravnava (Word Alignment) je izraz za tehnologijo statističnega pridobivanja leksikonov prevodnih ustreznih iz vzporednih korpusov.*  
*[Word alignment (Word Alignment) is an expression for the technology of statistical acquisition of lexicons of translation equivalents from parallel corpora.]*

- The demonstrative pronoun can precede the *genus* noun phrase, like in the examples below:

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

*[The root is that<sup>19</sup> part of the word that bears the concept meaning, while the ending is the one that describes the grammatical properties of the word.]*

vii. *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 /.../«.*

*[Agreement is “/t/hat type of grammatical, syntactic link between words of the noun phrase or between sentence elements in which a subordinate word (or sentence element) agrees in case, number, person or gender with its main constituent /.../”.]*

- The definiendum does not necessarily occur at the beginning of the sentence. In Example (viii) it is the word *sopomenke* [synonyms] that is defined by its hypernym *words*, placed at the beginning of the sentence.

viii. *Besede so sopomenke, če jih je v besedilu mogoče zamenjati, brez da bi pri tem spremenili njegov pomen. (sic!)<sup>20</sup>*

*[Words are synonyms, if in the text they can be replaced without changing its meaning.]*

- The definition scope can be explicitly limited to a domain, authors' theory or just a specific interpretation. In these cases, if the definition scope is provided at the beginning of the sentence, the copula verb comes first, and then the word being defined and the *genus*.

ix. *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.*

*[Both in philosophy and in computer science, the ontology is by definition a presentation of entities, ideas and events, together with their properties and relations – regarding the chosen system of categories.]<sup>21</sup>*

x. *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.*

*[As data structures, semantic networks are directed graphs, where concepts are presented by points or nodes, and the relations by arrows or links between them.]<sup>22</sup>*

<sup>19</sup> In natural English text one would use definite article the instead of demonstrative article that. This comment refers to both examples (vi and vii).

<sup>20</sup> *brez da bi* is a very common error in Slovene; correct grammatical construction is *ne da bi*. Another mistake in the sentence is the use of *njegovega* [its] instead of *njihov* [their].

<sup>21</sup> In Slovene the word order is as follows In philosophy and in computer science, is the ontology by definition a presentation of entities, ideas and events /.../.

## Genus et differentia structure: verbs other than verb *be*

Definitions with *genus et differentia* structure, where the verb is other than the verb *biti* [be]; alternative verbs may include *definirati, imenovati, opredeliti, meriti, predstavljati, biti znan pod imenom, veljati za, nanašati se, govoriti o* [define, designate, measure, present, be known under the name, be considered as, refer to, speak about]. The definiendum, *genus* and *differentia* do not necessarily occur in this order and all the other variations mentioned above are possible. Consider Example (xi) where the definition scope is limited to the field of *literary studies*.

- xi. *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.*

*[In literary studies, critical editions denote the editions, in which the text is checked, transcribed, reconstructed, commented and finally published according to the principles of textual critics or ecdotics as supporting literary discipline.]*

In the example above, verb *imenovati se* [be called, designate] is a reflexive verb, and thus the structure is even more complicated since the particle *se* is separated from the main verb part. If we analyze the elements of this sentence in more detail we observe the following:

[Znanstvenokritične izdaje]<sub>definiendum</sub> [se]<sub>pronoun of reflexive verb</sub> [v literarnih vedah]<sub>domain(scope)</sub> [imenujejo]<sub>verb</sub> [tiste]<sub>demonstrative\_pronoun</sub> [edicije]<sub>genus</sub>, [v katerih so besedila pregledana, prepisana, rekonstruirana, komentirana in naposled objavljena po načelih tekstne kritike ali ekdotike kot pomožne literarnovedne discipline]<sub>differentia</sub>.

[Critical editions]<sub>definiendum</sub> [<sub>pronoun of reflexive verb</sub>] [in literary studies]<sub>domain(scope)</sub> [designate]<sub>verb</sub> [those]<sub>demonstrative\_pronoun</sub> [editions]<sub>genus</sub>, [in which the text is checked, transcribed, reconstructed, commented and at last published under the concepts of textual critics or ecdotics or other literary field]<sub>differentia</sub>.

Consider Example (xii) concerning the verb *biti predstavljen* [be presented] and Example (xiii) for *biti definiran* [be defined]:

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

*[In classical theory (Katz and Fodor 1963), the meanings of words are presented as sets of necessary and sufficient conditions that include their conceptual meaning, explicated by words.]<sup>23</sup>*

---

<sup>22</sup> Same as in the example above, literal translation of the beginning of the sentence is As data structures, are the semantic networks directed graphs /.../.

<sup>23</sup> The word order in Slovene is as follows: In classical theory (Katz in Fodor 1963) *are* the meanings of words presented as sets of necessary and sufficient conditions, that /.../, meaning that the copula verb precedes the definiendum and the definiens (words serving to define the definiendum).

- xiii. *V tej shemi so diskurzni označevalci definirani kot izrazi, ki k vsebini diskurza ne prispevajo nič ali skoraj nič, pojavljajo pa se v naslednjih pragmatičnih funkcijah: -vzpostavljanje povezave z vsebino prejšnjega oziroma sledečega diskurza, - vzpostavljanje in razvijanje odnosa med sogovorniki, - izražanje odnosa govorca do prejšnje oziroma sledeče vsebine diskurza, - organiziranje poteka diskurza na ravni prehodov med temami pogovora, menjavanja vlog in strukture izjave.*

*[In this schema the discursive markers are defined as expressions that contribute nothing or nearly nothing to the discourse content, but appear in the following pragmatic functions: -forming a connection to the content of the preceding or following discourse, -building and developing the relation between co-speakers, -organizing discourse development regarding topic switching, role changing and utterance structure.]<sup>24</sup>*

In the example below *veljati* [*be considered*] and *biti znan pod imenom* [*be known as*] are used.

- xiv. *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.*

*[Ontology is traditionally considered as a branch of philosophy and was for a long time known under the name metaphysics, it considers the questions of so called entities that exist or are considered as existing; how these entities are grouped into larger classes and how they are hierarchically classified in these classes in terms of similarities and differences.]*

- xv. *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.*

*[According to the linguistic tradition, the concept of symbolic prosody refers to speech properties that do not refer only to one phonetic segment, voice, but to larger entities, that include several phonetic segments, such as words, phrases, clauses or even larger parts of spoken text.]*

In Example (xvi) the expression *govorimo o* [*we talk about*] is used.

- xvi. *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).*

*[When the ambiguity is concerned,<sup>25</sup> where different meanings of words are connected, we talk about polysemy (e.g., mouse can be a part of computer or a rodent).]*

### **Genus et differentia: without a verb**

This category does not correspond to the formal structure “X is Y” but still defines the concept by the same strategy (using hypernym and the *differentia*). The link between the

---

<sup>24</sup> Same as in the example above, the copula verb occurs immediately after the introductory part In this schema are the discursive markers defined as /.../ .

<sup>25</sup> The original Slovene structure starts with When “*it goes about*” ambiguity, /.../.

*definiendum* and the *definiens* (the defining part of a sentence) is not a defining verb, but the *definiens* is provided in an embedded clause, as part of a sentence that is itself not necessarily a definition. For this type, generally some manual refinement is needed. In the first example below, we have the definition introduced by *torej* [*hence*], while the second one is just the apposition without any introductory element. In both examples *oziroma* is also used to introduce two alternative synonymous expressions: *izhodiščni oziroma prvi jezik* [*source or first language*] and in the second example *leksikalne enote oziroma leksemi* [*lexical units or lexemes*].

- xvii. *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).*

*[In corpora of foreign language learning, the target language, hence the language "that someone learns with the purpose to master it as his first, second or foreign language" (Pirih Svetina 2005), and the source or first language, "from which the person learns all the other languages" (ibid.), are both important.]*

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

*[The core of every semantic collection, and of the present study, are the lexical units or lexemes, the main building blocks of meaning in a language.]*

Informal definitions can be provided also in parentheses, as illustrated below.

- xix. *Pri tem pristopu naletimo na posebnosti, ki jih lahko razdelimo v dve skupini: leksikalne vrzeli (pojma, 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 izvirnega jezika, vendar je nekoliko splošnejša ali nekoliko bolj specifična).*

*[In this approach we encounter two groups of special cases: lexical gaps (concepts that are in one language expressed with a lexical unit are in the other language only possible to express using a free combination of words) and denotational differences (in the target language there is a translation equivalent of the source language concept, but it is somewhat more general or more specific).]*

### **Proper nouns (named entities)**

All the possibilities mentioned above can be used for defining a named entity. This does not question the formal structure of a definition, but its semantics. Whether a named entity should be considered as a term to be defined or not, depends on the final application.

- xx. *FIDA je referenčni korpus slovenskega pisanega jezika in obsega 100 mio besed iz različnih tekstovnih virov iz obdobja 1990-1999.*

*[FIDA is the Slovene written language reference corpus and comprises 100 million words from different text sources from the period 1990-1999.]*

### 3.4.2 Defining by paraphrases, synonyms, sibling concepts or antonyms

As noted in Section 2.1.5, even if in lexicography the analytical (*genus-differentia*) definition type has a prestigious status of being ‘the best definition type’, alternative methods should also be considered. While in definition extraction research the most common focus is on analytical definitions, the analysis of the corpus definitions presented in this section (and in further experiments of this thesis) shows that the corpus definitions cover a large variety of definition types. In contrast to previous category of analytical *genus-differentia* definition, this section presents definitions, defining a term by means of synonyms and paraphrases, sibling concepts or other related terms such as antonyms.

#### Paraphrases and synonyms

Some of the terms in the Language Technologies Corpus are defined by paraphrases or synonyms. All the definitions that define a new term in relation to other terms might be problematic regarding circularity. Consider the example below, where the term *enopojavnica* [unique word] is defined through a synonym *hapax legomena*. If such a synonym is not defined in the same sentence or in the same collection or is not part of general knowledge, we get a cyclic structure where the term is in fact not defined at all. The definition of the synonymous term should therefore be (possibly manually) added into the resulting domain dictionary.

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

[Unique words in corpus linguistics are also called hapax legomena and represent an interesting domain of research].

#### Sibling concepts

Sibling concept definitions can be understood as what E. Westerhout (2010, p. 37 referring also to Borsodi, 1967) introduces as *analogic* definitions, which are in their classification a subtype of synonymous definitions (e.g., *Hyves is something like Myspace*)<sup>26</sup>. Similar to the synonymous definition type, circularity can be a problem; however if the sibling concept is part of general knowledge, the definition is less problematic than if both terms were domain specific. In a definition of a term, a sibling concept can be used alone (as in the above example of Westerhout). Alternatively it can be used in combination with other defining techniques where sibling concept can for instance replace the *genus* of a *genus-differentia* structure, or be used in addition to defining by *genus-differentia*, paraphrases, etc. (see Example xxii).

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

[Wikidictionary is a project similar to Wikipedia and is a free multilingual dictionary with definitions, etymologies, pronunciation and citations.]

In Example (xxiii) below, *classical dictionaries* are used as a sibling concept, but the *differentia* is stated by means of functional defining strategy (see below).

---

<sup>26</sup> It is debatable whether in this sentence the term Hyves is actually defined.

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

*[In contrast to classical dictionaries, semantic collections define the meaning of a word according to its relation to the meanings of other words.]*

### Antonyms and other relational definitions

In addition to synonyms or sibling concepts, other semantic relations can be used in definition sentences. In Section 2.1.5 we have already introduced *antonymic* and *meronymic* subtypes of relational definitions (where meronymic definitions as defined by Borsodi (1967) and Westerhout (2010) situate a term between two other terms). The next sentence is an example of relational definition by explaining that the term *hypernymy* is an inverse relation of *hyponymy*. However, as already noted, this is a circular definition since one term is defined by another term.

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

*[The most common relation is hypernymy, and with it also its inverse relation hyponymy.]*

The cases containing *circulus in definiendo*—meaning that one unknown term is defined by means of another unknown term within the sentence or within the collection of definitions—are borderline definitions. The circularity is known as problematic when defining a concept, and ideally these sentences should not be considered as valid definitions if e.g., *hyponymy* in the example above, is not defined by other means in the same collection. However, to some extent circularity is, as it was discussed in Chapter 2, acceptable and unavoidable. Moreover, in tasks addressing (semi-)automatic extraction of definitions from a specific domain consisting of very limited resources, one can consider these sentences as borderline cases, but still sort of definitions.

### 3.4.3 Extensional definitions

In the examples above we discussed *intensional definitions*, which provide the meaning of a term by typically specifying the necessary and sufficient conditions for belonging to the set being defined. On the other hand, *extensional definitions* define a term by enumerating the objects (all or typical examples) that fall under the term in question. In Section 2.1.5 we have enumerated several types of extensional definitions. Since in our setting, ostensive definitions are not relevant, we use the term *extensive definitions* mainly to refer to citational extensional definitions or partitive-concept definitions. Instead of specifying the hypernym, extensional definitions list (all/typical) realizations of a concept.

In the example below, different taxonomical relations are listed (*hypo-* and *hypernymy*, *meronymy*, *holonymy*, *troponymy*) and even some of these concepts are additionally defined within the sentence or illustrated by examples.

- xxv. *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).*

*[Besides hyper- and hyponymy, other taxonomical relations are meronymy and holonymy that express the relation between a part and a whole (e.g., steering wheel ↔ car), and troponymy between verbs that connects verbs based on the type of realization (e.g., to speak ↔ to whisper) (Fellbaum 2002).]*

### 3.4.4 Other types of definitions: defining by purpose or properties

#### Defining by purpose (functional definitions)

In this definition type the term can be defined without a hypernym: the term is defined by its purpose, why it is used, etc. This is illustrated in Examples (xxvi) and (xxvii).

- xxvi. *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.*

*[Lexical semantics deals with the meaning of words and investigates different aspects of word meaning, that are realized in typical (or untypical) usage in grammatically appropriate contexts.]*

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

*[Speech synthesizer can transform any Slovene text into comprehensible computer speech.]*

Note that the last definition is too specific because a *speech synthesizer* can transform any text (not Slovene text specifically) into comprehensive computer speech. These kinds of examples are borderline cases, since they need manual refinement.

Like in the *genus et differentia* examples, several varieties were noticed, such as naming the author of a definition, as shown in Example (xxviii) below.

- xxviii. *Po Corazzonu ponuja ontologija merila, ki razlikujejo med seboj različne vrste stvari (konkretne od abstraktnih, obstoječe od neobstoječih, realne od idealnih, neodvisne od odvisnih ...)*

*[According to Corazzon, an ontology provides the measures, with which different types of things can be differentiated (concrete from abstract, existing from non-existing, real from ideal, independent from dependent...)]*

Verbs such as *morajo* [*must*] or *skušajo* [*try*] can be used in these definition types.

- xxix. *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.*

*[Programs for part-of-speech tagging must assign to a word form all possible part of speech tags, and then choose the right one among them based on the context in which this word form occurs.]*

- xxx. *V zadnjem času so na področju računalniške obdelave naravnega jezika izjemno popularne statistične metode, ki z modeliranjem jezika s pomočjo velikih količin podatkov, dobljenih iz korpusov, in strojnem učenjem skušajo zaobiti potrebo po dragem in zamudnem ustvarjanju semantičnih virov.*

*[Recently, in the domain of computer processing of natural language, statistical methods have become very popular, as they model the language with the help of large amounts of data, and using machine learnin they try to overcome the need of expensive and time-consuming creation of semantic resources.]*

- xxxi. *Reprezentativnost je sicer relativna kategorija, saj je nemogoče predvideti in v korpus zajeti vse besedilne variante, vendar pa se skuša z merili reprezentativnosti zajeti vsaj ključne, ki pa morajo vključevati čim več jezikovnih variant.*

*[Representability is a relative category as it is impossible to foresee and include all text variations into the corpus, however with the measures of representability at least the most important ones try to be included, representing as much language variations as possible.]*

### **Defining by properties (typifying definitions)**

Also in this definition type the hypernym can be omitted. In the example below, the *antonyms* are defined by their property, while the expected hypernym *word* is not expressed (in contrast a formal *genus-differentia* definition of word *antonym* might begin with “*Antonyms are words that...*”). Even if these kinds of definitions can be considered incomplete, due to a missing hypernym, they are certainly good candidates for automatic extraction and further manual refinement.

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

*[It is characteristic for antonyms that they have in common the majority of elements of the meaning, with the difference that they take the extreme value of a certain dimension (e.g., hot ↔ cold).]*

Sometimes the limit between defining by purpose and defining by properties is not very clear, as in Example (xxxiii) where word *lastnost* [*property*] could be easily substituted by word *vloga* [*role*] (“*their main property is to establish and develop the relationship between co-speakers*” could as well be expressed as “*their main role is to establish and develop the relationship between co-speakers*”).

- xxxiii. *Tukaj za označevalce pragmatične strukture uporabljam izraz interakcijski označevalci, saj lastne predhodne raziskave (Verdonik et al., 2007; Verdonik et al., v tisku) kažejo, da je njihova osrednja lastnost vzpostavljanje in razvijanje odnosa med sogovorniki, izraz pragmatičen pa je lahko zelo široko in različno razumljen.*

*[Here, for the markers of pragmatic structure, I use the expression interaction markers, because my previous research (Verdonik et al., 2007; Verdonik et al., in press) shows, that their main property is to establish and develop the relationship between co-speakers, while the expression pragmatic can be very widely and non-uniformly understood.]*

Definitions discussed can also be embedded in a non-definitional sentence where the definition is introduced in a relative clause, as illustrated in the sentence below.

- xxxiv. *Najbolj ohlapna so asociativna razmerja, ki povezujejo besede iz istega pomenskega polja (npr. zdravnik ↔ bolnišnica) in jih psihologi ponavadi pridobivajo s pomočjo asociativnih testov (Kilgarriff in Yallop 2000).*

*[The loosest are the associative relations, that connect the words from the same concept field (e.g., doctor ↔ hospital) and that psychologists usually acquire using associative tests (Kilgarriff in Yallop 2000).]*

In this analysis we show that a simple “X is\_a Y” pattern corresponding to the *genus-differentia* definition type is far from being the only way of defining concepts; it offers a too restrictive view on definitions as occurring in running text and that even this pattern raises several questions for discussion, for instance when the hypernym Y is too general or too specific. Different categories introduced in this linguistic analysis are

used in the discussion of the results in the rest of this thesis. They also serve as a basis for defining hand-crafted patterns for automatic extraction of definitions presented in Sections 5.1.1 and 5.2.1 for Slovene and English, respectively. Further on in this thesis, Section 5.3.4 reconsiders the question of different definition types introduced here, while complementing this analysis with new examples and insights from a much larger set of analyzed Slovene and English automatically extracted definitions.

This leads us to a brief conclusion and overview of the entire chapter. In summary, in Section 3.1 we described the task of the thesis, i.e., modeling a domain from Slovene and English text corpora, mainly focusing on a definition extraction task. In Section 3.2 we presented the building of the Language Technologies Corpus, our domain of interest. After the presentation of the corpora, we described domain by initial models—separately for Slovene and English subcorpus—in form of *topic ontologies* (cf. Section 3.3). This domain modeling step provides an overall understanding of the topics covered by the corpus. In the last section, Section 3.4, we analyzed the same corpus from a different angle: for a better understanding of the definition extraction task from the Language Technologies Corpus, we analyzed a subset of definitions and classified them into different definition types depending on the strategies used for defining a term.



## 4 Methodology and background technologies

The main challenge tackled in this dissertation is to develop and implement a definition extraction methodology to be implemented as an online workflow. This chapter summarizes the main definition extraction methods (Section 4.1) and introduces the measures used for evaluating definition candidates (Section 4.2). The background technologies used in the definition extraction process are described in Section 4.3, together with the evaluation of the reimplemented lemmatizer and morphosyntactic tagger (ToTrTaLe) and term extractor (LUIZ) presented in Section 4.4. The evaluation of the background technologies is important, since their performance influences the definition extraction results.

### 4.1 Overview of the definition extraction methodology

The dissertation addresses domain modeling from specialized corpora, where the main challenge is to model the domain by definition extraction used as means of glossary construction by the user. This section presents a brief overview of the developed methodology. A top-level overview of the methodology is shown in Figure 10. Starting from a specialized corpus, the first step is text preprocessing (segmentation and tokenization, lemmatization as well as morphosyntactic annotation of the corpus). Next, the domain is modeled by means of terminology extraction (i.e., extraction of terms specific to a given domain), followed by definition extraction, which results in a set of definition candidates. The term and definition candidates are proposed to the user for the final manual glossary construction phase.

Note however, that the schematic representation below is simplified. It is a semi-automatic and not an automatic process and the user is in the loop all the time. She can intervene between different phases, e.g., after the corpus collection the user can inspect the corpus by the OntoGen topic ontology construction approach (not added to the scheme, since it is not part of the definition extraction methodology) and complete or filter the corpus based on her findings. After the preprocessing step the user can decide to manually correct some errors. After the term extraction, one could filter the list of terms to be input in the next phase, and it is always the user that defines the parameters for the definition extraction phase. The final step of glossary construction needs still quite some manual work but could be further automatized in the future. The methodology is available as a workflow (cf. Chapter 6), and most probably, for any real application, the user will run the workflow several times and actively participate in the process.



whereas for Slovene we use sloWNet (Fišer and Sagot, 2008), a Slovene counterpart of WordNet.

The three approaches, together with text preprocessing performed with a morphosyntactic tagging and lemmatization tool and term extraction, represent the main ingredients of the definition extraction methodology developed in the scope of this thesis. The main results are the list of definition candidates, as well as a list of terminological candidates. Since our approach is intended to be semi-automatic, the resulting lists of term and definition candidates can be the subject of further manual refinement when forming a domain glossary (terminological dictionary).

The definition extraction methodology was applied to the Slovene and English parts of the Language Technologies Corpus, consisting of academic papers in the area of language technologies. The corpus was presented in detail in Section 3.2.

The methodology extends our initial work (Fišer et al., 2010) and the recent work presented in Pollak et al. (2012a). In Fišer et al. (2010) only Slovene definition extraction was addressed and the corpus was less specialized, consisting mainly of popular science textbooks and articles. Highly specialized language, which characterizes the Language Technologies corpus modeled in this thesis, is known to be more complex on the semantic, syntactic and lexical level compared to the popular science discourse (Schmied, 2007). As a pattern-based method in Fišer et al. (2010) we used a single *is\_a* pattern, which yields useful candidates if applied to structured texts such as textbooks or encyclopaediae. However, if used on less structured specialized texts, such as scientific papers addressed in this thesis, a larger range of patterns yield better results. In Pollak et al. (2012a), we used eleven different patterns for Slovene and four different patterns for English. In the final methodology presented in this thesis, we extended the list to twelve patterns for Slovene and seven for English. In this thesis we examine different parameter settings for the term-based approach and elaborate the methodology in much greater detail, including its extensive evaluation.

## 4.2 Definition extraction evaluation methodology

This section presents the definition extraction evaluation methodology, which will be used for quantitative and qualitative evaluation of results in Chapter 5.

For quantitative evaluation of definition extraction performance we use the measures of precision and recall, defined below.

- *Precision* denotes the percentage of actual definitions in the set of all candidate sentences that were automatically extracted by the method.
- *Recall* denotes the percentage of successfully extracted definitions from all the definition sentences. In our case, the recall is calculated on a recall test set, consisting of 150 manually selected actual definitions from the corpus.

An important part of the thesis is also the qualitative evaluation of the extracted definition candidates, discussed throughout Chapter 5. Besides the binary annotation of definition candidates into definitions (Y=yes) and non-definitions (N=no), needed for measuring the precision and the recall, we assign specific tags to extracted candidates, such as Y? and N? for borderline cases and more specific notions, such as Ns? (too specific – borderline non-definition), Ys? (too specific – borderline definition), Yg (too general definition), Ye? (borderline extensional definition) etc. These evaluation tags, inspired by the initial analysis of various definition types in Section 3.4, were

systematically used in Section 5.3.4 in order help grouping the qualitative interpretation of results.

The decision whether a definition candidate is a definition or not is a complicated question in itself. We opted for the interpretation that a definition does not have to correspond to a predefined formal criterion, such as e.g., that a definition should have the hypernym and the *differentia*. Consequently, we left the formal criterion open and instead set the condition that the definition should be informative enough to be included into a glossary with no or minimal manual refinement. Moreover, being aware of possible differences in annotating candidate definition sentences, we performed a set of inter-annotator agreement experiments, which indicate the difficulty of the task as well as the reliability of results. The *inter-annotator agreement* measures how different annotators agree in their judgments about selected definition candidates. These results are discussed in Section 5.3.3 of this thesis.

## 4.3 Background technologies and resources

In this section we describe existing tools—the linguistic annotation tool ToTrTaLe and the LUIZ term-extraction tool—and resources (WordNet and sloWNet) that were used as background technologies in the methodology developed in the scope of this thesis.

### 4.3.1 ToTrTaLe morphosyntactic tagger and lemmatiser

The ToTaLe (Erjavec et al., 2005) tool, whose name denotes a script for the Tokenization, Tagging and Lemmatization pipeline comprising these three text processing steps, is available as a web application. ToTaLe has recently been extended with another module, Transcription, and the new edition is called ToTrTaLe (Erjavec, 2011). The transcription step is used for modernizing historical language (or, in fact, any non-standard language), and the tool was used as the first step in the annotation of a reference corpus of historical Slovene (Erjavec, 2012a). An additional extension of ToTrTaLe is the ability to process heavily annotated XML document conformant to the Text Encoding Initiative Guidelines (TEI P5, 2007). The three main modules of ToTrTaLe, tokenization, tagging and lemmatization, are presented below. As a result of the work performed in this thesis, linguistic annotation with ToTrTaLe is now made available as a web service and as a publicly available workflow (cf. Section 6 and Pollak et al., 2012c).

#### Tokenization

The multilingual tokenization module mlToken<sup>27</sup> is written in Perl and in addition to splitting the input string into tokens also assigns to each token its type, e.g., XML tag, sentence final punctuation, digit, abbreviation, URL, etc. and preserves (subject to a flag) white-space, so that the input can be reconstituted from the output. Furthermore, the tokenizer also segments the input text into sentences.

The tokenizer can be fine-tuned by putting punctuation into various classes (e.g., word-breaking vs. non-breaking) and also uses several language-dependent resource files: a list of abbreviations (words ending in a period, which is a part of the token and

---

<sup>27</sup> mlToken was written in 2005 by Camelia Ignat, then working at the EU Joint Research Centre in Ispra, Italy.

does not necessarily end a sentence); a list of multi-word units (tokens consisting of several space-separated words); and a list of (right or left) clitics, i.e., cases where one word should be treated as several tokens. The tokenization resources for Slovene and English were developed by hand for both languages.

### **Tagging**

Part-of-speech tagging is the process of assigning a word-level grammatical tag to each word in running text, where the tagging is typically performed in two steps: the lexicon gives the possible tags for each word, while the disambiguation module assigns the correct tag based on the context of the word. Most contemporary taggers are trained on manually annotated corpora, and the TnT tagger we use is no exception. TnT (Brants, 2000) is a fast and robust tri-gram tagger, which is able, by the use of heuristics over the words in the training set, to tag unknown words.

For languages with rich inflection, such as Slovene, it is better to speak of morphosyntactic descriptions (MSDs) rather than part-of-speech tags, as MSDs contain much more information than just the part-of-speech. For example, the tagsets for English have typically 20–50 different tags, while Slovene has over 1,000 MSDs. For Slovene, the tagger has been trained on jos1M, the 1 million word JOS corpus of contemporary Slovene (Erjavec et al., 2010), and is also given a large background lexicon extracted from the 600 million word FidaPLUS reference corpus of contemporary Slovene (Stabej et al., 2006; Arhar Holdt and Gorjanc, 2007). The English model was trained on the MULTEXT-East corpus (Erjavec, 2012b), i.e., on George Orwell’s novel “1984”. This is of course a very small corpus, so the resulting model is not very good. However, it does have the advantage of using the MULTEXT-East tagset, which is compatible with the JOS one.

### **Lemmatization**

For lemmatization the system uses CLOG (Erjavec and Džeroski, 2004), which implements a machine learning approach to automatic lemmatization of (unknown) words. CLOG learns on the basis of input examples (pairs word-form/lemma, where each morphosyntactic tag is learnt separately) a first-order decision list, essentially a sequence of if-then-else clauses, where the defined operation is string concatenation. The learnt structures are Prolog programs but in order to minimize interface issues a converter from the Prolog program into one in Perl was developed.

The Slovene lemmatizer was trained on a lexicon extracted from the jos1M corpus. The lemmatization of language is reasonably accurate. However the learnt model, given that there are 2,000 separate classes, is quite large: the Perl rules have about 2 MB, which makes loading the lemmatizer slow. The English model was trained on the English MULTEXT-East corpus, which has about 15,000 lemmas and produces a reasonably good model, especially as English is fairly simple to lemmatize.

We implemented ToTrTaLe as a web service and used it as a workflow component, since the annotated corpus was needed in both main steps of the methodology, i.e., in the term extraction and definition extraction step. The output of ToTrTaLe is illustrated in Table 4.

```

5451 <w lemma="on" ctag="Pp3fsa--y">jo</w>
5452 <w lemma="na" ctag="Sa">na</w>
5453 <w lemma="primer" ctag="Ncmsan">primer</w>
5454 <w lemma="miseln" ctag="Agmpn">miseln</w>
5455 <w lemma="vzorec" ctag="Ncmpn">vzorci</w>
5456 <pc ctag=",">,</pc>
5457 <w lemma="tehnika" ctag="Ncfpn">tehnike</w>
5458 <w lemma="vihar" ctag="Npmsn">vihar</w>
5459 <pc ctag=".">.</pc>
5460 <w lemma="jenjati" ctag="Vmer3s">jenja</w>
5461 <w lemma="možgani" ctag="Ncmpg">možganov</w>
5462 <pc ctag=",">,</pc>
5463 <w type="abbrev" lemma="lpd." ctag="Y">lpd.</w>
5464 <w nform="v" lemma="v" ctag="Sa">v</w>
5465 <w lemma="odločitven" ctag="Agpmsay">odločitvent</w>
5466 <w lemma="analiza" ctag="Ncfsl">analizi</w>
5467 <w lemma="skušati" ctag="Vmrip">skušamo</w>
5468 <w lemma="problem" ctag="Ncmpa">probleme</w>
5469 <w lemma="strukturirati" ctag="Vmbn">strukturirati</w>
5470 <w lemma="ln" ctag="Cc">ln</w>
5471 <w lemma="on" ctag="Pp3mpa--y">jih</w>
5472 <w lemma="razdeliti" ctag="Vmen">razdeliti</w>
5473 <w lemma="na" ctag="Sa">na</w>
5474 <w lemma="majhen" ctag="Agcmpa">manjše</w>
5475 <w lemma="ter" ctag="Cc">ter</w>
5476 <w lemma="bolj" ctag="Rgp">bolj</w>
5477 <w lemma="obvladljiv" ctag="Agpfpn">obvladljive</w>
5478 <w lemma="podproblem" ctag="Ncmpa">podprobleme</w>
5479 <pc ctag=".">.</pc>
5480 </s>
5481 <s>
5482 <w nform="pri" lemma="pri" ctag="Sl">Pri</w>

```

Table 4. A sample output of ToTrTaLe, annotating sentences and tokens, with lemmas and MSD tags on words.

### 4.3.2 LUIZ terminology extractor

LUIZ (Vintar, 2010) is a terminology extraction tool for English and Slovene. Terminology extraction is performed in two steps. First, the terminological candidates are extracted based on morphosyntactic patterns. Next, the terminological candidates are weighted and ranked by their ‘termhood’ value.

To get the list of candidates potentially relevant terminological phrases corresponding to predefined patterns (e.g., Noun + Noun; Adjective + Noun, etc.) are retrieved from a morphosyntactically annotated corpus.

In order to attribute a termhood value to these candidates, first a list of all word types (i.e., lemmas) from the corpus is extracted and their frequencies are calculated. The words with the highest frequency are mainly function words. Next, the keyness for each lemma of the lexicon is calculated. Very general words from the domain corpus are supposed to have approximately the same distribution in a reference corpus and in a domain corpus, while domain specific words have a considerably higher (relative) frequency in the domain corpus. The relative frequency of each lemma is calculated as the ratio between the relative frequency of a word (lemma) in the domain corpus and the relative frequency of this word in the reference corpus.

As the reference corpus, LUIZ uses the FidaPLUS corpus (Stabej et al., 2006; Arhar Holdt and Gorjanc, 2007) for Slovene and the British National Corpus (BNC, 2001) for English. FidaPLUS consists of texts of different types from the majority of Slovene daily newspapers, various magazines and books from a number of publishers (fiction, non-fiction, textbooks), etc. It contains approximately 621 million words and it is freely available. The BNC reference corpus used for English is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English.

In the last step, the relative frequencies of lemmas are used in order to compute the termhood value of noun phrase terminological candidates. The termhood value  $W$  of a

candidate term  $a$  consisting of  $n$  words is computed as:

$$W(a) = \frac{f_a^2}{n} \times \sum_i^n (\log \frac{f_{n,D}}{N_D} - \log \frac{f_{n,R}}{N_R})$$

Where  $f_a$  is the absolute frequency of the candidate term in the domain-specific corpus,  $f_{n,D}$  and  $f_{n,R}$  are the frequencies of each constituent word in the domain specific and the reference corpus, respectively, and  $N_D$  and  $N_R$  are the sized of these two corpora in tokens (see Vintar, 2010).

LUIZ exports two lists of terms, one for single word terms and the other for multi-word terms. After the termhood score, the lemmatized form and the canonical form of the term are provided, the latter meaning that the term's headword is in singular for English and in the nominative case singular for Slovene.

As explained in the workflow implementation (see Section 6.4.2) we implemented the LUIZ term extractor as a web services and changed a few details, as well as provide the unique output list, on which single and multi word terms are ranked. The term extraction was used for the term-based definition extraction.

### 4.3.3 WordNet and sloWNet

WordNet (Fellbaum, 1998; PWN, 2010) is a lexical database that groups words (called *literals*) into sets of synonyms called *synsets*; each synset expresses a distinct concept. Same word forms with different meanings are represented in different synsets. Moreover, WordNet provides short, general definitions, and records the various semantic relations, mainly hypernymy, between the synonym sets. Hypernymy relations are transitive and all noun hierarchy chains reach the ultimate root node. Other relations are meronymy (part-whole relation), troponyms denote relations between verbs, while adjectives are organized in terms of antonymy.

For example, the English concept with ID (03082979) has six different realizations (literals):

*{computer, computing\_machine, computing\_device,  
data\_processor, electronic\_computer,  
information\_processing\_system}.*

The concept is defined by a short gloss: *a machine for performing calculations automatically* and related terms are provided, such as its direct hypernymy: [machine], defined as *any mechanical or electrical device that transmits or modifies energy to perform or assist in the performance of human tasks*.

WordNet has become one of the most important resources used in a large variety of natural language processing applications. Its utility initiated the construction of wordnets for many other languages including Slovene. Instead of manually building a large database, the Slovene wordnet, called sloWNet (Fišer, 2009; Fišer and Sagot, 2008) was built nearly fully automatically, by exploiting multiple multilingual resources, such as bilingual dictionaries, parallel corpora and online semantic resources. We used the version of sloWNet (2.1, 30/09/2009) containing about 20,000 unique literals, which are organized into almost 17,000 synsets, covering—as claimed in Vintar and Fišer, 2013—about 15% of PWN. The most frequent domain in sloWNet 2.1 is Factotum, followed by Zoology, Botany and Biology domains. SloWNet is aligned with the English PWN and since the sloWNet does not cover the entire inventory of PWN

concepts, there are some gaps (empty synsets) in the network.

In our methodology, WordNet and sloWNet were used in the wordnet-based definition extraction method.

#### 4.3.4 ClowdFlows workflow composition and execution environment

The ClowdFlows platform (Kranjc et al., 2012) consists of a workflow editor (the graphical user interface) and the server-side application, which handles the execution of the workflows and hosts a number of publicly available workflows. It is a cloud-based application that takes the processing load from the client's machine and moves it to remote servers where experiments can be run with or without user supervision.

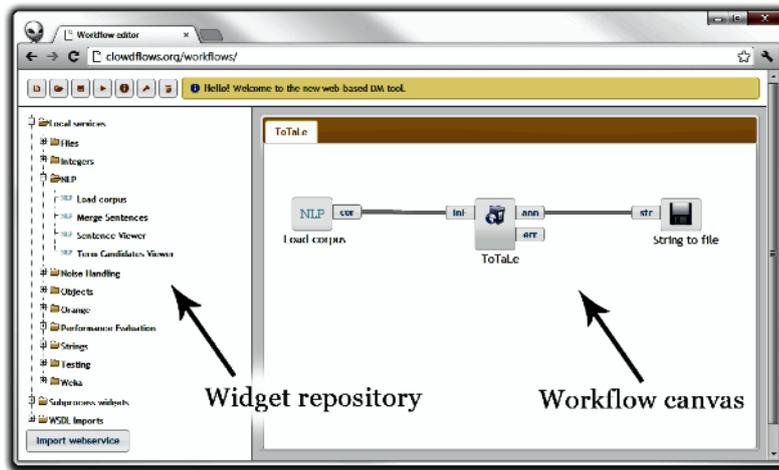


Figure 11. A screenshot of the ClowdFlows workflow editor in the Google Chrome browser.

As shown in Figure 11, the workflow editor consists of a workflow canvas and a widget repository, where widgets represent embedded chunks of software code, representing downloadable stand-alone applications which look and act like traditional applications but are implemented using web technologies and can therefore be easily embedded into third-party software. In this thesis, all NLP processing modules were implemented as such widgets, and their repository is shown in the menu at the left-hand side of the ClowdFlows canvas in the widget repository. The repository also includes a wide range of default widgets. The widgets are separated into categories for easier browsing and selection.

By using ClowdFlows we were able to make the workflow developed in this thesis public, so that anyone can execute it. The workflow is simply exposed by a unique web address which can be accessed from any modern web browser. Whenever the user opens a public workflow, a copy of the workflow appears in his private workflow repository in ClowdFlows. The user may execute the workflow and view its results or expand it by adding or removing widgets.

#### 4.4 Evaluation of selected background technologies

Note that the results of the definition extraction do not depend only on the quality of definition extraction methods themselves, but also on the methods used in the preceding steps of the definition extraction workflow. Therefore, in the next two subsections we discuss the preprocessing output and term-extraction results on our corpus. These two

background technologies were evaluated since we implemented them as separate workflow components. In further work, it would be useful to add the sloWNet and PWN evaluation, as well as perform a more complete ToTrTaLe evaluation.

#### 4.4.1 ToTrTaLe evaluation

ToTrTaLe—tokenization, lemmatization and morphosyntactic annotation tool—is the background technology used for all the further steps in the term and definition extraction process. In this subsection we present the observed ToTrTaLe mistakes, focusing on Slovene, and propose some corrections that were partly implemented in the post-processing step of the workflow (proposed as an optional parameter). The ToTrTaLe workflow implementation (and the proposed improvements), published by Pollak et al. (2012c), are presented in more detail in Section 6.2.

##### **Incorrect sentence segmentation**

We have detected errors in sentence segmentation, which originate mostly from the processing of abbreviations. Since the analyzed examples were taken from academic texts, specific abbreviations leading to incorrect separation of sentences are frequent.

In some examples the abbreviations contain the period that is—if the abbreviation is not listed in the abbreviation repository—automatically interpreted as the end of the sentence. For instance, the abbreviation *et al.*, frequently used in referring to other authors in academic writing therefore incorrectly implies the end of the sentence, and the year of the publication is mistakenly treated by ToTrTaLe as the start of a new sentence.

Note, however, that a period after the abbreviation does not always mean that the sentence actually continues. This is the case when an abbreviation occurs at the end of the sentence (*ipd.*, *itd.*, *etc.* are often in this position). Consequently, in some cases two sentences are mistakenly tagged by ToTrTaLe as a single sentence. This mistake was also observed with the abbreviations *EU* or measures *KB*, *MB*, *GB* if occurring at the last position of the sentence just before the period.

##### **Incorrect morphosyntactic annotations**

We have also identified morphosyntactic (MSD) tagging mistakes. Since several grammatical forms can have the same realization, there are examples where a wrong MSD is attributed to the token. In several cases these mistakes occur systematically. These mistakes can lead to lower performance in NLP tasks where MSD information is used (e.g., in the definition extraction task we can search for patterns “Noun-nominative is Noun-nominative”, and if the nominative case is not recognized this can result in lower recall).

For example, in Slovene, in the first masculine declension, the forms of nominative and accusative singular are the same for inanimate nouns. The same ambiguity—possibly leading to mistakes in MSD annotations—occurs for example in the second feminine declension singular and plural, as well as in the singular of the first feminine declension. Since also the gender/number can be wrongly assigned, other ambiguities can occur. In English, there are ambiguities for example between third person singular verb form and a noun in plural, both ending with “-s”. As an example, take the word *works* that was in the sequence *different reference works on Slovenian grammar* tagged as verb instead of noun (plural form). As mentioned in Section 4.3.1, the system was trained on a relatively small corpus for English, therefore wrong annotations are expected.

Another example is in subject complement structures. For instance, in the Slovene sentence *Kot podatkovne strukture so semantične mreže usmerjeni grafi.* [As data structures semantic networks are directed graphs.], the nominative plural feminine *semantične mreže* [semantic networks] is wrongly annotated as singular genitive feminine.

Another frequent type of mistake, easy to correct, is unrecognized gender/number/case agreement between adjective and noun in noun phrases. For example, in the sentence *Na eni strani imamo semantične leksikone...* [On the one hand we have semantic lexicons...], *semantične* [semantic] is assigned a feminine plural nominative MSD, while *leksikone* [lexicons] is attributed a masculine plural accusative tag.

Next, in several examples, *sta* (second person, dual form of verb *be*) is tagged as a noun. Even if—when written with capital letters—*STA* can be used as an abbreviation for *Slovenska tiskovna agencija* [Slovene Press Agency], it is much more frequent as the word-form of the auxiliary or copula verb.

### **Incorrect lemmatization**

Apart from common errors of wrong lemmatization of individual words (e.g., *hipernimija* being lemmatized as *hipernimi* [hypernoms] and not as *hipernimija* [hypernymy]), there are systematic errors when lemmatizing Slovene adjectives in comparative and superlative form, where the base form is not chosen as the lemma. Last but not least, there are typographic mistakes in the original text and due to end-of-line split words.

In summary, in this section we comment on several types of mistakes. It is obvious that these mistakes in corpus preprocessing influence the term and definition extraction results. For some of the systematically occurring mistakes we propose a post-processing script implemented in the definition extraction workflow. In further work other preprocessing tools (e.g., Tree Tagger (Schmid, 1994) or Obeliks (Grčar et al., 2012)) will be implemented and the influence of the preprocessing step will be tested.

## **4.4.2 LUIZ evaluation**

An important step of domain modeling is the terminology extraction step. We implemented the monolingual terminology extraction of the LUIZ system (Vintar, 2010), presented in more detail in Sections 4.3.2 and 6.3. The performance of the term extraction system also influences one of the three developed definition extraction methods, i.e., the term-based definition extraction that is evaluated in detail in Sections 5.1.2 and 5.2.2 for Slovene and English, respectively. The top ranked term from our corpus are provided in Table 5 (Slovene) and Table 6 (English).

Therefore, in this section, we evaluate the results of term extraction on the Language Technologies Corpus. The results of precision evaluation by two annotators and their agreement scores are presented in Table 7, while the recall results are presented in Table 8.

First, top 200 (single- or multi-word) domain terms for each language were evaluated by the domain expert (cf. A1 in Table 7). Each term was assigned a score of 1–5, where 1 means that the extracted candidate is not a term (e.g., *table*) and 5 that it is a fully lexicalized domain-specific term designating a specialized concept (e.g., *machine translation*). The scores between 2 and 4 are used to mark varying levels of domain-

specificity on the one hand (e.g., *evaluation* is a term, but not specific for this domain; score 3), and of phraseological stability on the other (e.g., *translation production* is a terminological collocation, not fully lexicalized, compositional in meaning, score 3).

1.000000	[korpus] <<korpus>>
1.000000	[diskurzen označevalec] <<diskurzni označevalec>>
0.756365	[govoren signal] <<govorni signal>>
0.746904	[strojen prevajanje] <<strojno prevajanje>>
0.561043	[slovenski jezik] <<slovenski jezik>>
0.414158	[jezikoven vir] <<jezikovni vir>>
0.367204	[jezik] <<jezik>>
0.343655	[besedilo] <<besedilo>>
0.319568	[beseda] <<beseda>>
0.311063	[spleten stran] <<spletna stran>>
0.294892	[beseden vrsta] <<besedna vrsta>>
0.289277	[naraven jezik] <<naravni jezik>>
0.284740	[govoren zbirka] <<govorna zbirka>>
0.279142	[pomnilnik prevod] <<pomnilnik prevodov>>
0.277600	[beseden zveza] <<besedna zveza>>
0.276101	[jezikoven tehnologija] <<jezikovna tehnologija>>
0.191862	[razpoznavanje govor] <<razpoznavanje govora>>
0.169757	[referenčen korpus] <<referenčni korpus>>
0.157759	[prevajanje govor] <<prevajanje govora>>
0.144995	[vzporeden korpus] <<vzporedni korpus>>
...	
...	

Table 5. Top ranked 20 terms from the Slovene Language Technologies Corpus.

1.000000	[language] <<language>>
1.000000	[machine translation] <<machine translation>>
0.979119	[word] <<word>>
0.833899	[corpus] <<corpus>>
0.650033	[translation] <<translation>>
0.536591	[translation memory] <<translation memories>>
0.291507	[mt system] <<mt system>>
0.260914	[system] <<system>>
0.251408	[target language] <<target language>>
0.244333	[language model] <<language models>>
0.226685	[text] <<text>>
0.218025	[speech recognition] <<speech recognition>>
0.172751	[sentence] <<sentence>>
0.171767	[rule] <<rule>>
0.163233	[natural language] <<natural language>>
0.153303	[data] <<>>
0.153056	[parallel corpus] <<parallel corpora>>
0.114266	[algorithm] <<algorithm>>
0.110055	[pos tag] <<pos tags>>
0.101847	[error] <<error>>
...	
...	

Table 6. Top ranked 20 terms from the English Language Technologies Corpus.

Precision <sup>28</sup>	Slovene terms			English terms		
Score	A1	A2	Average	A1	A2	Average
Yes (2-5)	0.905	0.710	0.860	0.815	0.980	0.940
Yes (5)	0.620	0.225	0.205	0.490	0.295	0.225
IAA-overall agr.	0.315			0.345		
IAA-kappa (unw.)	0.130			0.139		

Table 7. Precision of the reimplemented LUIZ term extraction method (on Slovene and English Language Technologies Corpus), evaluated by two annotators (the average denotes that the arithmetic mean of the two scores is above 2 in the first row and 5 in the second). The IAA scores for overall agreement and unweighted kappa are given in the last two rows and show low IAA agreement.

Next, we asked two other annotators—both experts in linguistics and familiar with the field of language technologies—to evaluate the same set of terms (their scores are given in columns A2 of Table 7). The mean of the two evaluators’ precision scores is given in column Average, where in the first row the precision is computed based on terms that have the average score of at least 2 and in next row provides the information about how many out of the top 200 terms were considered fully lexicalized domain terms by both annotators (assigned score 5).

Additionally, we performed an *inter-annotator agreement* experiment (using Lowry’s (2013) Vassarstats online kappa calculator). When calculating two evaluators’ agreement scores based on their evaluation on the scale from 1–5 the *overall agreement* is 31.5% for Slovene (where for 63 terms the two annotators gave the same score) and 34.5% for English (with 69 identically scored terms).

Next, we calculated different *kappa* coefficients (cf. Cohen, 1960) for measuring of agreement between the two annotators of each dataset. For Slovene the observed *unweighted kappa* is 0.13 and for English 0.1389. In contrast to overall agreement scores, kappa takes into account the agreement occurring by chance. Kappa lies on a scale, where 1 is perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate agreement less than chance, i.e., potential systematic disagreement between the observers (Viera and Garrett, 2005). On the scale from less than chance agreement and almost perfect agreement, our results show only slight agreement between the two annotators.<sup>29</sup> Since the annotation categories are in the ordinal scale, we computed also *kappa with linear* (0.3162 and 0.2591 for Slovene and English, respectively) and *quadratic weighting* (0.4536 and 0.3703 for Slovene and English, respectively). These scores indicate that the inter-annotator agreement is fair to moderate, if we consider taking into account not only absolute concordances, but also the distance between different categories (it is not the same if two annotators assigned scores 2 and 3 or 1 and 5).

To have a better idea of the evaluated terms we provide few terms for different scores. Term candidates assigned score 5 by both annotators are for example *memory-based machine translation*, *speech recognition*, *language resources*; scores from 2 to 4 were given by both annotators to terms *symbol error* (2), *rule* (3),

<sup>28</sup> Note that the results in Table 7 and Table 8 are slightly different from those reported in Pollak et al. (2012a) due to some minor revisions/improvements of the implementation.

<sup>29</sup> For interpreting kappa scores, note that score less than 0 indicates less than chance agreement, 0.01–0.20 slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement and 0.81–0.99 almost perfect agreement (Viera and Garrett, 2005).

*probability distribution* (4); score 1 was given to the expression *van den*, while 1.5 was the mean score for terms such as *problem*, *number* or *user*.

As already mentioned, one of the evaluators for English and Slovene terminology was the same and one was different in each case. We can note that the first evaluator who evaluated both datasets and whose results are reported in A1 columns tagged many term candidates with score 1 (not terms), but interpreted also many candidates as fully lexicalized domain-specific terms ('perfect' terms with score 5). The other two assigned score 5 less frequently, but did so also for score 1 (esp. the second evaluator of the English dataset who decided for score 1 only once). If we want to compare the performance of the two systems—the English and Slovene term extraction—we should rely on the results where the same person evaluated the two datasets (annotator A1). From these results one can see that the system performs slightly better for Slovene than for English.

The second part of term evaluation involved the assessment of recall. A domain expert (A1) annotated a random text sample of the Slovene and English corpus with all terminological expressions (fully lexicalized domain-specific terms (cf. score 5 in the previous part of the evaluation). Approximately 65 terms were identified for each language and these samples were then compared to the lists of terms extracted by the term extraction system. Table 8 shows the results for both samples using either all term candidates or just the top 10,000/5,000/200.

	Number of terms	Recall <sup>28</sup>
<b>Slovene</b>	33,978	0.714
	10,000	0.528
	5,000	0.443
	200	0.257
<b>English</b>	22,196	0.824
	10,000	0.719
	5,000	0.596
	200	0.246

Table 8. Recall of terminological candidates extracted from the LT Corpus.

The results of the term extraction step shown in this section have a big influence on the term-based definition extraction, presented in Sections 5.1.2 and 5.2.2. It was shown that the term extraction results are good but not perfect and that we can expect some errors due to the recognition of more general terms instead of fully lexicalized domain terms only. In future versions, we will consider performing term filtering and human evaluation before inputting the term list into the term-based definition extraction system. On the other hand, based on the recall evaluation, we decided not to limit ourselves to searching only for definitions of the extracted terms, since we might still miss approx. 20% of domain terms potentially defined in the text.

In summary, this chapter provided an overview of the definition extraction methodology, by first introducing the three definition extraction methods (Section 4.1), followed by the presentation of the methods for evaluating the extracted definition candidates in Section 4.2. The background technologies and resources were presented in Section 4.3 and those that reimplemented in the workflow were evaluated in Section 4.4.



## 5 Definition extraction from Slovene and English text corpora

This section describes the methodology and the results of definition extraction from Slovene and English text corpora. In Section 4.1 we have already presented an overview of the proposed definition extraction methodology. This chapter presents the core of this thesis, i.e., the developed definition extraction methodology for Slovene and English, where the main focus of our approach is on Slovene. The methodology was applied to the *Language Technologies Corpus*, which is—as briefly discussed in Section 3.4—characterized by very complicated constructions and presents difficult material for the given task. For each language, we developed and tested three different methods, i.e., the *pattern-based*, *term-based* and *wordnet-based approach*. Section 5.1 presents the methods and the results of definition extraction from the Slovene part of the corpus and Section 5.2 the methods and the definition extraction results obtained on the English subcorpus. The concluding section (Section 5.3) summarizes the results, proposes different original combinations of the three methods for each language and discusses the results in terms of text type and types of definition/non-definition sentences, the latter proposing a qualitative systematization of the results. Throughout the chapter, numerous examples of extracted definition candidates are presented, evaluated and analyzed, illustrating the challenge of the task and improving the understanding of the domain in terms of domain modeling. Moreover, this analysis represents a novel contribution from the linguistic perspective of improved understanding of definitions and defining strategies as occurring in running scientific text.

### 5.1 Extracting definitions from Slovene texts

This section presents the pattern-based, term-based and sloWNet-based method, applied to definition extraction from the Slovene part of the corpus. Since the evaluation of the term-extraction system showed that quite some terms (cca. 20%) are not extracted by the LUIZ system (cf. Table 8), we keep the setting more open and do not search only for definitions of a previously defined list of terms. In further work we could consider also experimenting with the alternative setting.

Developing the definition extraction methodology for domain modeling is the main focus of this thesis, however an additional aim is to semi-automatically construct a pilot Slovene glossary of the human language technologies domain. In our work, a *glossary* refers to a terminological resource, consisting of domain terms and their definitions. Compared to a terminological dictionary that should be as complete as possible (ideally defining ‘all’ the relevant concepts of a specific subject field, i.e., domain), a glossary conforming to our definition can also model a smaller domain (a corpus, a book, etc.), meaning that it can either model an entire subject field or not. Moreover, for a terminological dictionary one expects that variations, synonyms, etc. are properly handled, while a glossary can also be a less extensive resource. A glossary can also be viewed as a preliminary stage to building a proper terminological dictionary.

### 5.1.1 Pattern-based definition extraction

The pattern-based approach is the traditional approach, where we use predefined lexico-syntactic patterns. The simplest pattern is “X je Y” [“X is Y”], where X is the term to be defined and Y is its hypernym. This corresponds to the Aristotelian view of definition, with *genus* and *differentiae*, meaning that if we have term X to be defined, we define it by using its hypernym (Y) and by listing the differences from other types belonging to this class of entities (“X is Y that...”). In highly inflected languages, such as Slovene, we can add the condition that the noun phrases should agree in case and that the case should be nominative (i.e., “NP-nom is NP-nom”), where “NP” means *noun phrase* and “NP-nom” stands for *noun phrase in the nominative case*.

The simplest pattern is the above-mentioned “is\_a” pattern. Clearly, this basic “is\_a” pattern cannot cover all definition types. We can expect that in (semi-)structured texts, such as textbooks, encyclopaediae or Wikipedia, applying this pattern will yield good results. However, if used on less structured authentic specialized texts, such as scientific papers or theses, a larger range of patterns—capturing more definition candidates—should be considered. Therefore, inspired by the linguistic analysis of a small set of known definitions from the corpus, presented in Section 3.4, we crafted eleven additional patterns (or better said pattern types<sup>30</sup>) presented in Table 10. We used these patterns for automatic definition extraction and compared their performance in terms of precision and recall.

#### X is Y pattern type

First, we experimented with the basic “X je Y” [“X is Y”] pattern type (see Table 9). We evaluated different realizations of this pattern, as described below.

1. The most basic is “N\_je/sta/so\_N” pattern (cf. Pattern 1), where N denotes a “noun”. In English this patterns corresponds to “N\_is/are\_a/the\_N”, whereby Slovene does not use articles. The auxiliary verb *biti* [*be*<sup>31</sup>] can occur in third person singular (*je*) or in the forms for dual *sta* [*are*, dual form] (specific to Slovene) or plural *so* [*are*, plural form].
2. Next, we added the constraint that both nouns should agree in the nominative case (Pattern 2).
3. Since we are aware that a simple nominative does not cover all the types of noun phrases, we replaced the simple noun in the nominative case by a noun phrase (NP) in the nominative case (Pattern 3). Since no chunker was available for Slovene at the time of conception of these experiments, we manually defined different noun phrase types. All the noun phrases have a head noun that determines the noun phrase case (in our case we match noun phrases in the nominative case), while optional elements can precede the head noun (any number of adjectives) or follow it (nouns, adjectives and nouns, and/or prepositional phrases composed of a preposition introducing nouns, optionally preceded by adjective(s)). To illustrate it by examples, we extended the list from a simple noun *jezik* [*language*] in Pattern 2 to terms like

---

<sup>30</sup> We can say patterns or pattern types, since a pattern type can in fact have different realizations of the same type of pattern, in some cases due to the order of constituting elements and in other cases due to optional elements and various noun phrase structures.

<sup>31</sup> In the thesis, we had to choose between the bare infinitive and the full (or to-)infinitive form of verbs: for simplicity, the bare infinitive form is used in this thesis, e.g., we use *be* and not *to be*.

*računalniško jezikoslovje* [computational linguistics], *računalniška zbirka besedil* [electronic text collection] or *sistem za strojno prevajanje* [machine translation system], where the latter is in Slovene composed of a head noun, preposition, adjective and noun [system for machine translation].

4. In Pattern 4 we investigate whether a continuation of a sentence after the second NP—i.e., the *differentia* following the *genus* NP—improves the results; this is marked by “...” in Table 9 below.
5. Pattern 5 investigates how the condition of a noun phrase starting the sentence influences the precision and recall.
6. Pattern 6 is similar to Pattern 5 except that it obligatorily continues after the second NP.
7. In the last pattern (Pattern 7), we added two details. First, in a noun phrase we added the possibility that if several adjectives follow each other, the last element can be introduced by conjunction *in* [and] or *ali* [or] (*učna in testna množica* [training and test set]), in Table 9 we mark it with NP<sup>o</sup>-nom in order to differentiate it from the previous noun phrases without this option. Second, a noun phrase can be followed by an optional English noun phrase translation, introduced by the abbreviation *ang.* or *angl.* (e.g., *lematizacija (angl. lemmatization)*). The latter is very frequently used in our corpus, but also in Slovene Wikipedia or other texts when introducing terminology which is not yet established in Slovene.

X_is_Y type of patterns	# Extracted sentences	Precision (# of definitions in extract. sentences)	Recall estimate (# of def. in 150 recall test set)
1. N je/sta/so <sup>32</sup> N	1,711	0.1300 (est.) <sup>33</sup>	0.2600 (39)
2. N-nom je/sta/so N-nom	541	0.2052 (111)	0.2333 (35)
3. NP-nom je/sta/so NP-nom	1,255	0.1984 (249)	0.3933 (59)
4. NP-nom je/sta/so NP-nom...	1,199	0.2052 (246)	0.3867 (58)
5. ^NP-nom je/sta/so NP-nom	645	0.2356 (152)	0.2667 (40)
6. ^NP-nom je/sta/so NP-nom...	622	0.2411 (150)	0.2667 (40)
7. NP <sup>o</sup> -nom (ang./angl. NP)? je/sta/so NP-nom	1,281	0.2022 (259)	0.4000 (60)

Table 9. Evaluation of precision and recall for different variations of “X\_is\_Y” type of patterns on the Slovene corpus.

In Table 9, we provide the number of extracted sentences for each of the “X is Y” pattern type and provide the precision and recall for each pattern. For measuring precision, we evaluated all the extracted sentences.<sup>33</sup> Evaluating the entire sets of sentences all over the thesis has several reasons. Even if it represents a significant amount of time invested, this enabled us to actually identify the sentences that can be used as preliminary definitions for the language technologies glossary; from a simple

<sup>32</sup> We use *je/sta/so* [is/are] in the pattern list in order to facilitate the reading. The MSD tag corresponding to *je/sta/so* defines the category of auxiliary verb types in third person singular, dual or plural, present tense, without negative value (negative value differentiates *je* [is] from *ni* [isn't]).

<sup>33</sup> While for other patterns we evaluated all the candidate sentences, for Pattern 1 (which is the most basic pattern) we decided not to evaluate the entire number of candidates, since with other patterns, we expected to achieve higher precision and/or recall. Therefore, for the first pattern, the estimation is provided for 100 randomly selected sentences only.

proof-of-concept setting this approach enabled us to build an initial glossary of the domain, thus modeling the Slovene language technologies domain. Secondly, the more sentences we evaluate, the more complete is our overview of the variety of definition types in running, which is in line with the linguistic aims of this thesis.

As already mentioned in Section 4.2, recall was measured against the 150 definitions test set (the so-called *recall test set*), meaning that it is only an estimate of the actual recall. It is also interesting to observe the number of actual definitions extracted from the corpus, written in parentheses of the *Precision* column.

The following symbols are used in Table 9: “/” means alternative choices, “^” means beginning of a sentence, “?” means optional element, “...” means continuation of a sentence, “N” means noun, nom means nominative case, “NP” means noun phrase (details described in the text above) and “NP<sup>o</sup>” denotes noun phrases with optional *in/ali* [and/or] when enumerating the adjectives.

The precision of applying the first pattern is 0.13 and the estimated recall is 0.26. We can see that simply by adding a condition that a noun should be expressed in the nominative case, precision gets much higher (from 0.13 to 0.2052 as shown in the second row of Table 9). Pattern 3 improves especially the recall (0.2333 to 0.3933) and extracts 249 compared to 111 actual definitions from the corpus. Precision is slightly lower in this case (when a variety of noun phrases is taken into consideration instead of nominative noun only). Pattern 4 shows that definition extraction is more precise if the sentence continues after the *genus* part (in the majority of cases with a relative pronoun that introduces the *differentia* part, but still some of the definitions are lost when applying this condition). Patterns 5 and 6 show that precision can get as high as 24% if we look only at the sentences that start with a noun phrase. The last pattern (Pattern 7) has the extended definition of a noun phrase and has the best coverage of the recall test set and extracts more definitions than other methods.

To briefly discuss the results, we present a few examples. First, let us have a look at two examples of correctly extracted definitions:

- xxxv. *Lematizacija je postopek, pri katerem neki besedni obliki v besedilu tvorimo lemo (geslo, iztočnico).*  
[Lemmatization is a process, where we assign a lemma (keyword, source word) to a word form in a text.]
- xxxvi. *Ontologija je izraz, sposojen iz filozofije, ki na področju računalništva in informatike označuje formalno urejeno strukturo pojmov določenega področja in razmerij med njimi za namene inteligentnih aplikacij.*  
[Ontology is an expression, borrowed from philosophy, that in the domain of computer science and informatics defines a formally organized structure of concepts from a selected domain and relations between them, with the purpose of intelligent applications.]

We can see that in the first sentence the second noun is a (quite general) hypernym, while in the second sentence the second noun is a general term *izraz* [expression] and that the defining part comes only later introduced by *označuje* [denotes] (*formalno urejeno strukturo pojmov* [formally organized structure of concepts...]).

Compared to Pattern 2, Pattern 3 covers examples in which a noun phrase has a more complex structure than a simple noun. For example in sentence (xxxvii) the first noun phrase *programi s pomnilnikom prevodov* [programs with translation memory] is composed of a head noun in the nominative case plural, followed by a preposition, a noun in instrumental case and a noun in a genitive case. A similar example is sentence

(xxxviii) where the first nominative noun phrase has the head noun in the nominative followed by another noun in the genitive case.

- xxxvii. *Programi s pomnilnikom prevodov so integrirana orodja, ki združujejo najmanj dve komponenti, in sicer pomnilnik prevodov in terminološko banko, lahko pa vključujejo še druga od zgoraj omenjenih orodij.*

*[Translation memory programs are integrated tools that contain at least two components, namely a translation memory and a terminological bank, but can contain also other above-mentioned tools.]*

- xxxviii. *Pomnilnik prevodov je baza, ki vsebuje besedilne segmente v izvornem in ciljnem jeziku.*

*[Translation memory is a database, which contains word segments in the source and target languages.]*

Next, we analyze different types of false positive sentences (extracted non-definitions). In some examples, we observe a metaphoric use, where a sentence corresponds to the pattern, and is thus extracted, but cannot be used as a term definition because of its metaphorical meaning. Both sentences below have also the characteristics of defining a proper noun, which is, as already mentioned in Section 3.4.1, a complex case.

- xxxix. *Enciklopedija Britanika je kraljica med spletnimi enciklopedijami.*

*[Encyclopedia Britannica is the queen of web encyclopediae.]*

- xl. *Softpedia je zakladnica informacij za vsakega računalničarja.*

*[Softpedia is a treasury of information for every computer scientist.]*

In other examples the defined word is not a term. Looking at the sentence below, the term *WordList* denotes a function of the *WordSmith* tool, but we do not consider it a term. The tool where the defined function can be used (*WordSmith*) is not mentioned in the sentence, meaning that the definition is useless without background knowledge.

- xli. *Drugo osnovno orodje je WordList, ki sestavi seznam vseh pojavnih v korpusu in jih uredi bodisi po številu pojavitev bodisi po abecednem redu.*

*[Another basic tool is WordList, which composes a list of all corpus tokens and ranks them either by the number of occurrences or in alphabetic order.]*

As the last example, we provide a sentence which could be a definition if the definiendum were specified. Regardless of the missing definiendum, the reader can guess that the sentence describes the holonymy/meronymy relation.

- xlii. *Za tovrstne relacije nimamo slovenskega poimenovanja; gre za razmerja vključevanosti, in sicer X je del Y; v besediloslovju so bili uvedeni zaradi spoznanja, da razmerja NAD - in podpomenskosti zajamejo nekaterih v besedilnem svetu pomensko nepredvidljivo povezanih enot (Gorjanc 1999: 146).*

*[For these kinds of relations we do not have a Slovene naming; they concern inclusion relations, such as X is part of Y; in lexicography, these expressions were introduced because of the fact that the hyper- and hyponymy relations sometimes comprise conceptually unpredictably connected units (Gorjanc 1999: 146).]*

One of the reasons for not extracting definitions is the incorrect assignment of morphosyntactic descriptions by ToTrTaLe in text preprocessing. For example in

sentence (xliii) the first noun phrase *Sketch Engine* is tagged as noun in nominative for *Sketch* and adjective in accusative for *Engine* and is therefore not extracted by any of the patterns. Similarly, in sentence (xliv) below, the noun phrase *empirični pristop* is incorrectly tagged as accusative, and is therefore not extracted.

- xliii. *Sketch Engine je korpusno orodje, ki na vhodu sprejme korpus kateregakoli jezika ter njegove slovnčne vzorce, iz njih pa ustvari besedne skice (Word sketches) za besede tega jezika.*

*[Sketch Engine is a corpus tool that takes as its input a corpus in any language and its grammatical patterns, and creates word sketches (Word sketches) for the words of this language.]*

- xliv. *Konverzacijska analiza je empirični pristop, ki uporablja v glavnem indukcijske metode in išče ponavljajoče se vzorce v najrazličnejših posnetkih človeških pogovorov.*

*[Conversation analysis is an empirical approach that uses mainly induction methods and searches for repeating patterns in various human conversation recordings.]*

In summary, we have analyzed the extraction results of different types of “X is/are Y” patterns, together with their precision and estimated recall. The best pattern in terms of precision was Pattern 6, while the best pattern concerning the estimated recall is Pattern 7, which extracted the largest number (259) of actual definitions from the corpus. For this reason, Pattern 7 was selected as the basic pattern to be included in the final set of patterns presented in Table 10.

### Other pattern types

In addition to the selected “NP-nom je/sta/so NP-nom” pattern of the “X is Y” pattern type described above, we defined eleven other pattern types inspired by the analysis of the sentences presented in Section 3.4.

Table 10 lists the twelve pattern types. For each pattern we provide its English translation together with the number of extracted sentences, the precision and the recall (estimated on the recall test set). The last row (TOTAL) presents the results of applying all the patterns (i.e., the union of all the extracted definition sentences), which shows that the pattern-based approach has the overall precision of 0.2251.

A detailed description of the symbols used in the pattern list of Table 10 is as follows. “NP” denotes an extended notion of a noun phrase, covering different types of noun phrases, also the ones involving coordination of premodifiers (i.e., adjectives connected by *in* [*and*] or *ali* [*or*]); it also comprises the optional second noun phrase, introduced by the abbreviation *ang.* or *angl.* “NP-nom” denotes the noun phrase in the nominative case, “/” denotes alternatives, “.\*” means any number of words. All the words in the patterns denote word lemmas, except when the word is used between single quotes (‘), which in turn denotes a word form. “ADV” denotes adverbs, “PRT” a particle and “V-inf” a verb in infinitive. When there is an interrogation mark (?) it means that the element can occur one or zero times. When the pattern is split into a) and b), it belongs to the same pattern type, where just the word order is different.

As we can see from the results in of Table 10 the majority of examples are covered by the first pattern, which is an “X is/are Y” pattern type (i.e., “NP-nom je/sta/so NP-nom”). However, as shown in this table, better results can be achieved by using all the

twelve pattern types proposed. All the pattern types of Table 10 and some corresponding examples are described in more detail below.

1. The first pattern (NP-nom je/sta/so NP-nom) is a noun phrase in the nominative case, followed by third person singular, dual or plural of the verb *be*, followed by another noun phrase in nominative. Noun phrases (NP) have a variety of forms, such as “adjective+noun”, “noun+noun”, etc. An optional English noun phrase translation, introduced by abbreviation “ang.” or “angl.” can follow a NP. In all the following patterns we mean the same options when we use the term NP (*noun phrase*). (Note that this pattern is the best performing pattern in terms of recall among all the “X is Y” pattern types, evaluated in Table 9). An example sentence covered by this pattern is given below.

xliv. *Lematizacija je postopek pripisovanja osnovne oblike besedam v korpusnem besedilu.*<sup>34</sup>

[*Lemmatization is the process of assigning a base form to words in a corpus text.*]

2. The second pattern is similar to the first one but has an extra demonstrative pronoun *tisti* [*that/those*] before the second noun phrase. We selected the forms of the pronoun that cover examples in the nominative case (it can be in singular, dual or plural). It extracts sentences corresponding to the “NP is/are those NP ... that” pattern where *that* is a relative pronoun. An example is:

xlvi. *Morfologija ali oblikoslovje je tisti del slovnice (Top84), ki se ukvarja z notranjo zgradbo besed in njihovo funkcionalno vlogo v stavku.*

[*Morphology<sup>35</sup> is that part of the grammar (Top84) that is concerned with the internal structure of words and their functional role in a sentence.*]

3. The third pattern seeks for noun phrases in the nominative case that are defined by any realization of the lemmas of defining verbs other than the verb *be*: *definirati* [*define*], *opredeliti* [*determine*] *opisati* [*describe*], followed by a particle *kot* [*as*] and followed by another noun phrase. Example:

xlvii. *Dickinson (2009) v navezavi na Biberja (1993) reprezentativnost korpusa definira kot "mero, do katere vzorec vsebuje variabilnost celotne populacije" in ki nam omogoča, da pridobljene rezultate posplošujemo na celotno vzorčeno jezikovno zvrst.*

[*Dickinson (2009) referring to Biber (1993) determines the representativeness of a corpus as “the extent to which a sample includes the full range of variability in a population” and enables the generalization of the results to the entire linguistic genre that was sampled in the corpus.*]

---

<sup>34</sup> The sentence starts with a section number and title. To ensure easier reading we do not cover it in these examples, but we discuss the wrong segmentation issue in other sections.

<sup>35</sup> In Slovene the structure is *oblikoslovje ali morfologija* [*morphology or morphology*] where the two nouns are synonyms, the first one being the Slovene term and the latter the international form, and the two synonyms are connected by *or*:

Pattern	English translation of the pattern	# Extracted sentences	Precision (# of def. in extr. sentences)	Recall estimate (# of def. in 150 recall test set)
1. NP-nom je/sta/so <sup>32</sup> NP-nom <sup>36</sup>	NP-nom is/are NP-nom	1,281	0.2022 (259)	0.4000 (60)
2. NP-nom je/sta/so 'tisti/a/o/e' NP-nom .* ki	NP-nom is/are 'those' NP-nom .* that	10	0.5000 (5)	0.0133 (2)
3. NP .* definirati/opredeliti/opisati kot NP	NP .* define/describe/determine as NP	33	0.4848 (16)	0.0200 (3)
4. definirati/opredeliti/opisati NP kot NP	define/describe/determine NP as NP	8	0.6250 (5)	0.0067 (1)
5. a) pojem/beseda/termin/poimenovanje/izraz NP-nom .* nanašati/pomeniti b) nanašati/pomeniti .* pojem/beseda/termin/ poimenovanje/izraz NP-nom	a) concept/word/term/naming/expression NP-nom .* refer/mean b) refer/mean .* concept/word/term/ naming/expression NP-nom	40	0.2500 (10)	0.0067 (1)
6. 'V/Po' .* je/sta/so NP-nom definiran/opredeljen/predstavljen kot NP	'In/According to' .* is/are NP-nom defined/determined/presented as NP	5	0.8000 (4)	0.0133 (2)
7. NP .* imenovati/poimenovati (ADV/PRT)? NP	NP .* call/name (ADV/PRT)? NP	222	0.2793 (62)	0.0533 (8)
8. imenovan (ADV/PRT)? NP	called (ADV/PRT)? NP	87	0.2529 (22)	0.0400 (6)
9. NP .* znan pod ime NP	NP .* known under the name NP	2	1.0000 (2)	0.0067 (1)
10. 'Kot' NP .* je/sta/so .* NP .* NP	'As' NP .* is/are .* NP .* NP	67	0.2239 (15)	0.0600 (9)
11. naloga NP-gen je/sta/so NP/V-inf	role of NP-gen is/are NP/V-inf	9	0.3333 (3)	0.0067 (1)
12. a) kadar/ko/če .* 'govorimo' o NP b) o .* NP .* 'govorimo' .* kadar/ko/če	a) when/if .* 'we talk' about NP o b) about .* NP .* 'we talk' .* when/if	10	0.7000 (7)	0.0067 (1)
<b>TOTAL</b>		<b>1,728</b>	<b>0.2251 (389)</b>	<b>0.5867 (88)</b>

Table 10. Precision and recall of individual patterns on the Slovene data set. Second column contains the translated pattern in English, the third column indicates the number of extracted sentences with a number of true definitions between the parentheses. The fourth column gives the precision on all the extracted sentences and the last one the recall estimate, calculated on the 150 definitions dataset with the number of elements extracted from this dataset between the parentheses.

<sup>36</sup> This pattern corresponds to Pattern 7 of Table 9.

4. This pattern is similar to the previous pattern by using the same verb lemmas *definirati* [*define*], *opredeliti* [*determine*], *opisati* [*describe*], but has a different structure: “*definirati/opredeliti/opisati NP kot NP*” [“*define/determine/describe NP as NP*”]. Example:

xlvi. *Frank Austermühl [1] definira terminološki program kot orodje za izdelavo in vzdrževanje terminologije.*

[*Frank Austermühl [1] defines a terminology software as a tool for creating and maintaining the terminology.*]

5. This pattern has the condition that the sentence should contain one of the following words *pojem/beseda/termin/poimenovanje/izraz* [*concept/word/term/naming/expression*] followed by a noun phrase in the nominative case and one of the verbs *nanašati/pomeniti* [*refer/mean*]. The pattern has two different possible orders, either “*pojem/beseda/termin/poimenovanje/izraz NP-nom .\* nanašati/pomeniti*” [“*concept/word/term/naming/expression NP-nom .\* refer/mean*”] or “*nanašati/pomeniti .\* pojem/beseda/termin/poimenovanje/izraz NP-nom*” [“*refer/mean.\* concept/word/term/naming/expression NP-nom*”], where *.\** means possible other elements. This pattern models sentences of type “word X means/refers to Y ...”, such as:

xlix. *Termin govorni dogodek izhaja iz etnografije govora (njegov avtor je Hymes, povzeto po (Coulthard, 1985)) in pomeni največje jezikovne enote, za katere lahko ugotovimo jezikovno strukturo.*

[*The term speech event originates from ethnography of speech (its author is Hymes, summarized after (Coulthard, 1985)) and means the largest linguistic units for which we can determine the linguistic structure.*]

6. This pattern extracts sentences like Example (xii) of the 100 initially analyzed definition sentences (*In classical theory (Katz and Fodor 1963), the meanings of words are presented as sets of necessary and sufficient conditions that ...*). The first part introduces the author or the scope of the definition (*V [In]... / Po [According to]*) and the second part of sentence has the structure composed of the verb *biti* [*be*], noun phrase in the nominative case, participle form of the defining verb *definiran/opredejen/predstavljen* [*defined/determined/presented*] followed by *kot NP* [*as NP*].
7. The pattern matches sentences in which the noun phrase is defined by using the verbs *imenovati* or *poimenovati* [*name/call*]. Before the second noun phrase an optional adverb or particle (e.g., *tudi* [*also*], *kar* [*simply*]) can figure, see example below:

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

[*We call an inventory of linguistic denotations for concepts of a particular subject also a terminology, for example the geological, medical, alpine terminology.*]<sup>37</sup>

<sup>37</sup> Note that the original Slovene word order is different than the English translation: [the inventory of linguistic denotations for concepts of a particular subject] [we call also] [a terminology, /.../].

8. Similar to the previous pattern, the participle *imenovan* [*named/called*] can be used. An example is given below.

- li. *Avtomatsko pridobivanje leksikalnih podatkov iz korpusnih in primerljivih virov je v literaturi imenovano luščenje leksikalnih podatkov (ang. lexical data extraction).*

*[Automatic acquisition of lexical data from corpora and comparable resources is in the literature called lexical data extraction (Eng. lexical data extraction).]*

9. In this search pattern the definiendum is introduced by the expression *znan pod imenom* [*known under the name*] and defined by a noun phrase at the beginning of the sentence. Example:

- lii. *Pri skladijskem označevanju gre tako za funkcijsko- kot tudi za pomenskoskladijske oznake, ki seveda zahtevajo poglobljeno jezikoslovno analizo s pomočjo razvejane analize v drevesne strukture, znane pod imenom treebank.*

*[Syntactic tagging is concerned both with assigning functional- and semantic-structural tags that certainly need deep linguistic analysis with the aid of analysis into tree structures, known under the name treebank.]*

10. The next pattern matches sentences like Example (x) (“*As data structures, the semantic networks are pointed graphs, where the concepts are presented by points or nodes, and the relations by arrays or links between them*”).

11. Since in the analyzed set we observed that several sentences are also definitions defining the concept by its purpose, we introduce only one simple pattern “the role of NP-gen is/are” followed by a noun phrase or verb in the infinitive form. Example:

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

*[The role of part-of-speech taggers is to assign part-of-speech tags (English “part-of-speech”) or even more detailed forms of part-of-speech to words in a text.]*

12. In the last pattern we used the structure *kadar/ko/če* [*when/if*] and *govorimo o NP* [*we talk about NP*]. We did not use the lemma *govoriti* [*talk*], which is not specific to defining contexts but only the form *govorimo o*. The order can be inversed as can be seen in versions a) or b) of the pattern.

- liv. *Ko danes govorimo o korpusu, nam to pomeni računalniško zbirko besedil oz. delov besedil, zbranih po enotnih kriterijih za namene različnih, predvsem jezikoslovnih raziskav (Atkins et al. 1992: 1).*

*[Nowadays, when we talk about corpora, we mean electronic collections of texts or parts of texts, selected according to explicit design criteria for the purpose of various, especially linguistic research (Atkins et al. 1992: 1).]*

### Discussion on definition candidates extracted by the pattern-based approach

The examples of definition sentences outlined in Section 3.4 and the quantitative results of pattern-based definition extraction presented in Table 10 indicate the complexity of the actual task of definition extraction from the Slovene Language Technologies Corpus. Having in mind the goal of building a pilot glossary of the Slovene language technologies domain, the pattern-based approach—covering a variety of different pattern types—resulted in a reasonable number of candidate definition sentences to be inspected for inclusion in the glossary. All 1,728 candidate sentences were manually inspected and validated in terms of defining relevance, of which 389 were deemed acceptable as preliminary glossary definitions.

From Table 10 we can see that not all the patterns are productive to the same extent. For the moment they were evaluated on our corpus only, but in the future, when evaluated on a larger set of text types, we can decide to keep in the methodology only the most productive patterns, having a good precision-recall balance. However, compared to the simple straightforward pattern “NP-nom je/sta/so NP-nom” (Pattern 3 of Table 9), we extract 150 definitions more and improve both precision and recall.

Even if some effort was needed for defining different patterns, they can from now on be used for extracting definitions from any new corpus. Moreover, since the method is implemented in the modular workflow environment, the list of patterns can quickly be adjusted, while the pattern-based method can itself be combined with other methods.

Compared to creating a list of definitions from scratch, we think that the user can benefit from our approach, since it is quicker to filter out and edit sentences than to write definitions, the user can get the context of the sentences from the corpus since each sentence is associated with the source article ID, and less expert knowledge is needed (e.g., the system is used as a support in translation).

Next, we comment on some examples of extracted definition candidates, discuss the reasons for extracting false positive examples (i.e., sentences that correspond to one of the patterns but are not definitions) and discuss some other limitations of (semi-)automatic definition extraction.

Let us first provide an example definition extracted with the first (“NP-nom is/are NP-nom”) pattern (Example (lv)) and an example definition including a verb other than the verb *be* (Example (lvi)). Example (lv), extracted by the first pattern (“NP-nom je NP-nom” definition type), is basically the *genus* and *differentia* definition. The definiendum *reference corpus* is defined after the hinge *is* by the *genus* (*collection of texts*) that already contains the specific element that it is a *monolingual* collection of texts, and is followed by the (rest of) *differentiae*. The *differentiae*, i.e., what differentiates a *reference corpus* from other *monolingual collections of texts* is, as stated by the given definition, its representativeness of a certain language, as well as its specific use (that can serve for fundamental linguistic research). We can already see that the *differentia* structure is not as simple as when one provides typical (made up) examples of definition types. Instead of the structure ... *that is representative, balanced*, etc., the structure contains a relative clause *ki naj bi ... [that is supposed to ...]* which is already less clear because of the expressed modality. After the first part providing the typical characteristics of the definiendum, the second part of the *differentia* mentions the purpose/use. We can therefore speak of a combined type, in which the *genus-differentia* structure has the *differentia* of the *functional definition type*, as defined in Section 2.1.5.

- lv. *Referenčni korpus je enojezikovna zbirka besedil, ki naj bi predstavljala celovito podobo nekega jezika in tako služila kot izhodišče za temeljne jezikovne raziskave.*

*[A reference corpus is a monolingual collection of texts, which is supposed to present an integral representation of a given language and hence serve as a basis for fundamental linguistic research.]*

A definition that is extracted by a verb other than the verb *be* is given below. The sentence is extracted by Pattern 3 on the basis of the verb *definirati* [*define*]. The sentence is again a *genus-differentia* definition type with a functional definition element in the *differentia* structure. *Discourse markers* is the definiendum and *words* or *phrases* is a complex *genus*, while the *differentia* is given by the functional definition (the difference between discourse markers and other words is claimed to be in their function). In this sentence, which is taken out of the context, the verb *defines* is not attributed to a scientific authority; the subject is expressed by a verb in third person singular.

- lvi. *Diskurzne označevalce definira kot besede ali fraze, ki so uporabljene s primarno funkcijo usmeriti naslovnikovo pozornost k posebni vrsti povezave med izjavo, ki bo sledila, in trenutnim diskurznim kontekstom.*

*[He defines discourse markers as words or phrases that are used with a primary role of focusing the recipient's attention on a specific kind of relation between the utterance that will follow and the current discursive context.]*

As shown in Section 3.4 a definition in a corpus does not always correspond to the Aristotelian *genus and differentia* formula. Even sentences with no hypernym can be considered definitions. Example (lvii), extracted by Pattern 8, is a *functional definition*, defining the definiendum by its use/purpose.

- lvii. *S tako imenovanimi pregledovalniki lahko poiščemo zelene dele korpusa.*

*[With so-called corpus processing tools, we can find specific parts of the corpus.]*

Compare sentences (lviii) and (lix). The first one, extracted by Pattern 3, does not contain a hypernym, but is a definition sentence. The latter (extracted by Pattern 1), even if a hypernym is provided, is not an actual definition. It could be used as a second or third sentence in a dictionary entry but it cannot function as a definition itself. On the other hand it provides a synonym, so it can be considered a knowledge-rich context, but in order to be a definition at least one of the two synonyms would need to be defined.

- lviii. *Leibniz besedi definira kot sopomenki, če zamenjava ene z drugo nikoli ne spremeni pomena stavka, v katerem je do zamenjave prišlo.*

*[Leibnitz defines two words as synonyms, if substituting one with the other never changes the meaning of the sentence, in which the substitution was performed.]*

- lix. *Sopomenskost oziroma sinonimija je horizontalni pojav, relacija pa je simetrična: če je x sopomenka besede y, je tudi y sopomenka besede x /.../*

*[Synonymy<sup>38</sup> is a horizontal phenomenon and the relation is symmetrical: if x is a synonym of y, y is also a synonym of word x /.../]*

---

<sup>38</sup> In Slovene, the sentence starts with *Synonymy or synonymy*, where the first term is the Slovene term

We can observe also the *extensional definition type* (Example (Ix) below). Given that for defining *metadiscourse elements* all their categories are enumerated, the sentence can be interpreted as an enumerative extensional definition.

- lx. *Pri analizi sledi Hylandovi tipologiji, po kateri so metabesedilni elementi razvrščeni v deset kategorij (povzeto iz Pisanski, 2005): logični povezovalci (predvsem vezniki in prislovne besedne zveze), označevalci okvira (npr. najprej, nato, prvič, drugič, če zaključimo, moj namen je), endoforični označevalci (npr. glej spodaj, kot je bilo omenjeno zgoraj), dokazovalci (npr. citiranje), tolmači (npr. to se imenuje, z drugimi besedami), omejevalci in ojačevalci (npr. morda, možen, jasno), označevalci odnosa do vsebine (npr. žal, strinjam se), označevalci odnosa do bralca (npr. iskreno, bodite pozorni), označevalci osebe (npr. jaz, mi, moj, naš).*

*[The analysis follows Hyland's typology that classifies metadiscourse elements into ten categories (after Pisanski, 2005): logical connectors (especially conjunctions and adverbial phrases), frame markers (e.g., first, next, firstly, secondly, in conclusion, our aim is), endorhic markers (see below, as mentioned above), evidentials (e.g., citations), code glosses (e.g., called, in other words), hedges and boosters (maybe, possible, clearly), attitude markers (e.g., unfortunately, I agree), engagement markers (e.g., sincerely, be careful), person markers (e.g., me, we, my, our).]*

On the other hand, we should mention that several types of sentences, formally corresponding to definition patterns, are not definitions. We have already briefly discussed different reasons for extracting non-definition sentences. We here continue this line of analysis. This can be caused by the fact that the sentence (or its hypernym) is too general or too specific and therefore the sentence is not actually defining the term (these two deficiencies are discussed in Sections 2.1.5 and 2.1.6).

First we mention several examples of non-definitions, because of the 'too general' meaning. E.g., sentence (lxi) does not define the tool *Emacs* and sentence (lxii) is not a definition of *(English) lexicography*. Both correspond to Pattern 1, extracting *genus* and *differentia* definitions, but the hypernyms are too general (*Emacs–contemporary product; English lexicography–dynamic field*). Example (lxiii) for instance describes the *bag of words representation (BoW)* and provides the term's hypernym, but does not provide sufficient information to act as a definition. Of course, what is a definition is itself a complex question, discussed in more detail in Section 5.3.3 on inter-annotator agreement. In Example (lxiv), a sibling concept is used instead of a hypernym (cf. *analogic definition*). However, the definition is insufficient, because neither the differences between the two concepts nor their characteristics are explained. Moreover, we should be sure that the sibling concept is also defined in the same glossary. These examples also prove that the extracted sentences—even if not definitions—can contain useful information for further manual refinement of automatically extracted sentences.

- lxi. *Urejevalnik emacs je sodoben izdelek, ki je v marsikaterem tehnološkem pogledu pred komercialnimi izdelki najbolj znanih proizvajalcev.*  
*[Emacs editor is a contemporary product that is in many technological aspects better than commercial products of best-known producers.]*
- lxii. *Angleška leksikografija je dinamično področje, ki predvsem v zadnjih dvajsetih letih res sprotno spremlja spremembe na leksikalni ravni.*

---

and the second one is the international term.

*[English lexicography is a dynamic field that especially in last 20 years continually follows the lexical changes.]*

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

*[Bag of words (Engl. bag of words or BOW) is a simple technique of text transformation for the needs of classification.]*

- lxiv. *Lematizaciji podoben postopek se imenuje krnjenje (angl. stemming).*

*[A process similar to lemmatization is called stemming (Engl. stemming)].*

In some examples (see lxv) the definition candidate is too specific. In this example, *translation* is defined only in a specific setting (*statistical machine translation*) and it does not define the general concept of *translation*.

- lxv. *Besedilo je prevedeno glede na verjetnostno porazdelitev – prevod je tisto besedilo, ki ima najvišjo verjetnost, ta pa se običajno računa po posameznih povedih.*

*[Text is translated according to probability distribution – translation is the text with the highest probability, which is usually computed for each individual sentence.]*

An important number of false positive examples are out-of-domain sentences, meaning that even if the sentences are ‘true’ they cannot be used for domain modeling in terms of glossary construction. Some examples are come from very general expressions such as *problem*, *result* or *exception* which are very frequent in the structures *the results are...*, *the exception is...*, *the problem is...* (cf. sentence lxvi). Numerous out-of-domain examples can be identified as being extracted from corpus articles about Slovene wordnet construction or in papers providing examples for semantic relations. Another type of false positive examples are sentences that are not ‘true’, such as the example sentences (lxviii) and (lxix), taken from a Master’s thesis on automatic construction of logic exercises.

- lxvi. *Naslednji problem je velika dinamičnost človeškega govora – pri hitrem govoru je odstotek napak pri razpoznavi večji.*

*[The next problem is high dynamics of human speech – the faster the speech the higher the number of recognition errors].*

- lxvii. *Miza je kos pohištva.*

*[A table is a piece of furniture].*

- lxviii. *Vsa majhna telesa so kocke.*

*[All small bodies are cubes.]*

- lxix. *D je vitez.*

*[D is a knight.]*

For avoiding the extraction of out-of-domain sentences, we see several solutions. The first is to predefine a list of terms that we want to define, the second is to compute the domain termhood value for each sentence (we investigate this in Section 5.3 where we combine pattern-based and term-based definition extraction methods), and the third is to invest more time in corpus preprocessing, by filtering out the noisy parts of the text, (such as table contents, examples, etc.) and keeping only the main body text. We did perform a lot of preprocessing on the small LTC proceedings corpus, but not on the

main LT corpus. Note also that the last three examples below would not be extracted, if a more restrictive “is\_a” pattern, making the continuation of a sentence after the second noun phrase obligatory (cf. Patterns 4 and 6 in Table 9), was applied.

In summary, in Section 5.1.1 we defined and evaluated different patterns for pattern-based definition extraction from Slovene corpora. As it was shown in Table 10, the majority of examples are extracted by an extended version of the “X is\_a Y” pattern type. Other patterns that extract a non-negligible number of definitions are using verbs such as *definirati*, *opredeliti*, *opisati* [*define*, *describe*, *determine*], *imenovati* in *poimenovati* [*call*, *name*] and have higher precision than the first pattern. However, also the patterns extracting fewer definition candidates contribute to improved overall recall, but their utility should be tested on other types of corpora. The estimated recall of the union of different patterns is 58.67% and 389 definitions were extracted from the given Slovene corpus, as a result of manual evaluation of the entire set of 1,728 extracted definition candidates. Finally, a qualitative analysis was performed, discussing different definition types extracted by pattern-based methods from the corpus, in reference to the various definition types identified in Section 2.1.5 and Section 3.4. In addition, different reasons for extracting non-definitions or not extracting definitions were discussed (partially in line with Section 2.1.6 describing potential problems in definition construction).

### 5.1.2 Term-based definition extraction

The second approach, named term-based definition extraction, is primarily tailored to extract knowledge-rich contexts as it focuses on sentences that contain at least  $n$  domain-specific single or multi-word terms.

The first step of this approach is the extraction of domain terms. The term extraction module identifies potentially relevant terminological phrases on the basis of predefined morphosyntactic patterns (e.g., Noun+Noun; Adjective+Noun, etc.). These noun phrases are then filtered according to a weighting measure that compares normalized relative frequencies of single words between the domain-specific corpus and a general reference corpus of Slovene, i.e., the FidaPLUS corpus (Stabej et al., 2006; Arhar Holdt and Gorjanc, 2007). The term extraction tool LUIZ was briefly described in Section 4.3.2, while our implementation of the tool is presented in detail in Section 6.3.

Once the domain terminology has been extracted, we can use different parameters to extract definition candidates. The least selective condition is that the sentence contains at least two domain terms (a term pair). For this setting we expect lower precision since the condition is very loose, but it can lead to a better recall than the one achieved in the pattern-based approach. If only this basic condition is applied, the method can be used to measure the domain relevance of a sentence, however it is too imprecise to be used on its own and is therefore expected to be useful only in conjunction with other methods (i.e., the pattern-based or the wordnet-based approach).

In order to improve the precision of the term-based definition extraction, additional conditions can be specified (e.g., termhood value, verb between two terms, nominative case for terms, position of a term at the beginning of a sentence, etc.). Therefore we performed several sets of experiments by adding other conditions to the one stated above that a definition candidate contains at least two domain terms.

We tested several settings based on a number of hypotheses listed below. These are related to different parameter settings tested in the experiments, as will be described later in this section.

1. Precision will increase if—when matching the domain terms in a sentence—only higher scored terms are taken into account. This hypothesis was checked by experimenting with a *threshold* parameter (e.g., reducing the list of automatically extracted terms to 1% is better than considering top 10% of the extracted terms),
2. Taking into account more domain terms in a sentence will lead to higher precision (e.g., setting the *number of terms* to 3 instead of 2).
3. Imposing a *verb* condition, i.e., requesting that at least one verb occurs between two domain terms, will improve the performance. If more than two domain terms are considered we check also whether requesting that a verb occurs between the first two terms (VF) increases the precision compared to a setting where a verb may occur between any terms (VA).
4. Precision will increase if one term occurs at the *beginning of the sentence* (we consider the beginning of a sentence as the first or the second word in a sentence and do not limit it to first position only, because in English an article's position is before the term, while in Slovene, a preposition can be used).
5. Precision will increase if the first term is a multi-word expression (*multi-word first* parameter).
6. Increasing the *number of multi-word terms* will yield higher precision (e.g., if the number of terms is set to 3 and number of multi-word terms to 2, it means that at least two out of three domain terms detected in the sentence should be multi-word expressions).
7. Having terms in the *nominative* case will increase precision. (Note that this condition, applicable to Slovene only, is related to the pattern-based approach where “X-nom is Y-nom” is used, however allowing for any kind of verbs or other elements between the terms and not only the verb *be* or other predefined verb; the terminological noun phrases in nominative were identified based on the nominative case of the NP's head noun.)

For testing the above hypotheses we performed numerous experiments. Due to the complexity of testing numerous parameter settings and for the ease of reading this section, we decided to skip the tables of experimental results and their extensive discussion from this section, while describing the entire set of experiments in Appendix A (Table 23 and Table 24). This section describes only a selected set of experiments, reported in Table 11 below. In all these tables, the columns correspond to the parameters related to testing the seven hypotheses above: as the table columns correspond to the parameters and the rows represent different experiments (implementing different parameter settings).

Overall, the results presented in Table 23 and Table 24 mostly confirm the above hypotheses. A general trend is that the higher the termhood value<sup>39</sup> and the number of nominatives in the sentence, the better the precision and the lower the recall. Moreover, the more terms and multi-word terms in a sentence, the better the precision. In addition, other constraints, such as having a verb between two terms, having a term at the

---

<sup>39</sup> Terms extracted by the term extraction method are ranked by their termhood value, meaning that if e.g., 1% of terms are used, the termhood value is higher than if 2% of all extracted terms are used, etc.

beginning of a sentence and a multi-word term preceding a single-word term improve the results.

	<i>Threshold</i>	<i>Terms (#)</i>	<i>Verb (VA-VF)</i>	<i>Begin. sent.</i>	<i>Multi-word first</i>	<i>Multi-word (#)</i>	<i>Nominatives (#)</i>			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	<b>Extr. sent.</b>	<b>Precision</b>	<b>Recall-est.</b>
<b>A</b>	1%	2	no	no	no	no	no	28,215	0.0520 ♣	0.8533 (128)
<b>B</b>	10%	2	no	no	no	no	no	35,624	0.0300 ♣	0.9733 (146)
<b>ee</b>	1%	5	VF	yes	yes	3	2	102	0.2647 (27)	0.0067 (1)
<b>w</b>	1%	4	VA	yes	yes	2	1	499	0.1944 (97)	0.0333 (5)
<b>R</b>	1%	5	VA	yes	yes	no	no	594	0.1751 (104)	0.0467 (7)
<b>R &amp; w</b>								721	0.1747 (126)	0.0467 (7)

Table 11. Selection of settings of term-based methods from Table 23 and Table 24. (A: basic; B: highest recall; ee: highest precision; R: best precision-recall tradeoff<sup>40</sup> (settings without nominative); w: best precision-recall tradeoff (settings with nominative); R&w: union of w and R, set as a suggested combination. (For the less restrictive settings, we evaluated the precision on 1,000 randomly selected definition candidates (sign ♣), the others were evaluated in totality.)

The results presented in the tables show the precision and recall achieved in individual experiments. Precision evaluation was performed by extensive manual evaluation of several thousand definition candidates (in Table 11, Table 23 and Table 24) we evaluated all the extracted sentences in the majority of settings, and only in setting with sign ♣ we evaluated a subset of 1,000 randomly selected sentences. On the other hand, recall was evaluated on the 150 definitions test set, hence only providing the estimated recall. However, to get a feel of actual performance, the exact number of actual definitions found among the extracted definition candidates is also provided (these numbers are given in parentheses in the precision column). Extensive manual evaluation was not only done for showing the results presented in these tables but also for

<sup>40</sup> A measure that combines precision and recall is called the F-measure. The basic variant is  $F_1$ -score which is the harmonic mean of precision and recall. In our case we opted for  $F_{0.5}$  which attributes more weight to precision score. The formula is as follows:

$$F_{0.5} = (1 + 0.5^2) * \frac{\text{precision} * \text{recall}}{0.5^2 * \text{precision} + \text{recall}}$$

We have decided for this option because the precision results are in our settings generally not very high and we prefer settings with slightly higher precision. We have calculated the  $F_{0.5}$  taking the actual precision based on the evaluation of all definition candidates extracted by each setting (except in several cases with very loose constraints, where precision, marked with sign ♣, is an estimate). The recall is given as an estimate, where we have taken an average of two recall estimates. One recall measure was the one presented in Section 4.2 and used in all the tables, i.e., the estimate on the 150 definitions test set. The other estimate of the recall was motivated by the observation that even if the number of actual definitions extracted by a specific setting was higher than when using a different setting, the recall on the 150 definitions test set was sometimes the same since it depends on the particular test set. Therefore another estimate of recall was proposed, measured in the following way: we randomly selected 1,000 sentences from the corpus and evaluated them with definition vs. non-definition tags. The number of definitions, i.e., 24 was used to estimate the number of estimated definitions in the entire corpus. The estimated recall was then computed as the number of actually extracted definitions divided by the estimated number of definitions in the corpus. The recall based on which the  $F_{0.5}$  was computed is the average of these two recall estimates.

determining the candidate entries for the pilot Slovene language technologies glossary which is one of the results of this thesis.

Table 11 provides only a selection of the entire set of performed experiments, reported in Table 23 (experiments denoted by upper-case letters) and Table 24 (experiments denoted by lower-case letters). The most basic setting in all the experiments is Setting (A) with a simple condition that a sentence should contain 2 domain terms above a threshold set at 1%; the second simple setting (B) is the one with the same condition but considering top 10% of terms with the highest termhood value. We can see that with (B) nearly all the candidates from the 150 definition recall test set are selected, but both, (A) and (B), select too many sentences with too low precision to be used in practice. The highest precision is achieved in Setting (ee); in this setting the strictest parameters are applied (e.g., at least 5 domain terms out of which 3 must be multi-word expressions and 2 in nominative case), leading to the best precision in our experiments. The next two settings (R and w) were selected as the best compromise between precision and recall—with a slight preference for the precision— where (R) is selected from the experiments without the nominative case constraint and (w) having a condition of at least one term in the nominative case. The last row of Table 11 (R&w) gives the results for the union of sentences of these two settings (R and w)<sup>40</sup> and is well-suited to be used as a default setting, given that it achieves a suitable tradeoff between the precision and the number of extracted definitions (with a higher weight imposed on precision than on recall). Note that based on the objective of the user's application, one can choose to tune the method for higher precision or recall by selecting different parameter settings. Moreover, it also depends on the decision whether the approach will be used on its own or in combination with other approaches. The results for combining different approaches will be further considered in Section 5.3.1.

Note again that an extensive comparison of different parameter settings is provided in Table 23 and Table 24, used as a basis for a detailed quantitative evaluation of each of the hypotheses presented in this section (see Appendix A for details).

### **Discussing definition candidates extracted by the term-based approach**

Following a brief quantitative evaluation of different methods summarized above, we shall qualitatively examine the term-based approach. We first outline several advantages of the term-based approach. Next, we illustrate the influence of different parameter settings by examining several examples of extracted sentences that were extracted by selecting certain parameter values.

#### ***Advantages of the term-based approach***

An advantage of the term-based approach compared to the pattern-based approach is the extraction of sentences which do not have a typical defining verb of any of the predefined patterns. For example, with a term-based approach one can extract sentences with a verb used for describing the purpose of a term, i.e., functional definitions. Note that in the first example below, *translation memory* is defined by its function, i.e., it is a functional definition within which also the term *translation unit* is defined by means of a paraphrase. The next example is also a functional definition. (Note that in both examples the nominative constraint was used).

- lxx. *Pomilnik prevodov hrani prevodne enote, tj. segmente (ponavadi povedi) nekega originala in njihove prevode.*

[Translation memory saves translation units, i.e., segments (usually sentences) of an original, and their translations.]

- lxxi. *Besedna skica prikazuje leksikalni profil izbrane iztočnice s podatki o njenem tipičnem sobesedilnem okolju (Gantar in sod. 2009: 33).*

[A word sketch presents the lexical profile of a selected word together with its typical contexts (Gantar et al. 2009: 33)].

The next example illustrates that the term-based approach can find defining verbs that were not explicitly included in the pattern-based approach, e.g., *predstavljati* [present] in the sentence below. The term-based approach could therefore in the future be used for enlarging the set of verbs and typical defining structures of the pattern-based approach.

- lxxii. *Naglasno mesto predstavlja zlog, na katerem ima beseda tonsko ali jakostno izrazitost.*

[Accentuation position represents the syllable on which the word has a tonal or intensity stress.]

Example (lxxiii) shows another characteristic of the term-based approach; compared to the pattern-based approach, it is less sensitive to word order (which is more flexible in Slovene than in English) and to complex noun phrase construction issues. For instance, the pattern “NP je/sta/so NP” does not predict that the part after the copula verb *be* can start with a preposition (e.g., preposition *na* in *na računalništvo vezana raziskovalna smer*).

- lxxiii. *Obdelava naravnega jezika (ONJ) je na računalništvo vezana raziskovalna smer, ki jezik in jezikoslovne ugotovitve uporablja predvsem za (pol) avtomatsko pridobivanje raznovrstnih podatkov, potrebnih za razvoj računalniških aplikacij, ki so z jezikom povezane (jezikovnih tehnologij).*

[Natural language processing (NLP) is a research area related to computer science, that uses the language and linguistic findings mostly for (semi-)automatic extraction of various data, needed for the development of computer applications related to language (language technologies).]<sup>41</sup>

An interesting definition subtype mentioned in Section 3.4.2 are definitions in which a term is defined through a sibling concept (and *differentia*), also called analogical definitions. (One should note that for actual terminological dictionary (glossary) creation, a definition has a full meaning only if a sibling concept is defined in the same resource.)

- lxxiv. *Temeljne razlike med terminologijo in leksikologijo Kot se terminologija ukvarja s termini in terminotvorjem, se tudi leksikologija ukvarja z leksemi in postopki tvorjenja leksikalnih enot.*

[Fundamental differences between terminology and lexicology Similar to terminology that investigates terms and term formation, lexicography investigates lexemes and lexical units formation.]

- lxxv. *Leksikalne semantične mreže so zelo podobne semantičnim mrežam v umetni inteligenci, glavna razlika pa je v izhodišču, ki je pri leksikalnih mrežah leksikalna raven jezika, pri semantičnih mrežah pa pojmovna raven.*

---

<sup>41</sup> The English translation does not show the above-mention problem. A literal translation would be “Natural language processing (NLP) is to computer science related research area /.../”.

[Lexical semantic networks are very similar to semantic networks in artificial intelligence, while the main difference is in the starting point, which is for lexical networks the lexical language level, whereas for semantic networks it is the concept level.]

- lxxvi. *Statistično strojno prevajanje Statistična metoda strojnega prevajanja v nasprotju z metodami na osnovi pravil temelji na večji količini vzporednih besedil, iz katerih se s statističnimi algoritmi izračunavajo verjetnosti prevodne ekvivalence za posamezne jezikovne enote.*

[Statistical machine translation The statistical machine translation method is, in contrast to rule-based methods, based on larger amounts of parallel texts from which statistical algorithms calculate the probabilities of translation equivalents for individual language units.]

The term-based method also extracts more complex sentences, where an informal definition is embedded in a sentence and introduced by a relative pronoun.

- lxxvii. *Načela gradnje semantičnih leksikonov Predstavljen semantični leksikon je oblikovan v skladu z načeli teorije relacijskih modelov (Evens: 1988), kjer pomene besed opredeljujejo (paradigmatska) pomenska razmerja, ki veljajo med besedami in jih združujejo v pomenske mreže.*

[Principles of semantic lexicon construction The presented semantic lexicon is constructed in accordance with the principles of relational models theory (Evens: 1988) where word meanings are determined by (paradigmatic) meaning relations that hold between words, connecting them into semantic networks.]

The pattern-based approach is highly dependent on morphosyntactic descriptions assigned in text preprocessing. Since morphosyntactic corpus annotation does not perform without mistakes, an advantage of the term-based approach is to extract also sentences where the pattern-based approach fails because of these mistakes. E.g., sentence (lxxviii) should have been extracted by the “NP-nom is/are NP-nom”<sup>42</sup> pattern, but since the second noun phrase is wrongly annotated as genitive instead of nominative, it was not extracted. In contrast, it figures between definition candidates extracted by the term-based method with no nominative condition (or if the nominative condition was set at 1 instead of 2 nominatives). Note that the given example is a borderline case, since the term is defined by its hypernym, however the *differentia* is not specific enough.

- lxxviii. *Referenčni korpusi so enojezikovne zbirke besedil, ki pomenijo obsežen vir informacij o jeziku in njegovih lastnostih.*

[Reference corpora are monolingual text collections, representing a meaningful source of information about the language and its properties.]

A similar case is Example (lxxix). The sentence has a structure “NP-nom .\* označuje [denote] NP-acc”. Since this pattern type does not correspond to any of the manually crafted patterns, it illustrates the advantage of the term-based method extracting the sentences with other verbs and structures than the ones manually defined (based on the performed linguistic analysis). Moreover, the first noun phrase *izraz strojno prevajanje* is incorrectly tagged as accusative and therefore could not be extracted based on the

---

<sup>42</sup> NP-nom: noun phrase in the nominative case.

nominative patterns. However, these kinds of errors also influence the term-based extraction if the nominative parameter is applied.

- lxxx. *Izraz strojno prevajanje (MT – Machine Translation) navadno označuje računalniške sisteme za prevajanje naravnih jezikov, pri katerih je prevajalski proces do največje možne mere avtomatiziran.*<sup>43</sup>

*[Expression machine translation (MT – Machine Translation) usually denotes computer systems for translating natural languages, where the translation process is as much as possible automated.]*

As the examples above illustrate, the term-based approach can successfully complement the pattern-based approach.

### ***Effects of parameter settings***

In the perspective of using the term-based approach on its own, it is important to use additional conditions, complementing the main condition that a sentence should contain at least two terms, in order to achieve higher precision. However, each additional condition limits the recall, and below we discuss the effects and characteristics of different constraints/parameter settings. (Note that if the term-based approach is used in combination with other methods (intersection), the settings with better recall should be considered, as will be discussed in Section 5.3.1). This section thus illustrates the effects of different parameter settings, in line with the seven hypotheses postulated at the beginning of this section, while the quantitative results are provided and interpreted in Appendix A. (Note again that different parameters correspond to the columns of Table 11, Table 23 and Table 24.)

Ad Hypothesis 1. The term-extraction threshold parameter influences the precision and recall. For instance, when setting the parameter to the top ten percent of the automatically extracted domain terms, out-of-domain terms are often included. And even the termhood value set at 1% does not completely solve the problem, since the term extraction methods do not perform without mistakes. The most common examples are those starting with *tabela* [table] or *slika* [figure], which are highly ranked in the extracted term list. Also *difference* [razlika] or *description* [opis] included in top 10% are often used in articles and are not domain terms. E.g., Example (lxxx) includes two highly ranked words (*description* and *chapter*), which are actually not domain terms (they are specific to scientific writing but not to the language technologies domain).

- lxxx. *Kratek opis programa prinaša poglavje IV-3.1.4.*

*[A short description of the program is described in Chapter IV-3.1.4.]*

Ad Hypothesis 2. The higher the number of domain terms, the higher the precision. However, we should note that in a number of sentences the segmentation was faulty, causing that the wrongly separated title eventually increased the number of terms in the sentence. While the sentences with wrong segmentation are all borderline cases

---

<sup>43</sup> Note that the originally extracted sentence contains the title *STROJNO PREVAJANJE NEKOČ IN DANES [MACHINE TRANSLATION IN THE PAST AND TODAY]*, due to wrong sentence segmentation. Wrong segmentation is addressed in Section 5.3.4, while for simplicity we omit the wrongly segmented title parts of sentences from some of the examples in the text.

(because of wrong segmentation and therefore requesting manual refinement), a side effect of this error is the increased number of terms in the sentence following the title; therefore this feature may even be considered beneficial for definition extraction.

lxxxii. *[4.1.2 Kalkiranje Kalkiranje je terminotvorni postopek, kjer slovensko ustreznico tvorimo neposredno po tujejezični predlogi, ali drugače rečeno, kadar prevzeti izraz dobesečno, včasih celo po posameznih morfemih, prevedemo (npr. internet v med-mrežje, viktimologija v žrtv-o-slovje, escape character v ubežni znak).]*

*[4.1.2. Calquing Calquing is a term-formation process, in which a Slovene equivalent is formed directly from the foreign word, or said differently, when we translate the borrowed term literally, sometimes even based on morphemes (e.g., internet to med-mrežje, victimology to žrtv-o-slovje, escape character to ubežni znak).]*

Ad Hypothesis 3. The constraint of a verb occurring between two domain terms results in higher precision, but experiments show that in the majority of cases where exactly the same settings but where in addition the verb condition is applied, the recall remains the same (cf. Appendix A). The only examples where the recall is lower are the examples with the most basic settings—two domain terms—is applied and when we add the verb condition. For example, sentence (lxxxii) has the structure where a copula verb precedes the definiendum and the definiens. These types of sentences cannot be extracted with ‘verb constraint’ setting.

lxxxii. *Na splošno je akustična segmentacija členitev zvočnega niza na homogene odseke po nekem vnaprej določenem pravilu.*

*[In general, acoustic segmentation is the segmentation of an acoustic sequence into homogenous parts based on a predetermined rule.]<sup>44</sup>*

Ad Hypothesis 4. The condition of having a term at the beginning of a sentence generally increases the precision. However, certain types of definitions are excluded because of their introductory part that defines the source or the scope of the definition. See Example (lxxxiii) below.

lxxxiii. *Kot navaja Britanika,<sup>45</sup> je korpus definiran kot zbirka besedil določene teme, ki se uporablja za lingvistično analizo.*

*[As cited in Britannica, a corpus is defined as a collection of texts for a specific domain, used for linguistic analysis.]*

Ad Hypotheses 5 and 6. The condition that the first appearing domain term is a multi-word term and the parameter selecting a higher number of multi-word expressions influence the recall by eliminating sentences like Example (lxxxiv). In this sentence, when the threshold is set at 1%, the considered terms are *lema*, *osnovna oblika*, *beseda* and *oblika*, meaning that we have only one multi-word expression (which is not the first

---

<sup>44</sup> The English translation does not reflect the Slovene structure where the verb precedes both terms. Slovene structure has the copula verb *is* at the following place: *[In general, IS acoustic segmentation the segmentation of acoustic sequence into homogenous parts based on a predetermined rule.]*

<sup>45</sup> In the original corpus, a footnote number stands after *Britannica*.

term)—therefore if the parameter of multi-word terms in a sentence is set higher than 1, the sentence will not be extracted.

lxxxiv. *Lema je kanonična, osnovna oblika besede (npr. lema za besedo „hitrega“ je „hiter“, za „pogledali“ „pogledati“ itd.).*

*[A lemma is the canonical, elementary form of a word (e.g., lemma for the word “quick” is “quick”,<sup>46</sup> for “looked” is “look”, etc.).]*

Ad Hypothesis 7. Nominative condition is generally highly correlated with better precision. However, there are sentences that are definitions without nominative terms that we miss if the nominative condition is applied, or a sentence is not extracted if it has only one nominative, but the setting is set to two nominatives. A type of sentence that we can extract with the condition of a multi-word term at the beginning of a sentence (cf. Hypotheses 4 and 5), but is excluded if the number of nominatives is set to 2 is e.g., Example (lxxxv), where the first term is in the accusative case.

lxxxv. *Na diskurzne označevalce med prvimi na kratko opozori (Kranjc, 1999: 65), in sicer navaja, da diskurzni označevalci, npr. veš, ja, aha, »/o/pravljajo vlogo sredstva preverjanja pozornosti, hkrati pa so tudi sredstvo označevanja oziroma kazanja različnih vrst udeleževanja in pritrjevanja«.*

*[Discourse markers were initially noted by (Kranjc, 1999: 65), who states that discourse markers, such as e.g., you know, yes, aha, »/h/ave the function of checking attention and at the same time are the means of annotation or showing of different kinds of engagement or agreement«.]*

### ***Inspecting borderline cases (mostly knowledge-rich context sentences)***

Since the conditions of this approach, regardless of the selected parameter settings, are still very loose, there are many definition candidates that cannot be accepted as definitions. However, in many cases we can still talk of knowledge-rich contexts often resulting in categories of borderline definitions or non-definitions (as discussed in more detail in Section 5.3.4). These borderline knowledge-rich context sentences (KRC) can be attributed to two main reasons: either we have a domain term that the provided candidate definition sentence does not properly define, but still provides useful information about it (the definition candidate is too general or too specific, only a hypernym is provided, wrong segmentation etc.); or the definiendum is not (fully) considered as a domain term. It is worth considering different types of extracted borderline definitions or non-definitions, as well as the reasons for extracting non-definition sentences. Often the decision about the definition/non-definition label is not easy and many cases are complex and depend on the evaluator's choice (as will be further discussed in Section 5.3.3).

Several definition candidates were annotated as non-definitions because the hypernym is too general and the sentence does not define the term (an example is given below).

lxxxvi. *Naravni jezik je kompleksna in živa tvorba in ustrezna pravila za opisovanje so temu primerno zapletena, če jih je sploh mogoče vsa zapisati.*

---

<sup>46</sup> *hitrega* [quick] is in Slovene in the genitive case, therefore the English translation does not illustrate the difference between the base and inflected form.

*[Natural language is a complex and live construct and adequate rules for describing are therefore complicated, or even impossible to be defined.]*

The sentence can also be too specific. For instance, Example (lxxxvii) can be understood as a knowledge-rich context, since it provides information about lemmatization, but it is not a definition, since *lemmatization* cannot be limited to two specific transformations, such as deleting the information about the declination and the plural.

- lxxxvii. *Lematizacija pa je proces, kjer odstranimo besedam sklanjatev in množino.*  
*[Lemmatization is the process, where we delete from the words the declination and the plural.]*

If the sentence contains a meaningful hypernymy or meronymy relation it provides a knowledge-rich context, but may be insufficient to act as a definition. Such a borderline KRC sentence with the expressed ‘part-of’ relation is Example (lxxxviii). Sentences containing only a hypernym—the so called *exclusively genus* or *classificatory definitions*—or sentences where the *differentia* part is too imprecise are thus in our opinion KRC, borderline cases that we mainly tagged as non-definitions (see Example (lxxxix)).

- lxxxviii. *Luščenje leksikonov je sestavni del pri metodah statističnega strojnega prevajanja.*  
*[Lexicon extraction is an integral part of statistical machine translation methods].*
- lxxxix. *Iskanje informacij, besedil oz. dokumentov (angl. » information/text /document retrieval«) je najenostavnejša in najbolj pogosto uporabljena oblika tekstovnega rudarjenja.*  
*[Information/text/document retrieval (Eng. »information/text/document retrieval«) is the simplest and the most frequently used form of text mining.]*

A knowledge-rich context is also a sentence in which a term is illustrated by enumerating several instances of a concept, but is insufficient to be considered as an extensional definition. For this type of borderline cases, see Example (xc). Moreover, as the verb *mrgoleti* [*be rife with*] is expressive and typical of figurative language, it is not appropriate for forming definitions, as mentioned in Section 2.1.6.

- xc. *Na svetovnem spletu kar mrgoli tovrstnih virov, ki jih lahko najdemo s pomočjo splošnih iskalnih orodij, kakršni so Google, Altavista, Najdi.si itd.*  
*[The web is rife with such resources, which one can find with the help of general search tools, such as Google, Altavista, Najdi.si, etc.]*

Another borderline case, with wrong segmentation, providing a knowledge-rich context but not being considered a definition, is Example (xci). It contains information that could be, after manual refinement, reformulated into a functional definition.

- xci. *Računalniško podprto prevajanje in strojno prevajanje Obstaja kar nekaj prevajalskih orodij, ki skušajo vsaj deloma, če ne (skoraj) popolnoma avtomatizirati prevajalski proces.*  
*[Computer-assisted translation and machine translation There are several translation tools that try to, at least partly if not (nearly) fully, automate the translation process.]*

Another borderline case, which was evaluated as definition but merits a discussion, is Example (xcii). In this example two functions (functions *WordList* and *KeyWords*) of the corpus linguistics software *Wordsmith Tools* are defined by means of functional definitions. The definition is a borderline case for several reasons. First, the definienda are named entities, special tools within *Wordsmith Tools*, which is in itself a relevant question, but we have decided that since they are important for the domain that we model they should be considered domain terms to be defined in the glossary. Secondly, the sentence should be manually refined for being a real glossary entry since it should be split into two definitions, while the relative clause after the first noun phrase should be erased. This example shows the real difficulty of the task of automatic definition extraction from unstructured texts when for limited domain corpora we cannot opt to extract only ‘perfect’, well-formed definitions.

- xcii. *Orodje WordSmith Tools, ki je podrobneje predstavljeno v 5.2, s funkcijo WordList izdelava seznam pojavnih v posameznem korpusu, funkcija KeyWords<sup>47</sup> pa omogoča izbor ključnih besed, ki to besedilo ločijo od referenčnega korpusa.*

*[The WordSmith Tools, presented in more detail in 5.2, can with function WordList construct the list of tokens in a specific corpus, while the function KeyWords enables the selection of keywords that differentiate this text from the reference corpus.]*

In contrast, a borderline case defining one of the two definienda, but not evaluated as definition is Example (xciii). The reason for being evaluated as a borderline non-definition is that for defining the specific tool *Wordlist tool* the context of *Wordsmith Tools* should have been specified. However, it is still a knowledge-rich context sentence.

- xciii. *Orodje Wordlist nam omogoča vertikalni vpogled v korpus, se pravi vpogled v besedni inventar izbranih besedil.*

*[The Wordlist tool enables a vertical insight into the corpus, which means an insight into the word inventory of selected texts.]*

To conclude, the term-based approach strongly depends on the term extraction system, which does not perform perfectly. If the threshold parameter is non-restrictive, set e.g., at 10%, the list of terms that are not real domain terms is extensive. But even if it is set at 1% it may extract terms that are not specific to the language technologies domain (e.g., terms characteristic for scientific texts such as *chapter, figure, table, etc.*). On the other hand the system even at 10% does not include important domain terms (e.g., *stemming*). Another deficiency is that with less restrictive settings—not considering many of the syntactical characteristics of definition structures—the term-based approach is quite imprecise and accepts a large number of candidates. However, this deficiency can at the same time be considered an advantage as the approach may find definitions that could not have been extracted by a more restrictive pattern-based approach and manual filtering is fast and easy.

In summary, when developing the term-based definition extraction approach we performed a large set of experiments with different parameter settings (presented in Appendix A). By evaluating different parameter settings, we showed that a higher number of terms, multi-word terms, the nominative condition, as well as the verb

---

<sup>47</sup> In the original corpus, a footnote number stands after *KeyWords*.

condition, finding a term at the beginning of a sentence the first term being a multi-word expression) lead to higher precision in the majority of cases. Next, we identified some advantages of the term-based approach and illustrated the influence of different parameter settings on several definition candidates by examining also which types of sentences are excluded when applying different constraints. In the last part we discussed different types of borderline cases. The parameter settings should be tuned differently based on the task: more restrictive if we want the sentences to extract mainly the definitions, even though missing many, or less restrictive if we want the approach to be used in combination with other methods, more for eliminating out-of-domain sentences. In further work, we may consider also manual filtering of terms after the term extraction step, in order to improve the term-based definition extraction.

### 5.1.3 SloWNet-based definition extraction

This approach exploits the *per genus et differentiam* characteristic of definitions and therefore seeks for sentences where a term occurs together with its hypernym. For this task the sloWNet (Fišer and Sagot, 2008) lexical database, presented in Section 4.3.3 was used. We aim to extract sentences that contain a sloWNet term together with its direct hypernym. The problem is that sloWNet suffers from low coverage of terms specific to the language technologies domain.

In addition to the main condition that a sentence should contain a pair of sloWNet terms, where one term is a direct hypernym of the other, there is a further condition that there should be at least one word between these two terms. This additional condition prevents the extraction of sentences only because of the embedded terms; for illustration see an example from the English WordNet (Fellbaum, 1998): a two word term *computer system* already contains the word *system* which is a hypernym of *computer system*, so any sentence with the occurrence of *computer system* would have been extracted if the extra condition were not applied. On the other hand, we have a maximum window condition: we consider the window of maximum size seven, meaning that we consider a term pair relevant if there are a maximum of five other words in between.

# Extracted sentences	Precision (# of definitions in extract. sentences)	Recall estimate (# of def. in 150 recall test set)
4,670	0.057 (270)	0.2533 (38)

Table 12. sloWNet-based definition extraction results.

One of the disadvantages of this method is that it extracts also sentences that are completely out-of-domain (this can be avoided by using the method in combination with the term-based approach and not as an individual method, see Section 5.3.1). For example, in sentence (xciv) *dividenda* [*dividend*] and *dobiček* [*earning*] are in direct hypernymy relation, but are not relevant to the language technologies domain.

xciv. *Dividenda je v SSKJ definirana kot 'del dobička delniške družbe, ki ga dobi delničar na posamezno delnico'; v sami definiciji se nam tako pojavijo leksikalni elementi pojmovnega polja, ki sodijo skupaj: dividenda, dobiček, delniška družba, delničar, delnica.*

[In SSKJ<sup>48</sup> a dividend is defined as 'part of company's earnings that a shareholder gets for his share'; /.../]

<sup>48</sup>The general dictionary of Slovene language.

The extraction of out-of-domain sentences is partly due to the nature of the corpus itself. Since several articles and one doctoral thesis in our corpus deal with the construction of the Slovene wordnet, the hypernymy pairs are often given as examples in these texts. For instance, *daddy* and *grandpa* are hypernyms irrelevant for our domain, as shown in Example (xcv) below. This could be in the future filtered by a more detailed text preprocessing phase (that was for now more detailed only on the small subcorpus), taking the body text only and not the examples in the text.

xcv. *Zato smo v slovensko besedno mrežo leksem ata vključili dvakrat, enkrat v sinset {ata1, atek, ati, oče, očka, tata}, drugič pa v sinset {ata2, ded, dedek, stari ata, stari oče}.*

*[This is why we included in Slovene wordnet the lexeme ata<sup>49</sup> twice, once in synset {ata149, father, daddy /.../} and the other time in synset {ata249, grandfather, grandpa /.../}]*

Next, there are sloWNet hypernyms that are the cause for extracting many non-definitions. For example, the hypernymy relation between *figure* and *example* illustrated in Example (xcvi) or pair “*zaključek–poglavje*” [*conclusion–chapter*] in Example (xcvii) are very frequent cases (but could be avoided if list of terms to be defined was selected in advance).

xcvi. *Slika 10: Primer vnosa v MultiTerm z vsemi podatki.*

*[Figure 10: Example of MultiTerm input with all the data.]*

xcvii. *Zaključek bomo podali v petem poglavju.*

*[The conclusions are presented in the fifth chapter.]*

Another problem are irrelevant hypernymy pairs. Even if the Slovene wordnet follows the English WordNet structures and is linked to it, it was semi-automatically constructed and there are several occurrences of irrelevant hypernymy pairs. For instance in Example (xcviii), the sloWNet pair responsible for the extraction of the sentence is “*govor–rezultat*”, where the English translation of the pair is “*speech–result*”. The Slovene sloWNet IDs for this pair are 07081177 (*govor*) and 07069948 (*rezultat*), where the first ID corresponds to the English WordNet concept “*parlance*” or “*idiom*” with the definition “*a manner of speaking that is natural to native speakers of a language*”. The second (hypernym) term has three synonyms in Slovene, one of them being *rezultat* [*result*] but the corresponding English synonym literals are *formulation* or *expression* with the definition “*the style of expressing yourself*”. From this example we can see that the pair on the basis of which the sentence was extracted from our corpus “*govor–rezultat*” [*speech–result*] is not a relevant hypernymy pair. This type of errors could be avoided by working with manually validated subset of sloWNet.

xcviii. *Članek sklenemo z rezultati preskusa razumljivosti in naravnosti sintetiziranega govora.*

*[We conclude the article by results of testing the understandability and naturalness of synthesized speech.]*

In general, when analyzing the sloWNet pairs on the basis of which the definition candidates were extracted, we observed that only in few cases the wordnet pair corresponds to the *definiendum* and its hypernym. Much more frequent are the cases

---

<sup>49</sup> *Ata* is not translated in English for illustration purposes.

such as in Example (xcix)—which is itself a borderline case—where the sentence was extracted because of the wordnet pair “*language–text*” and is not related to the definiendum in the sentence (*transfer system*).

- xcix. *Transforni sistemi (ang. transfer systems): Čeprav so vsi prevajalniki na nek način transforni, se poimenovanje uporablja za jezikovno odvisne sisteme, pri katerih je rezultat analize abstraktna predstavitev (govorjenega) besedila v vhodnem jeziku, vnos za sintezo besedila pa je abstraktna predstavitev besedila v ciljnem jeziku.*

*[Transfer systems (Eng. transfer systems): Even if all translation systems are in a way transfer systems, the naming is used for language-dependent systems, where the result of the analysis is an abstract representation of (spoken) text in a source language, while the input for speech synthesis is the abstract text representation in the target language.]*

The same goes for the sentence below, where *translation memory* is defined, but the sentence was extracted based on a very general pair (“*part–unit*”). As already mentioned, sloWNet suffers from low coverage of terms specific to our domain.

- c. *Natančneje pa pomnilnik prevodov opiše Špela Vintar (Vintar 1998): »Pomnilnik prevodov je podatkovna zbirka prevodnih enot, navadno povedi ali krajših delov besedila, ki so v izvorniku in prevodu shranjeni v pomnilnik in so ob morebitni ponovitvi enakega ali zelo podobnega dela besedila na razpolago za ponovno uporabo.*

*[Translation memory is described in more detail by Špela Vintar (Vintar 1998): “Translation memory is a database of translation units, usually sentences or shorter text parts, that are saved as a source text and translation in the memory, and if the same or similar parts of text occur, they are available for reuse.]*

Even true hypernyms from the language technologies domain (pair “*strojno prevajanje–umetna inteligenca*” [*machine translation–artificial intelligence*”]) in Example (ci) give no guarantee that the sentence is a definition.

- ci. *Čeprav imajo semantične mreže dolgo zgodovino v filozofiji, sociologiji in jezikoslovju, so danes priljubljene predvsem v umetni inteligenci in za strojno prevajanje.*

*[Even though semantic networks have a long history in philosophy, sociology and linguistics, today they are popular especially in artificial intelligence and for machine translation.]*

Immediately when we switch to a more general domain, from computational linguistics to linguistics, we extract more relevant hypernymy pairs. Sentence (cii) can be considered a borderline case of extensional definition (since it is embedded in running text sentence), in which the enumeration of examples of *punctuation marks* is provided between parentheses. The hypernymy pairs for this example are “*period–punctuation\_mark*” and “*semicolon–punctuation\_mark*”.

- cii. *V zapisih besedil jih predstavljajo ločila (pika, vejica, podpičje, klicaj, vprašaj).*

*[In written texts, they are represented by punctuation marks (period, comma, semicolon, exclamation mark, question mark).]*

The next definition is extracted because of the hypernymy pair “*thesaurus–dictionary*”. We were interested in why the first word *corpus* was not extracted and found out that the only sloWNet synset containing the word *corpus* is in the military domain. The definition is also a borderline definition, because it does not contain any other

information about what is a corpus, except that it is an “electronic collection of texts”.

- ciii. *Korpus je računalniška zbirka besedil in je izrednega pomena za izdelavo jezikovnih orodij bodisi za slovarje, slovnice, črkovalnike, tezavre, terminološke banke bodisi za pomnilnike prevodov.*

*[A corpus is an electronic collection of texts and is extremely important for the construction of linguistic tools, either for dictionaries, grammars, spell-checkers, thesauri, terminological bases, or for translation memories.]*

In the next given example, the hypernymy pair for extracting the definition was “*homonym–word*”.

- civ. *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.*

*[Homonyms<sup>50</sup> are words that have the same sound representation but different meanings, between which there is no obvious metonymic or metaphorical relation and one cannot imagine that such a relation ever existed.]*

To get an insight into the domain coverage, we analyzed the first 20 terms that were automatically extracted by the term-extraction method presented in 4.3.2: *korpus* [*corpus*], *diskurzni označevalec* [*discourse marker*], *govorni signal* [*speech signal*], *strojno prevajanje* [*machine translation*], *slovenski jezik* [*Slovene language*], *jezikovni vir* [*language resource*], *jezik* [*language*], *besedilo* [*text*], *beseda* [*word*], *spletna stran* [*web page*], *besedna vrsta* [*part-of-speech*], *naravni jezik* [*natural language*], *govorna zbirka* [*speech database*], *pomnilnik prevodov* [*translation memory*], *besedna zveza* [*phrase*], *jezikovna tehnologija* [*language technology*], *razpoznavanje govora* [*speech recognition*], *referenčni korpus* [*reference corpus*], *prevajanje govora* [*speech translation*], *vzporedni korpus* [*parallel corpus*]. Out of these domain terms only 8 are covered by sloWNet. One term, *korpus* [*corpus*], exists in sloWNet but only in the military domain as shown in Example (ciii), so we do not count it as being covered, while one of the eight covered terms is the term *web page*, which is not completely from the language technologies domain.

In conclusion, we observe that sloWNet-based definition extraction has very low precision. We showed that the extracted terms with hypernymy semantic relation are not the ones we expected. One of the main deficiencies of the sloWNet-based approach is that domain coverage is very low and we can expect the method to provide better results if it were applied to more general domains. On the other hand, even if sloWNet’s main goal is to cover general domain, we could first extend sloWNet with domain specific terms (cf. Vintar and Fišer, 2013) or at least experiment with a mini proof-of-concept handmade ontology. In the future, sloWNet will also have better coverage, since new terms (e.g., domain specific terms, multi-word terms) are continuously added as shown in Vintar and Fišer (2009 and 2013); already today we could experiment with more recent versions of sloWNet. Moreover, the combination of wordnet-based definition extraction may yield better results in combination with a pattern-based approach or with syntactic analysis that could determine the position of terms between which the relation should occur.

---

<sup>50</sup> In Slovene, there are two words provided “*homonyms or homonyms*”, the first term being (*homonim*) being foreignism of the second Slovene term (*enakozvočnice*).

## 5.2 Extracting definitions from English texts

We demonstrate the potential of adopting the proposed methodology—initially developed for Slovene, which is the main focus of the thesis—to other languages. Therefore we also present methods for extracting definitions from English text corpora and analyze different types of definition sentences on a subcorpus of texts from the language technologies domain.

### 5.2.1 Pattern-based definition extraction

The pattern-based approach is the only of the three methods that is language dependent.

Based on Slovene patterns, we manually created a list of patterns for English by using adequate regular expressions. We compare two settings; in the first one the extra condition is set, that the pattern (usually starting with a NP) should occur at the beginning of a sentence (see Table 13) and the second one without this condition (Table 14). The beginning of the sentence criterion is used in order to increase the precision, since in English we cannot use the nominative case condition used to achieve the same purpose of better precision in Slovene. Moreover, in

Table 15 we also evaluate the setting in which some typical variations of sentence beginnings preceding a NP are evaluated (such as *According to* .\*, *In* .\*, *As* .\*); these variations were defined in order to improve the recall compared to the setting of Table 13. Except for the sentence beginning, which varies in the three tables, the tables explore the same seven pattern types; for easier reading, in

Table 15 we also write in bold a more intuitive understanding of each pattern type, valid for all the three tables.

At this stage, we have not used any chunker or parser, since we keep the method as similar to the Slovene pattern-based definition extraction as possible and demonstrate that it could be used also for other languages with no readily available sophisticated NLP tools. However, this will be revised in further work.

In developing the pattern-based approach, we defined different noun phrase structures with regular expressions, varying from a simple noun (e.g., *corpus*) to compound noun phrases, such as *machine translation* or *computational linguistics* and even longer noun phrases (e.g., *Association for Computational Linguistics*) composed of noun+preposition+adjective+noun. In all the patterns, the noun phrase can start with a determiner *a/an/the* or occur without it. We add the possibility that an alternative designation is used with another noun phrase introduced by *or* (NP or NP) and the same is true for adjectives within the noun phrases (e.g., ADJ or ADJ N). In the tables below, NP thus denotes all these varieties and ADV is used for adverb. For easier reading of the simplified patterns provided in tables below, note that “/” denotes alternatives (e.g., *is/are* matches *is* or *are* in the sentence, but not both) and “?” denotes zero or one occurrence of the preceding element. Two other symbols should be specified: “^” denotes the beginning of a sentence, while “.\*” denotes any number and type of elements. Words in single quotes (e.g., ‘is’) denote inflected word forms, while words without quotes are word lemmas covering all the possible inflected forms of the word (e.g., *refer* covers *refers*, *refer*, *referred*, etc.).

One can see that with setting the patterns at the beginning of the sentence, much higher precision can be achieved, i.e., 0.3292 in Table 13, compared to 0.1196 in Table 14, however one third less definitions are extracted compared to the setting without the beginning condition (273 definitions are extracted with all the patterns as shown in

Table 14 compared to 185 in Table 13; the estimated recall results calculated against the 150 definition test set are 0.2533 and 0.3733, respectively). Compared to the setting, in which a noun phrase occurs at the beginning of a sentence (Table 13), the recall can be slightly improved (at the cost of precision) if in addition to a NP at the beginning of a sentence other variations of a sentence beginning are accepted, as presented in Table 15. In this setting 200 definitions are extracted (precision is 0.2849 and recall 0.2733).

Pattern (beginning)	# Extract. sent.	Precision (# of def. in extr. sent.)	Recall estimate (# of def. in 150 recall test set)
1. ^NP 'is/are' NP	480	0.3312 (159)	0.2067 (31)
2. ^NP 'is/are'? ADV? refer to (as)? NP	14	0.4286 (6)	0.0067 (1)
3. ^NP 'is/are'? ADV? define ((in.*/by.*)?as)? NP	29	0.3103 (9)	0.0333 (5)
4. ^NP mean .* NP	14	0.2857 (4)	0.0067 (1)
5. ^the 'concept/word/term/naming/expression' NP .* use/describe/denote .* NP	4	0.25 (1)	0 (0)
6. ^the role of NP is to/NP	2	0.5 (1)	0 (0)
7. ^NP .* 'known' .* as .* NP	19	0.2631 (5)	0 (0)
<b>TOTAL (beginning)</b>	<b>562</b>	<b>0.3292 (185)</b>	<b>0.2533 (38)</b>

Table 13. Pattern-based definition extraction for English with the beginning of the sentence condition. Precision is evaluated on all the extracted sentences, while we used 150 definition test set for evaluating the recall.

Pattern (no beginning)	# Extract. sent.	Precision (# of def. in extr. sent.)	Recall estimate (# of def. in 150 recall test set)
1. NP 'is/are' NP	2,004	0.1098 (220)	0.3 (45)
2. NP 'is/are'? ADV? refer to (as)? NP	65	0.2461 (16)	0.0133 (2)
3. NP 'is/are'? ADV? define ((in.*/by.*)?as)? NP	102	0.147 (15)	0.04 (6)
4. NP mean .* NP	69	0.0869 (6)	0.0067 (1)
5. the 'concept/word/term/naming/expression' NP .* use/describe/denote .* NP	25	0.28 (7)	0.0067 (1)
6. the role of NP is to/NP	2	0.5 (1)	0 (0)
7. NP .* 'known' .* as .* NP	45	0.2667 (12)	0.0067 (1)
<b>TOTAL (no beginning)</b>	<b>2,283</b>	<b>0.1196 (273)</b>	<b>0.3733 (56)</b>

Table 14. Pattern-based definition extraction for English without the beginning of the sentence condition.

As already mentioned, In Table 14 patterns can occur anywhere in a sentence, Table 13 covers only the patterns occurring at the beginning of a sentence, while in Table 15 we tested other beginnings of sentences, specified in variations b), c) and d) in the first five patterns. The examples below illustrate these variations.

The sentence in Example (cv) begins with a simple noun phrase (*parallel corpora*) and corresponds to the pattern NP are NP. The second one (cv) provides the reference of the definition (*According to ISO 9126 software standards*) before the definiendum *usability*. In the same way the third example (cvii) defines the scope (*In corpus linguistics*). Example (cvii) uses the beginning "As .\*", which is the least productive, not correctly extracting nearly any definitions from our corpus, but should be retested on some other corpus in order to proof its (non-)usefulness.

- cv. *Parallel corpora are texts and their translations, human translations, we should add, or at least translations supervised and post-edited by translators who understood both the source and the target text.*
- cvi. *According to ISO 9126 software standards ([EAGLES96]) usability is a quality characteristic that is composed of three subcategories: • understandability • learnability • operability /.../*
- cvii. *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.*
- cviii. *As said previously, morphological analysis is the process of deducing the base form (lemma) from the inflectional forms of the words (word-forms), while morphological synthesis is the process of producing the inflectional forms given the base form.*

The last example is a borderline definition, since *morphological analysis* is a broader concept than simply “*deducing the base form from the inflectional form of the words*”, known as the process of *lemmatization*. We anyway tagged the sentence as a (borderline) definition, since in our computational linguistics domain, the definition remains valid, but in linguistics *morphological analysis* denotes a broader concept than just lemmatization, since it refers to the *morphology as the identification, analysis and description of the structure of a language’s morphemes and other linguistic units*. Therefore, the definition in our corpus defines the concept from the computational linguistics perspective but not from the broader linguistics perspective, where the definition would be considered to be too specific.

Next, we explain the seven pattern types and provide several examples of extracted definitions for each pattern, as well as some incorrectly extracted definition candidates. We can see that the majority of English definition sentences in our corpus are extracted by the first pattern, while other patterns still contribute to better recall.

The first pattern uses the structure “NP is/are NP”, meaning that the copula verb can be in third person singular or plural. The second pattern uses the lemma of the verb *refer to*, where also the passive constructions *is/are referred to as* can be used. The third pattern is based on the verb *define* (also with the possibility of passive construction) and the fourth pattern uses the verb *mean*. In the fifth pattern the lemmas *use*, *describe* and *denote* are used, but the definiendum is preceded by the introducing expression *the term, the word, the expression*, etc. The sixth pattern describes the role of the defined term (cf. functional definitions) and the last pattern “NP known as NP” in the majority of cases provides the *definiendum*’s synonym.

Pattern (variated beginning)	# Extract. sent.	Precision (# of def. in extr. sent.)		Est. Recall (# of def. in 150 recall test set)	
<b>1. NP is/are NP</b>					
a) ^NP 'is/are' NP	480	0.3312	(159)	0.2067	(31)
b) ^'A/according' to .* NP is/are NP	3	0.3333	(1)	0	(0)
c) ^in .* NP 'is/are' NP	88	0.0795	(7)	0.0133	(2)
d) ^as .* NP 'is/are' NP	28	0.0357	(1)	0	(0)
<b>2. NP refer to NP (active/passive)</b>					
a) ^NP 'is/are'? ADV? refer to (as)? NP	14	0.4286	(6)	0.0067	(1)
b) ^'A/according' to .* NP 'is/are'? ADV? refer to (as)? NP	0	0	(0)	0	(0)
c) ^in .* NP 'is/are'? ADV? refer to (as)? NP	4	0.25	(1)	0	(0)
d) ^as .* NP 'is/are'? ADV? refer to (as)? NP	0	0	(0)	0	(0)
<b>3. NP define NP (active/passive)</b>					
a) ^NP 'is/are'? ADV? define ((in.*/by.*)?as)? NP	29	0.3103	(9)	0.0333	(5)
b) ^'A/according' to .* NP 'is/are'? ADV? define ((in.*/by.*)?as)? NP	0	0	(0)	0	(0)
c) ^in .* NP 'is/are'? ADV? define ((in.*/by.*)?as)? NP	8	0.25	(2)	0	(0)
d) ^as .* NP 'is/are'? ADV? define ((in.*/by.*)?as)? NP	1	0	(0)	0	(0)
<b>4. NP mean NP</b>					
a) ^NP mean .* NP	14	0.2857	(4)	0.0067	(1)
b) ^'A/according' to .* NP mean .* NP	0	0	(0)	0	(0)
c) ^in .* NP mean .* NP	6	0	(0)	0	(0)
d) ^as .* NP mean .* NP	0	0	(0)	0	(0)
<b>5. the concept/word/term/naming/expression NP use/describe/denote (active or passive)</b>					
a) ^the 'concept/word/term/naming/expression' NP .* use/describe/denote NP	4	0.25	(1)	0	(0)
b) ^'A/according' to .* the 'concept/word/term/naming/expression' NP .* use/describe/denote .* NP	1	1	(1)	0	(0)
c) ^in .* the 'concept/word/term/naming/expression' NP .* use/describe/denote .* NP	3	0.6667	(2)	0.0067	(1)
d) ^as .* the 'concept/word/term/naming/expression' NP .* use/describe/denote .* NP	0	0	(0)	0	(0)
<b>6. the role of NP is to.... /the role of NP is NP</b>					
a) ^the role of NP is to/NP	2	0.5	(1)	0	(0)
<b>7. NP ... known as NP</b>					
a) NP .* 'known' .* as .* NP	19	0.2631	(5)	0	(0)
<b>TOTAL (variated beginning)</b>	<b>702</b>	<b>0.2849</b>	<b>(200)</b>	<b>0.2733</b>	<b>(41)</b>

Table 15. Pattern-based definition extraction for English with the beginning of the sentence condition – extended starting patterns. Precision is evaluated on all the extracted sentences, while we used 150 definition test set for evaluating the recall.

1. The first pattern uses the structure “NP is/are NP”. With this pattern we aim to extract the well-formed definitions in which the first noun phrase is usually the term to be defined and the second one its hypernym. We can see that both, singular and plural form of the verb *be* are used. See the examples below.

- cix. *"Resource-poor" languages are languages for which few digital resources exist; and thus, languages whose computerization poses unique challenges.*
- cx. *A text-to-speech system (TTS system) is an application that converts a written text into a speech signal.*
- cxii. *A syntactic parser is a tool that gives the structural composition of a sentence in the form of a tree.*

It is not in all the cases that the first noun after the verbal pattern represents a hypernym. In Example (cxii) the hypernym of *corpus linguistics* is not *sub-discipline* (a very general noun) but a noun from our domain, i.e., *linguistics*. This is a separate—although not infrequent—pattern, in terms of searching for hypernyms, but since the words *kind*, *discipline*, *sub-discipline*, are nouns, it still corresponds to the basic “NP is/are NP” pattern in terms of definition extraction.

- cxii. *Corpus linguistics is the sub-discipline of linguistics that deals with extracting language data from corpora and processing them for various applications such as grammars for human users and for computers, dictionaries, and lexicons.*

In some cases, the definition is incorrectly separated from the title; in these cases we evaluate that the definition is still correctly extracted, but needs minimal manual refinement as shown in the examples below (these cases are borderline definitions). In many cases this is due to preprocessing, and more examples are extracted when we do not limit the search to sentences where the pattern occurs at the beginning of the sentence (difference between Table 13 and Table 14). See some examples below:

- cxiii. *4 Speech Recognition Speech Recognition is a pattern classification problem in which a continuously varying signal has to be mapped to a string of symbols (the phonetic transcription).*
- cxiv. *1.1 Natural Language Generation Natural language generation is the task of producing natural language surface forms from a machine representation of the information.*
- cxv. *Considering Synonyms Synonyms are the words that have identical or very similar meanings but are written differently [1].*

With the analysis of the examples that differ in the two settings (applying the beginning condition or not), one could also extend the list of acceptable sentence beginnings provided in Table 15 (see Examples (cxvi) and (cxvii)).

- cxvi. *Accordingly, a topic ontology (Fortuna, 2007) is a hierarchical organization of documents' topics and their sub-topics.*
- cxvii. *For corpus linguistics, words are symbols with two aspects, the aspect of expression and the aspect of meaning.*

A less typical example is (cxviii). It is a borderline definition, since it contains two definitions, but the explanation is cyclic and thus only partially defining the concept. Note also that the order of the definiendum and the hypernym is inverted, since a *vertex* is a hypernym of *authority* and *hub*.

- cxviii. *A vertex is a good hub, if it points to many good authorities, and it is a good authority, if it is pointed to by many good hubs.*

It is clear that there are also non-definitions that are extracted by the pattern-based approach. For instance, references to *figures* or *tables*, as in Example (cxix), or very frequently occurring sentences referring to *examples* or *results*, such as appearing in Examples (cxx) and (cxxi), represent a recurrent category of incorrectly extracted definition candidates.

- cxix. *Figure 6 is an example of a record editing window and illustrates the presentation of lexical information in English (terms, links between the terms and attested contexts of usage) and the list of descriptive fields for a given term.*
- cxx. *The result is a list of place names occurring in the text with their offset and length, plus latitude and longitude, as well as information on the country they belong to and probably information about the hierarchical organisation of the country (e.g., town, province, region, country).*
- cxxi. *Examples are Van Gogh IS-A painter, Seles IS-A tennis player.*

But even when the sentences are not definitions, we often deal with knowledge-rich contexts, e.g.:

- cxxii. *The availability of semantic information is a crucial issue in the interpretation of texts, and therefore it is important for many tasks related with Natural Language Processing such as Information Extraction, Question Answering or Information Retrieval.*

In several cases a domain term is followed by a structure corresponding to one of the defining patterns, but the hypernym is too general and the rest of the sentence does not define the term (e.g., Example (cxxiii)).

- cxxiii. *Machine translation is a hard problem with highly structured inputs, outputs, and relationships between the two.*

A large range of sentences was incorrectly extracted because of the mathematical abbreviations, which are tagged as noun phrases (see Example (cxxiv)).

- cxxiv. *V is a sequence of vertices  $hhv, vl$ .*

We also have out-of-domain sentences, such as the one below.

- cxxv. *A dog is a kind of dog.*

2. For the second pattern using the verb *refer to*, a good example is for instance:

- cxxvi. *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.*

The passive structure can be used as in Example (cxxvii), where the adverb *usually* also illustrates the usefulness of an optional adverb in the pattern.

- cxxvii. *The identification of words—and punctuation marks—is usually referred to as tokenisation, while determining sentence boundaries goes by the name of segmentation.*

One of the reasons for extracting non-defining sentences is that a sentence complies with a pattern but is too specific to be used as a definition. See Example (cxxviii), where *accuracy* is defined only for a specific part-of-speech tagging application.

- cxxviii. *Accuracy refers to the percentage of words (i.e. word tokens) in a corpus which are correctly tagged.*

Another incorrectly extracted sentence is (cxix). The anaphora *these two factors* indicates that the definition is spanning over several sentences, but at the sentence level we cannot consider it being a definition. For instance, in the text preceding the extracted sentence, the defining context of term *faithfulness* is given: *a translation should be true to the meaning of the original.*

- cxix. *These two factors are often referred to as faithfulness and fluency.*

3. Definitions extracted by patterns with the verb *define* are illustrated in Example (cxxx).

- cxxx. *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 /.../ " .*

4. Definitions extracted by patterns with the verb *mean* are illustrated in Examples (cxxxii) and (cxxxiii).

- cxxxii. *Localisation means not just translation into the vernacular language, it means also adaptation to local currencies, measurements, and power supplies, and it means more subtle cultural and social adaptation.*

- cxxxiii. *Backup means taking a periodic copy of a file store.*

5. The next pattern has one of the highest precision scores, since it introduces the definiendum by expressions such as *the term*, *the expression*, etc., in addition to using the defining verbs *denote* or *describe* or the more general verb *use*. The definitions extracted by this pattern are illustrated by Examples (cxxxiii) and (cxxxiv).

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

- cxxxiv. *We use the term "domain knowledge" in the way commonly used in machine learning or inductive logic programming as knowledge that is not or cannot be expressed by learning examples themselves.*

6. The next pattern using the expression *the role of* covers only one definition from our corpus and even this example is a borderline case, since it is too specific (given the low performance, in further developments we should therefore test this pattern on some other corpora and omit this pattern from the pattern list if its use does not prove to be useful).

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

7. To conclude, we provide some definitions extracted by the last pattern. The pattern “NP .\* known .\* as NP” mainly extracts the *definiendum*’s synonyms, which is not necessarily sufficient for a sentence to be classified as a definition (and also depends on the interpretation of what is a definition). However, as shown by the precision score, it is a relatively good indicator for finding a defining sentence. We provide two examples where after a synonymous expression a partitive concept definition (Example (cxxxvi)) or a functional definition (Example (cxxxvii)) are given. The last example (cxxxviii), on the other hand, shows that just alternative namings are not enough to define a concept, illustrating why non-definitions are also extracted by this pattern.

cxxxvi. *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.*

cxxxvii. *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.*

cxxxviii. *As such, in Microsoft applications, UTF-16 is known simply as "Unicode", while UTF-8 is known as "Unicode ( UTF-8)".*

We can see that the precision is quite high regarding the complexity of the corpus we are dealing with, but we still extract only one quarter of the definitions, therefore other methods are examined to improve the recall.

### 5.2.2 Term-based definition extraction

The method has been introduced in Section 5.1.2 in the context of definition candidates extraction from Slovene corpora.

The basic hypothesis is the same as in the Slovene counterpart. The basic condition for a sentence to be extracted as a definition candidate is to contain at least two domain terms. With this—very loose—setting we aim to extract knowledge-rich contexts, not necessarily the definitions. In combination with other methods, this simple approach can also be used as a domain filter. The first step is term extraction, where a weighting measure compares normalized relative frequencies of single words in a domain-specific corpus with those in a general reference corpus. A detailed description of the term-extraction methodology is described in Vintar (2010) and in Section 4.3.2, while our workflow implementation of the service is presented in Section 6.4.2.

If we want to extract fewer definition candidates, but achieve better precision, additional constraints should be applied. In contrast to the Slovene term-based approach, we cannot use the morphosyntactic tag-based nominative condition, since the

English language does not express the cases in the same way. Therefore, for English we tested the same additional hypotheses as in the Slovene counterpart (except for the nominative condition): the influence on precision (and recall) of the verb condition, the beginning of the sentence condition, the condition of the first term being a multi-word expression, as well as the influence of the main parameter settings, i.e., the threshold parameter, the number of terms and the number of multi-word terms.

The results of different experiments are provided in Table 16. General trends are the same as in Slovene definition extraction, i.e., the higher the threshold, the number of terms and multi-word terms, as well as the number of additional constraints, the higher the precision and the lower the recall. The highest precision of term-based definition extraction is around 13% (Experiments (j) and (w)) but the number of extracted definitions is very small for these settings. If we want a considerable number of definitions, we get much lower precision, e.g., 0.0824 and 90 definitions extracted in Experiment (m).

Note that the results are much worse than for Slovene (see the results of Slovene term-based definition extraction in Table 23 and Table 24 of Appendix A). Already a brief comparison shows not only that the best results for Slovene applying the nominative condition are much better in terms of precision reaching up to 0.2647 accuracy for the 1% threshold setting and up to 0.1381 for the 10% threshold setting, but also that without the nominative condition the results for Slovene can reach more than 20% precision.

By comparing different parameter settings of Table 16 we observe that if we use the additional *verb*, *multi-term first* and *beginning of the sentence* conditions the precision continuously increases with the number of terms and that if the verb figures between the first two terms (VF setting) results are better than if the verb occurs between any terms (VA). The continuous growth of precision can be observed for the 1% threshold setting in Experiments (a)–(e) and (h)–(j) with the best precision result of 0.1295—without applying the number of multi-word terms parameter—however extracting only 29 definitions and 0.0533 estimated recall (Experiment (j)). If we want to extract more definitions, a better solution is to use the 10% threshold, where we can also see that by applying the condition of minimum 4 domain terms in a sentence results in lower precision than the 5 terms condition (0.0755 precision for 4 terms in Experiment (g) extracting 119 definitions) compared to 90 extracted definitions with 5 terms (0.0824 precision in Experiment (m)).

Additional conditions were verified and as already noted the VF setting leads to better precision compared to the VA condition setting (e.g., pairs of experiments (b)–(c); (d)–(e); (v)–(w) and all other experimental settings). Not applying any of the two verb condition variations was tested in Experiments (h) and (x), showing that—similarly as in the evaluation of term-based definition extraction for Slovene—the verb condition generally improves the precision without significantly decreasing the recall. The beginning of the sentence condition and the condition of the first term being a multi-word expression both positively influence the precision (at the cost of decreased recall): if one of the two constraints is not applied, the precision decreases from 0.1295 in Experiment (j) to approximately 0.05 precision in Experiments (k) and (l), confirming the hypothesis that these two conditions are very useful if applied together.

By increasing the number of requested multi-word terms the precision increases (e.g., the best for 5 terms above 1% threshold, 3 multi-terms and all the additional constraints) up to 0.1392 in Experiment (w), but the number of definitions is very low. Also in all the other cases the precision is higher when the number of multi-word terms

is applied (e.g., pairs of experiments (a)–(n), (c)–(p), (e)–(t), (j)–(w), (q)–(r)) and when increasing the number of multi-word terms (cf., Experiments (q)–(r)). However, in practice we may opt for settings without the multi-word term condition given that the increased precision comes at the cost of decreased number of extracted definitions, e.g., in Experiment (m) only 48 definitions are extracted compared to 90 definitions extracted in Experiment (z).

There are several reasons why the results of definition extraction for English are lower than for Slovene. First, in English the nominative condition cannot be applied, while this feature significantly improved the precision of definition extraction in Slovene. Next, the background technologies used (the ToTrTaLe morphosyntactic tagger and lemmatizer, and the LUIZ term extractor) were mainly designed for Slovene and not for English. In future work other state-of-the art preprocessing tools, such as Tree Tagger (Schmid, 1994), will be considered, and other term extraction systems can be used for English than for Slovene (e.g. Sclano and Velardi, 2007). One of the reasons for differences in Slovene and English results could be also the smaller number of extracted terms in the term extraction step for English.

Even if the precision results are not satisfactory, we can still observe interesting definition structures. Different verb types for introducing definitions are used and not only formal definitions with a term and its hypernym are extracted. This is illustrated by Examples (cxxxix) and (cxli). In Example (cxxxix) the terms *translation memory* and *translation process* are related and even if there is no hypernymy relation, the sentence clearly defines the *translation memories*. A more straightforward definition for the same term is sentence (cxli).

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

cxli. *A Translation Memory (TM) is a type of computer-aided translation tool which stores previously translated texts alongside their corresponding source texts (ST) and allows translators to 'recycle' or re-use these texts, or parts of them, in new translations.*

Both sentences are extracted even if several restrictions are used, since they contain multi-word terms listed in top 1% of the terms, such as *translation memory* and *translation process* in (cxxxix) or *translation memory*, *translation tool* and *source text* in (cxli), there is also a high number of domain terms in both examples, a term figures at the beginning of the sentence (where the beginning is considered as the first or the second position, in that way accepting e.g., the article *a* having the first position) and the first appearing term is a multi-word. We can see that if the second sentence uses the standard “is-a” relation, where the hypernym is introduced after the expression “is a type of”, the first sentence defines the term by its purpose (*functional definition*).

Two other examples where the term is defined by its purpose and the sentence does not contain the “is-a” pattern or its variation are listed below. Notice that (cxli) was extracted only by 10% threshold settings since the main term *information retrieval* is not included in the 1% term list.

cxli. *Information retrieval is concerned with locating documents relevant to the user's information needs from a collection of documents.*

	Threshold	Terms (#)	Verb (VA-VF)	Beginning sent.	Multi-word first	Multi-word (#)	Extract. sent.	Precision (and # of definitions)	Recall on 150 test set (# of def.)
Beginning and First multiword term									
a)	1%	2	yes	yes	yes	no	803	0.0772 (62)	0.1133 (17)
b)	1%	3	yes-VA	yes	yes	no	657	0.0852 (56)	0.09333 (14)
c)	1%	3	yes-VF	yes	yes	no	567	0.0935 (53)	0.09333 (14)
d)	1%	4	yes-VA	yes	yes	no	475	0.0947 (45)	0.08 (12)
e)	1%	4	yes-VF	yes	yes	no	365	0.1068 (39)	0.0733 (11)
f)	10%	4	yes-VA	yes	yes	no	1,904	0.0693 (132)	0.1933 (29)
g)	10%	4	yes-VF	yes	yes	no	1,576	0.0755 (119)	0.1667 (25)
h)	1%	5	no	yes	yes	no	336	0.0952 (32)	0.06 (9)
i)	1%	5	yes-VA	yes	yes	no	319	0.1003 (32)	0.06 (9)
j)	1%	5	yes-VF	yes	yes	no	224	0.1295 (29)	0.0533 (8)
k)	1%	5	yes-VF	no	yes	no	1,003	0.0518 (52)	0.0867 (13)
l)	1%	5	yes-VF	yes	no	no	1,461	0.0513 (75)	0.16 (24)
m)	10%	5	yes-VF	yes	yes	no	1,092	0.0824 (90)	0.1333 (20)
Number of multiword terms									
n)	1%	2	yes	yes	yes	2	388	0.0928 (36)	0.0667 (10)
o)	1%	3	yes-VA	yes	yes	2	349	0.1003 (35)	0.06 (9)
p)	1%	3	yes-VF	yes	yes	2	306	0.1078 (33)	0.06 (9)
q)	10%	3	yes-VF	yes	yes	2	1,480	0.0757 (112)	0.1933 (29)
r)	10%	3	yes-VF	yes	yes	3	779	0.0821 (64)	0.0933 (14)
s)	1%	4	yes-VA	yes	yes	2	278	0.1079 (30)	0.0533 (8)
t)	1%	4	yes-VF	yes	yes	2	219	0.1187 (26)	0.0467 (4)
u)	10%	4	yes-VF	yes	yes	2	1,173	0.0793 (93)	0.14 (21)
v)	1%	5	yes-VA	yes	yes	3	108	0.1111 (12)	0.0267 (4)
w)	1%	5	yes-VF	yes	yes	3	79	0.1392 (11)	0.0267 (4)
x)	1%	5	no	yes	yes	3	111	0.1081 (12)	0.0267 (4)
y)	10%	5	yes-VA	yes	yes	3	683	0.0791 (54)	0.0733 (11)
z)	10%	5	yes-VF	yes	yes	3	550	0.0872 (48)	0.0733 (11)

Table 16. Term-based definition extraction on the English part of the corpus.

- cxlii. *In speech recognition, the objective is to predict the correct word sequence given the acoustic signals.*

The term-based approach extracts other definition structures, complementary to those listed in the pattern-based approach, e.g.:

- cxliii. *Classifying new data according to the closest training data example is often called nearest neighbor classification.*

Another advantage of this method compared to the pattern-based approach in terms of recall is that if the beginning of a sentence has a complex structure due to alternative denominations or additional given references, the sentence is still extracted, see sentences (cxliv) and (clxiv).

- cxliv. *POS Tagging Part-of-speech tagging (POS tagging or POST), also called grammatical tagging, is the process of marking words in a text that correspond to a particular part of speech, based on both their definition and their context, i.e. the relationship between adjacent and related words in a phrase, sentence, or paragraph.*
- cxlv. *Machine learning can be defined (after Witten & Frank (2002)) as the (semi)-automatic discovery of common patterns in a substantial amount of data, which results in the emergence of meaningful new information that was previously not apparent.*

Named entities represent more borderline cases, since their terminological value depends on specific application. However, if we consider the *Language Technology team of the Joint Research Centre* as a domain term, the definition of the aim of the team is a valid definition (cf. cxlvi). Specific parts of a program, such as that denoted by the term *Automatic Rule Refiner* were, in the majority of cases, not evaluated as terminological expressions and thus the definition sentences in which they occur were marked as non-definitions (cf. cxlvii).

- cxlvi. *The Language Technology team of the Joint Research Centre (JRC) has the aim to produce a number of text analysis applications for ideally all official EU languages (and more) that help users to navigate in large multilingual document collections and that provide them with cross-lingual information access.*
- cxlvii. *The Automatic Rule Refiner is the one in charge of executing linguistic rule manipulation inside the MT system.*

Extensional definitions (defining a term by listing all or at least the representative examples of a definiendum category) can be written as a list in the parentheses (see Example (cxlviii)) or after a colon punctuation mark (Example (cxlix)). These types of definitions (sometimes embedded in a longer sentence) have a higher number of domain terms.

- cxlviii. *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.*
- cxlix. *A dialog system typically is composed of the following components: a speech recognizer, a natural language understanding module, a dialog manager, a natural language generation module and a speech synthesizer.*

Especially with less restrictive methods, many extracted sentences are not definitions, which is not surprising since the conditions of the term-based approach are very loose. See for example the two sentences below, where terms such as *evaluation* and *method* are domain terms, but the sentence is not defining a term (*automatic evaluation* or *translation model*).

- cli. *Automatic evaluation shows that the method works best for nouns, which is why we focus on them in the rest of this section.*
- cli. *The translation model is based on word alignment.*

A large number of extracted sentences, even if not being definitions represent knowledge-rich contexts:

- clii. *Speech synthesis starts with the linguistic analysis of the input text, where the orthographic text is converted into phonemic representation.*
- cliii. *Translation model was trained using GIZA ++ (Och and Ney, 2003).*
- cliv. *Statistical language models can be applied in several tasks of language technologies (Manning & Schütze, 1999), including automatic speech recognition, optical character and handwriting recognition, machine translation, spelling correction.*
- clv. *Language model plays an important role in statistical machine translation systems.*

We also encounter non-definition sentences that even provide a hypernym for the term to be defined but lack listing specific properties of a definiendum:

- clvi. *String kernels (§ 4.3.1.3) are a general and efficiently computable similarity measure that is smoother than edit distance.*

There are examples with expressed hypernyms that are borderline cases closer to definitions (e.g., clvii) and those closer to non-definitions, since the hypernym is too general (e.g., non-informative hypernym *important step* in clviii).

- clvii. *Memory-based machine translation, as described in van den Bosch et al. (2007), is considered a form of Example-Based Machine Translation.*
- clviii. *Part of speech tagging, often referred to as tagging, is an important step in many language technology systems.*

Some examples are classified as borderline definitions because they are too specific to be very good definitions, e.g.:

- clix. *Symbol error rate—in our case, phoneme error rate—is based on the minimal number of edit operations (insertions, deletions and substitutions) necessary to transform the predicted string into the reference string, ignoring the reference alignment.*

As in the Slovene term-based definition extraction, one of the main reasons for extracting non-definitions is that numerous terms that are too general are extracted as terminological candidates. These expressions are e.g., *basic idea*, *future work* or *related work*, figuring between top 10% of terms. Slightly more justified but still not specific for this domain is the term *evaluation* that is in the first 1% of terminological candidates (there are many other terminological candidates that are not good terms and figure in top 1% of domain terms, such as *result*, *good result* (base form of *best result*), *chapter*,

*figure, model, table*. For example, this is one of the 18 sentences extracted because of the term *future work* (with four domains and the “verb”, the “multi-word first” and the “beginning of the sentence” conditions):

- clx. *As future work we plan one evaluation of the different methods to obtaining information, as well as a comparative between all of them, with particular attention in how the information is received and adapted.*

One of the reasons for low precision can therefore be the term extraction step (cf. Section 4.4.2). If evaluated by the same person (cf. A1 columns of Table 7) LUIZ performs worse for English than for Slovene (in terms of precision). As shown above even a term specific to scientific writing but not a real domain term can lead to extracting an important number of definitions. In further work this observation should be systematically examined and the solution of possible manual filtering of the list of terminological candidates by the user before proceeding to the term-based definition extraction step should be applied. The second reason for low precision is the morphosyntactic tagging which results in many annotation tagging mistakes when applied to the English corpus. Therefore in further work, we should consider replacing the ToTrTaLe annotation tool with one of the main taggers for English, e.g., TreeTagger (Schmid, 1994).

On the other hand, compared to the pattern-based approach, the term-based approach generally handles morphosyntactic tagging errors better than the pattern-based approach; while the pattern-based approach is highly dependent on the morphosyntactic annotation, the term-based approach depends on its quality only in the term extraction step (whose results could also be improved by additional manual filtering) and in applying the verb condition. See for example, sentence (clxi) that should not have been extracted because of the verb condition, but remains in the set of definition candidates since the author *Nivre* was wrongly tagged as a verb. Similarly, in Example (clxii) proper noun *Brown* is also tagged as a verb.

- clxi. *The Malt parser (Nivre et al., 2004), with a gold standard of 10,000 words.*
- clxii. *Automatic machine translation system construction in case of corpus-based machine construction systems such as Statistical Machine Translation (SMT) (Brown et al., 1993; Och and Ney, 2003) or Example-Based Machine Translation (EBMT) (Nagao, 1984) and (Hutchins, 2005)*

In conclusion, the term-based approach has lower precision than the pattern-based approach. As shown, the errors can often be attributed to the term extraction or morphosyntactic errors, as well as quite loose constraints (depending on the settings). However, it provides an interesting complementary method to extract knowledge-rich sentences. Additional usefulness of this method is that, since it requires no manually crafted verb patterns, we can use it to analyze a larger variety of definitions in running text, which is a very interesting task from the linguistic perspective. If the corpus does not offer any sentence corresponding to a well-formed definition extracted by patterns, these term-based candidates are usually still precious knowledge-rich contexts.

### 5.2.3 WordNet-based definition extraction

In this approach, the same conditions were used as for the sloWNet-based definition candidates extraction, meaning that the sentence should contain at least two terms which are identified in WordNet (PWN, 2010) as one being a direct hypernym of the other. As in the settings for the experiments on the Slovene dataset, the additional conditions are that there should be at least one word between the identified terms in order to avoid extracting embedded term pairs, and that window size 7 is used, i.e., a maximum of 5 words can be between two terms in order to consider them a relevant WordNet pair.

Using the WordNet-based method, we have extracted 3,258 candidates that we evaluated and the results are 0.043 precision (140 definitions were correctly extracted) and 0.2666 recall, the latter being estimated on the 150 definitions test set.<sup>51</sup>

It is essential to qualitatively analyze these results, and to identify the reasons for such low precision.

Number of extracted definition candidates	Precision	Recall (on 150 definition test set)
3,258	0.043 (140)	0.2667 (40)

Table 17. WordNet-based definition extraction candidates

Firstly, we provide some examples of correctly extracted definitions based on WordNet direct hypernyms. The system extracts sentence (clxix) due to the correctly identified hyperonymy pair “*metadata-data*”, whereas in sentence (clxx) *machine translation* is identified as a hyponym of *computational linguistics*. We can see that in contrast to the pattern-based approach the use of adverbs in the middle of the defining patterns does not harm the extraction of the definitions (for instance in the example below the adverb *usually* occurs in the middle of the pattern *is defined as*). The variety of defining verbs is much larger than when the verbs are enumerated in a predefined list (e.g., *remains* in the second example is not a typical verb used in definitions).

- clxiii. *Metadata is usually defined as 'data about data'.*
- clxiv. *Machine translation remains the sub-division of computational linguistics dealing with having computers translate between languages.*

Despite the fact that the English WordNet is much bigger than its Slovene counterpart sloWNet, the specific domain coverage remains low. For instance, definition in Example (clxv) is not extracted by the pair “*text\_mining-data\_mining*”, since WordNet does not include the term *text mining*, but only the term *data-mining*. The pair based on which the sentence was extracted is “*pattern-model*”. Similarly, in Example (clxvi) the good domain term to be defined *lemmatisation* is not a WordNet literal, but the sentence is extracted anyway, because of the pair “*form-word*”.

- clxv. *Text mining (Feldman and Sanger, 2007) is a variant of data mining in which models and patterns are extracted from unstructured natural language text.*
- clxvi. *Lemmatisation is the process of finding the normalised forms of words appearing in text.*

---

<sup>51</sup> The preliminary evaluation on 100 randomly selected sentences was much better (see e.g., results of a similar settings in Pollak et al., 2012), which proves that 100 sentences are definitely not enough to have a good precision estimation.

In Example (clxvii), the pair of WordNet hypernyms is “*linguistics-science*”. Even if the definiendum is in fact *corpus linguistics* and not *linguistics*, the system correctly extracts the definition, and the hypernym *science* is a valid hypernym also for *corpus linguistics*. As already mentioned the domain coverage is not sufficient, since the term *corpus linguistics* is very representative for our corpus and does not figure in WordNet. In WordNet we find different compound nouns including linguistics, such as *computational linguistics*, *descriptive linguistics*, *diachronic linguistics*, *historical linguistics*, *prescriptive linguistics*, *structural linguistics* or *synchronic linguistics*, but not *corpus linguistics*. Moreover, note that the sentence starts with *because*, which should be deleted in manual definition refinement, but the rest of the sentence is a definition. These types of examples are not the best definition sentences, since they cannot be included in the definition repository without any changes, but need manual intervention. However, we decided to classify them as definitions, since the manual refinement does not need any expert knowledge.

- clxvii. *Because corpus linguistics is an empirical science, in which the investigator seeks to identify patterns of linguistic behaviour by inspection and analysis of naturally occurring samples of language.*

As in the Slovene part, a large number of sentences are incorrectly extracted. A very common non-definition type of sentence is the one containing two terms in hyperonymy relation, but not defining any of the two terms. For example, sentence (clxiii) contains the words *word* and *language* but does not define any of them.

- clxviii. *Indeed, it does not work for most of the words that make up our general language vocabularies.*

To better understand the complexity of definitions in running text as well as the WordNet hypernymy condition, consider the following sentence:

- clxix. *The European Union today has 15 Member States, and 11 official languages (Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish).*

Firstly, the sentence is not a definition, since it is “out-of-date”. Even if the sentence were a valid extensional definition when the article was written, today there are more official languages in European Union and the definition is no longer true. Secondly, the term *official language* is not completely relevant for our domain. Anyhow, one would expect that the sentence would be extracted based on the general term *language* and the specific languages (*Danish, Dutch...*). However, the identified hypernymy pair was “*union-state*”, since the specific languages are not direct hyponyms of *language*. For instance, the direct hypernym for Greek is *Indo-European language* and there are two more levels before reaching the node *language*: *Indo-Hittite* and *natural language*. Even if the sentence above is not a definition, this example raises the question of loosening the hypernymy condition by taking into consideration all hypernyms of a term and not just a direct one, but since the precision is already very low, increasing the recall at the price of precision is not a good solution.

Many sentences may not qualify as definitions, yet represent knowledge-rich contexts. The embedded candidate definition for *specialization* in sentence (clxx) is not precise enough for defining the concept. In this sentence we can also again notice that the sentence is extracted based on a different term pair than the definiendum and its

hypernym (the detected WordNet pair in the sentence is “*database-information*”).

- clxx. *First of all, he is responsible for managing all language-independent data, which are identified by the term «concept» within a given field of specialization, i.e. the area to which the concept belongs, the possible semantic links that connect the concept to other concepts in the database as well as the information illustrating the concept.*

In Example (clxxi) a knowledge-rich context is extracted, i.e., *word-based statistical machine translation* is attributed the hypernym *machine translation system*, even if the WordNet pair is a more general pair “*system-method*”. Also in the next two sentences (clxxii) and (clxxiii) we find knowledge-rich contexts. In the first one, the *named entities recognition and categorisation* is illustrated by *recognition of names of streets*, etc. but the sentence is not a definition. The pairs based on which the extraction was performed are again not from the language technologies domain, but “*boulevard-street*” and “*avenue-street*” in the first sentence and “*corpus-part*” in the last one.

- clxxi. *Word based statistical machine translation has emerged as a robust method for building machine translation systems.*
- clxxii. *Recognition of names of the streets, boulevards, avenues, roads, etc., is an integral part of the problem of recognition and categorization of named entities (Chinchor et al, 1999).*
- clxxiii. *The spoken corpus is a very important part of the national linguistic infrastructure.*

Extracting knowledge-rich contexts instead of definitions can be based also on more domain specific concepts, such as “*divergence-variant*” in the first example below and “*word-language*” in the second. We notice that both examples do not have a hypernym:

- clxxiv. *Kullback-Leibler Divergence (and Symmetrized Variant) The Kullback-Leibler divergence is specific to probability distribution.*
- clxxv. *Inflectional languages usually have free word order.*

Even a correct hypernym does not necessarily provide a definition. In Example (clxxvi) the provided knowledge-rich context is not enough for defining the *Hungarian language*, since one does not get the necessary information about where the language is spoken, by how many people or what is the language origin:

- clxxvi. *Hungarian is an agglutinative, free word order language with a rich morphology.*

Further, there are many sentences which are not defining in any aspect (in Example (clxxvii), it is the pair “*document-representation*” that is the cause for extraction).

- clxxvii. *Size of the representation of a document collection.*

Arbitrary sentence extraction based on WordNet hypernyms can be seen also from the out-of-domain sentence below, where the WordNet pair is “*time-example*”.

- clxxviii. *For example, for a long time in every newspapers article in Serbia it was common to use the word Kosovo for the Serbian province called Kosovo i Metohija (Kosovo and Metohija).*

To sum up, many examples are good definitions with correctly identified hypernyms (such as hypernymy pairs “*morpheme-linguistic\_unit*” and “*word-linguistic\_unit*” in definition (clxxix)). These are the most informative examples, because in addition to defining a concept, they already extract a hypernym from the manually crafted resource

WordNet, allowing for better understanding of the domain and providing an excellent starting point for domain modeling. Other definitions are based on very general domain hypernyms, such as “*language-text*” in Example (clxxx), but still provide meaningful definitions. As previously discussed in the thesis, the distinction between a definition and non-definition is not always clear and there are many borderline cases. For instance, a sentence that is not general enough to be a perfect definition was as a borderline case still classified as a definition in (clxxxii) however an equally borderline case in (clxxxii) was rejected. In the first example *morphological tagging* is defined as *the process of assigning morphological information to a word*, and even if the listed categories (e.g., the case) are not language independent, since not all the languages have cases for example, the sentence gives a good idea of the *morphological tagging* process. In the second example (clxxxii), the candidate for extensional definition can be valid for some languages, but since it is not generally valid, it was classified as a non-definition, despite the fact that it provides a very relevant knowledge-rich context.

- clxxx. *A morpheme is the smallest semantically meaningful linguistic unit from which a word is built.*
- clxxx. *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.*
- clxxxii. *“Morphological tagging” is the process of assigning POS, case, number, gender, and other morphological information to each word in a corpus.*
- clxxxii. *There are three moods: the indicative, the imperative, and the conditional.*

Other discussed examples of borderline cases which can be classified as definitions are those explaining mathematical formulas as in Example (clxxxiii) or extensional definitions providing all or typical realizations of a concept as in Example (clxxxiv). In the two sentences the definition is embedded in a non-definitional sentence (we underline the definition part). In Example (clxxxv) the proper noun used for the name of the tool, *SimFinder*, is well defined but it could be further discussed whether the concept should be part of a definition lexicon or not (the category of borderline definitions of named entities).

- clxxxiii. *Since a conditional probability can be expressed as a joint probability divided by a marginal probability like, for example, in equation (4.1), deriving a conditional model from a joint model usually requires first deriving a marginal model.*
- clxxxiv. *However, also language checking tools such as spell checkers, grammar and style checkers have to meet the user's requirements and should easily be extensible to the specialised terminology, the text structure properties and the in-house writing conventions.*
- clxxxv. *In the context of multidocument summarization, *SimFinder* (Hatzivassiloglou et al., 1999) identifies sentences that convey similar information across input documents to select the summary content.*

Sentence (clxxxvi), which explains the concept of *sublanguage* through examples, was not labeled as definition, neither was the sentence (clxxxvii) where a candidate definition is introduced by *i.e.* but in our opinion does not define the concept (for comparison look at the definition of *information filtering system* from Wikipedia:

“An Information filtering system is a system that removes redundant or unwanted information from an information stream using (semi)automated or

computerized methods prior to presentation to a human user. Its main goal is the management of the information overload and increment of the semantic signal-to-noise ratio).”

- clxxxvi. *It may be restricted (or adapted) to a particular domain or sublanguage – the language of medicine is different from the language of engineering and the language of theatre criticism, etc. – by means of the definitions and constraints specified in the databases of domain or sublanguage information.*
- clxxxvii. *An area where MT is already involved is that of information filtering (often for intelligence work), i.e. the analysis of foreign language texts by humans.*

To conclude, even if the method provides a huge amount of knowledge-rich contexts, it does not provide a lot of definitions compared to the number of extracted sentences. This is due to the very low domain coverage of WordNet and we can expect the method to be more relevant for more general domains. We achieved better results on popular science texts (cf. Fišer et al., 2010), which are also known to have better WordNet coverage (e.g., Schmied, 2007). In further work, additional experiments will be performed in order to see how the preprocessing with a chunker/parser or limiting the WordNet pairs to domain terms only could improve the results.

### 5.3 Results of Slovene and English definition extraction methods and their combinations

In this section we summarize the results of definition extraction methods described in the two previous sections and examine the possibility of combining different methods in order to improve the definition extraction results in terms of precision or recall. The quantitative evaluation of the results of method combination is presented in Sections 5.3.1 and 5.3.2 for Slovene and English, respectively. On the other hand, the qualitative evaluation is discussed in Section 5.3.4 in which a systematization of definition types and problems related to extracting definition candidates from running text is provided.

#### 5.3.1 Combining different approaches on the Slovene subcorpus

Let us first summarize the results of definition extraction methods on the Slovene subcorpus (see Table 18).

- Using the pattern-based approach we extracted 1,728 definition candidates, of which 389 were true definitions, i.e., the precision is 0.2251 and the recall evaluated on the 150 definitions recall test set is 0.5867 (where the pattern-based approach takes the union of the basic “NP-nom is/are NP-nom” definition pattern, with all other patterns of Table 10, using verbs such as *define*, *determine*, *describe* with which we get better results).
- Different settings of the term-based approach can be tuned to achieve higher precision, higher recall or the best compromise between the two. In Table 18 below, we summarize a combination of two settings, i.e., the union of settings (R) and (w) (see (R & w) of Table 11). The union of these two settings is selected as a suitable compromise between the precision and recall,<sup>52</sup> where (R)

---

<sup>52</sup> For finding the best tradeoff we calculated the  $F_{0.5}$ -score for all the results of Table 23 and Table 24.

is selected from the experiments without the nominative case constraint (reported in Table 23) and (w) having a condition of at least one term in the nominative case (cf. Table 24).

- We consider setting (R & w) as well-suited to be used as the default setting, given that it achieves a suitable tradeoff between the precision and the number of extracted definitions.
- The sloWNet approach (see Table 12) has very low precision, i.e., 0.057 and extracts 270 definitions.

The three methods, the pattern-based, the term-based and the wordnet-based definition extraction, are first combined in the following straightforward way (see the results in Table 18).

- *Union* contains all the sentences that were extracted by at least one of the three methods, leading to nearly 650 extracted definitions with high recall of 0.7, but low precision (with approximately only each tenth sentence being a definition).
- *Intersection* contains the sentences that are extracted by at least two out of three methods; this results in higher precision of 0.26, with 129 extracted definitions. (We also tested the intersection of all the three methods, but even if the precision was above 0.40 the number of extracted definitions was too small to be considered.)

<b>Methods on the Slovene subcorpus: Summary and main combinations</b>	<b># Extracted sentences</b>	<b>Precision (# of definitions in extr. sentences)</b>	<b>Recall estimate (# of def. in 150 recall test set)</b>
Pattern-based (Total of Table 10)	1,728	0.2251 (389)	0.5867 (88)
Term-based (R & w of Table 11)	721	0.1747 (126)	0.0467 (7)
sloWNet-based (sloWNet of Table 12)	4,670	0.0570 (270)	0.2533 (38)
<b>Union</b>	6,606	0.0978 (646)	0.7000 (105)
<b>Intersection</b>	489	0.2638 (129)	0.1800 (27)

Table 18. Summary of definition extraction methods on the Slovene subcorpus. For the term-based approach a combined setting of two methods (R) of Table 23 and (w) of Table 24 is taken, the same as selected as the most appropriate term-based setting shown in Table 11. *Union* denotes the sentences extracted by any of the three methods, while *Intersection* denotes the sentences extracted by at least two out of three methods.

In addition to the above straightforward combinations of the three definition extraction methods we performed several other experiments, trying to achieve improved precision or recall (cf. Table 19 with the selection of the most useful and intuitive method combinations tested). As already mentioned, especially with the term-based approach, the users can choose the settings according to their desire of giving more importance to the precision or to the recall. As a basis we took the pattern-based approach, ensuring a good compromise between the precision and the number of extracted sentences, and combined it with other methods depending on whether we opted for better precision or recall.

---

Union of settings (R & w) is union of the settings with the highest  $F_{0.5}$ -score. For more details on  $F_{0.5}$  see footnote 40.

Methods on Slovene subcorpus: Other combinations	# Extract. sent.	Precision (# of definitions in extr. sentences)	Recall estimate (# of def. in 150 recall test set)
Pattern-based (Total of Table 10)	1,728	0.2251 (389)	0.5867 (88)
• <b>Enlarged with term best (ee)</b>	1,814	0.2255 (409)	0.5867 (88)
• <b>Enlarged with term (R &amp; w)</b> (biased to improved recall)	2,382	0.2040 (486)	0.6133 (92)
• <b>Filtered by term (R &amp; w) or WNet</b> (biased to improved precision)	336	0.3184 (107)	0.1733 (26)

Table 19. Summary of definition extraction methods on the Slovene subcorpus, presenting the results of a selection of combined methods.

- We can improve the precision and the number of extracted definitions (409) by complementing the pattern-based results with the sentences extracted by the term-based definition extraction using the setting with the best precision (setting ee).
- Alternatively, we can extract nearly 100 additional definitions if the candidate definitions extracted by the term-based (R & w) approach are added, while the precision still stays above 20%.
- On the other hand, if one opts for higher precision, the candidates extracted by the pattern-based approach can be filtered by the term-based or the slowNet definition candidates. In this case the precision can be nearly 10% higher than with the pattern-based approach, however, just slightly more than 100 definitions—instead of nearly 400 pattern-based definitions—are extracted.

To conclude, the combination of various definition extraction methods can be useful to improve the precision or the recall compared to using the methods individually. In summary, the union of three methods contains 646 definitions but the precision is under 10%, with precision over 30% precision just about 100 definitions (107) were extracted, and with 20% precision 486 definitions were extracted (for this setting, the  $F_1$  measure is 0.3062,  $F_{0.5}$  measure is 0.2354 and  $F_2$  measure is 0.4377).

### 5.3.2 Combining different approaches on the English subcorpus

Let us summarize the results of definition extraction methods on the English subcorpus (see Table 20).

- The pattern-based approach has three variants. With the patterns starting at the beginning of a sentence (*beg.-novar.*), we extracted 185 definitions with the precision of 0.3292. If we accept variations of the beginning of a sentence (*beg.-allvar.*), we extracted 200 definitions with a lower precision (0.2849). For substantially higher recall, the third variant of the pattern-based method (*novar.*, i.e., the patterns can occur anywhere in the sentence and not necessarily at the sentence beginning) can be used; in this way 273 definitions were extracted, however the precision is below 12%. We can see that compared to the Slovene pattern-based approach, the pattern-based approach for English can achieve a 10% higher precision.
- The term-based methods have lower precision than for Slovene (one of the reasons being that the nominative constraint is not applicable). In Table 20 we report the results for setting (m) of Table 16, which has a good compromise

between the precision and the number of extracted definitions.<sup>53</sup> The precision is 0.0824 and the estimated recall is 0.1333 with 90 extracted definitions.

- For WordNet-based definition extraction the precision is below 5%, while the estimated recall is 0.2667, meaning that (the same as for Slovene) this method is not useful if not applied in combination with other methods.

Straightforward combinations of the results of the three approaches are *Union* (including all the sentences that were extracted by at least one of the three methods) and *Intersection* (the union of definition candidates that were extracted by at least two out of three approaches). For the pattern-based approach we took variant 2 (in which patterns need to occur at the beginning of a sentence but different beginning variations are considered), while for the term-based approach we took the results of setting (m). However, none of the two combinations is appropriate for being used on its own since the pattern-based approach itself results in much higher precision. The only interesting observation is that with the union of the three methods, one can extract more than half of the definition test set sentences (good recall) and 344 definitions from the entire corpus.

Methods on English subcorpus: Summary and main combinations	# Extracted sentences	Precision (# of definitions in extr. sent.)	Recall estimate (# of def. in 150 recall test set)
Pattern-based			
1: beg.-novar. (Total of Table 13)	562	0.3292 (185)	0.2533 (38)
2: beg.-allvar. (Total of Table 15)	702	0.2849 (200)	0.2733 (41)
3: nobeg. (Total of Table 14)	2,283	0.1196 (273)	0.373 (56)
Term-based (setting m of Table 16)	1,092	0.0824 (90)	0.1333 (20)
WordNet-based (WNet of Table 17)	3,258	0.0430 (140)	0.2667 (40)
<b>Union</b>	4,727	0.0728 (344)	0.54 (81)
<b>Intersection</b>	318	0.2579 (82)	0.1333 (20)

Table 20. Summary of results of three definition extraction methods on the English subcorpus, as well as *Union* (i.e., sentences extracted by any of the methods) and *Intersection* (i.e., sentences extracted by at least two out of three methods).

Next, some other combinations of the three approaches were tested (see Table 21). As the basis we took the three variants of the pattern-based approach, as the pattern-based approach leads to the most satisfying results if used individually. Then we combined the three variants with other methods in different ways:

<sup>53</sup> We calculated the  $F_{0.5}$ -score for all the settings of Table 16. Setting (m) has the highest  $F_{0.5}$ -score at a 10% threshold. As explained in footnote 40, in the formula for calculating the  $F_{0.5}$ -score, the precision is the actual precision, while the recall is an average of two different recall estimates (one evaluated on the 150 definition recall test set—this recall estimation is also the one in all of the tables of the thesis—the other estimated recall is based on 1,000 randomly evaluated sentences out of which 20 were definitions, leading to the estimation of 860 definitions in the English subcorpus). We selected the best two settings in terms of the  $F_{0.5}$ -score, one for the 1% termhood value (setting j) and the other for the 10% termhood value (setting m). (Note that this is different than for Slovene where we took the best two settings: one with and the other without the nominative condition.) Since the selected setting (j) is included in the results of the selected setting (m), only the results of (m) are outlined in Table 17.

Methods on English subcorpus: Other combinations	# Extracted sentences	Precision (# of definitions in extr. sentences)	Recall estimate (# of def. in 150 recall test set)
Pattern-based 1 (beg.-novar.)	562	0.3292 (185)	0.2533 (38)
• <b>A: Filtered by term (m) or WNet</b> (biased to improving precision)	107	0.5420 (58)	0.0867 (13)
Pattern-based 2 (beg.-allvar.)	702	0.2849 (200)	0.2733 (41)
• <b>B: Enlarged with term best (w)</b>	778	0.2673 (208)	0.2867 (43)
Pattern-based 3 (nobeg.)	2,283	0.1196 (273)	0.3733 (56)
• <b>C: Filtered by term (m) or WNet</b>	390	0.2231 (87)	0.1400 (21)
• <b>D: Enlarged with term (m)</b> (biased to improving recall)	3,253	0.0987 (321)	0.4667 (70)
<b>Combined E (Union of B and C)</b>	1,022	0.2250 (230)	0.3333 (50)

Table 21. Summary of different combinations of definition extraction methods on the English subcorpus. As the basis three variants of the pattern-based approach are taken that are filtered or enlarged in different ways by term-based or WordNet-based definition extraction.

- *Combination A*, favoring precision, filters the sentences extracted by the most restrictive variant of the pattern-based approach (Pattern-based variant 1): it keeps only those pattern-based definition candidates that were also extracted by either the term-based approach (setting m) or by the WordNet-based definition extraction approach. This combination is a good choice if we strongly prefer high precision (precision is nearly 55%) over good coverage of the domain (lower recall).
- *Combination B* is a compromise between fairly good precision and recall. To the sentences extracted by the pattern-based approach (variant 2), it only adds sentences that were extracted by the best performing term-based setting (in terms of precision), i.e. setting (w). In this way we extracted 208 definitions with the precision above 28%.
- *Combination C* repeats the filtering of pattern-based extracted candidates as in combination A, except that the third variant of the pattern-based approach (variant 3—without the beginning of a sentence limitation—that has a lower precision but a higher recall) is taken as the basis. *Combination D* improves the recall by adding term-based extracted candidates (setting m) to the pattern-based approach (variant 3): 321 definitions at nearly 10% precision were extracted.
- *Combination E* has a good balance between the precision and the recall. It takes all the sentences extracted by the pattern-based definition extraction method in which different beginnings of the sentence are considered (variant 2), all the sentences with the best-performing (in terms of precision) term-based method as well as those sentences extracted with the less restrictive pattern-based method (variant 3) that were also extracted either by term-based (setting m) definition extraction or WordNet-based definition extraction. It extracted 230 definitions (more than the pattern-based method variant 2) and the precision is 22.5%. This setting can be proposed as a suggested setting if the user does not have other preferences for favoring precision or recall.

In summary, combining different definition extraction methods on the English subcorpus may lead either to improved precision or recall (extracting 58 definitions at

54% precision, 185 definitions at approx. 33% precision, 321 definitions at less than 10% precision but 46% recall and 230 definitions at 0.225 precision and 0.3333 recall). (For this setting, the  $F_1$  measure is 0.2686,  $F_{0.5}$  measure is 0.2406 and  $F_2$  measure is 0.304). In comparison with extracting definitions from the Slovene subcorpus, we can see that with 10% or 20% precision half less definitions were extracted in English. However, with above 30% precision more definitions are extracted in English and in English definition extraction settings can be set also to achieve precision above 50%.

### 5.3.3 Subjectivity of evaluation results

Relatively low results in terms of precision and recall can be partly attributed to the nature of our corpus and the subjectivity of the quantitative evaluation. The complexity of the task has been illustrated by citing numerous complicated sentences extracted from the corpus. Given that our corpus mainly consists of academic articles, a high level of prior knowledge is presupposed: basic terms are considered common knowledge and additional information is provided through references to related work, not always through definitions and explanations. Moreover the definitions are encoded in linguistically intricate structures. Ideally, the corpus would be extended by more popular science articles and textbooks, which prove to contain less complex sentences.

When analyzing the definition candidates, we noted that in many cases, the decision whether a sentence should be considered a definition or not is not obvious. This means that the quantitative results are highly dependent on the annotator. We therefore performed a quick inter-annotator agreement experiment. Fifteen sentences were evaluated by 21 annotators. We calculated *kappa* statistics, i.e., a chance-adjusted measure of agreement. Randolph's free-marginal multi-rater *kappa* (see Randolph, 2005; Warrens, 2010) was used, implemented in Randolph's (2008) *Online Kappa Calculator*. Values of *kappa* can range from -1.0 to 1.0, where -1.0 indicates perfect disagreement below chance, 0.0 indicates agreement equal to chance, and 1.0 denotes perfect agreement above chance. In our experiment, the inter-annotator agreement *kappa* value is 0.36, which is far from 0.70 which usually shows adequate inter-rater agreement. We provide two examples, Example (clxxxviii) is the one where all the evaluators agreed that the sentence is a definition, while for Example (clxxxix), 10 evaluators tagged it as definition and 11 as non-definition.

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

*[The Lombard effect is a phenomenon, where a speaker increases the intensity of the speech when the level of background noise raises.]*

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

*[Google Translate is a typical representative of statistical machine translation systems (Statistical Machine Translation-SMT), presented in Section 3.]*

Another experiment by which we relativize the value of quantitative results is when instead of taking the binary *definition/non-definition value*, we evaluated the definition candidates with scores from 1 (non-definition) to 5 (a perfect definition). For instance, when classifying the candidates only with *definition/non-definition* labels, out of 1,022 English candidate sentences (setting E of Table 21), 230 sentences were evaluated as

definitions. In contrary, when evaluating definitions on the 1–5 scale, 345 out of 1,022 sentences were attributed a positive score (in the range between 2 and 5). This shows that for English with less restrictive evaluation the precision could immediately be reported 0.3376 instead of 0.225 as reported in Table 21.

### 5.3.4 Analysis of different types of definition candidates

In this section we systematically analyze all the definitions of the two settings that we chose as final settings for English and Slovene. For Slovene this was setting B (biased to improved recall, cf. Table 19) and for English setting E (cf. Table 21). For Slovene we consider 486 definitions that were extracted by 0.2040 precision and 0.6133 recall, while for English 230 definitions—extracted with 0.2250 precision and 0.3333 recall—were analyzed. The two settings were selected based on the criterion of the largest number of definitions extracted at a relatively high precision, leading to 716 analyzed definitions out of 3,404 sentences analyzed for both languages in total.

Already in the previous chapters of this thesis, we have discussed different interesting phenomena related to the extraction of definitions from running text. Here we provide different tags for marking different types of extracted candidate definitions and discuss various problems with their classification into definition/non-definition categories. In total, approx. 19,300 Slovene and 14,000 English sentences were evaluated as definitions or non-definitions (out of which more than 1,500 sentences (approx. 1,050 for Slovene and more than 500 for English) were classified as definitions). However, the 3,404 annotated sentences—which we analyze in this section—are tagged more systematically, double-checked and reclassified into the definition/non-definition category after a thorough analysis.

In Table 22 we list the tags for marking different types of sentences and discuss different problems with their classification into definition/non-definition categories. The basic tags are “Y” denoting *yes* for definitions and “N” for non-definitions, but several other categories were used for (non-)definition labeling. Different letters denoting specificities of definition candidates can also be combined. If the sentence has an interrogation mark beside the tag, it means that the sentence is a borderline case. In addition to different tags, explanations and examples illustrating each category are provided.

This analysis mainly summarizes the observations discussed in previous chapters, reflected in the tags grouped in the following way.

- *Definition form* related tags mainly distinguish between main definition types, i.e. *genus* and *differentia* or paraphrases, extensional definitions and definitions with no hypernym (e.g., functional definitions) or with hypernym only.
- *Definition content* tags discuss different types of problems as encountered in the analysis of extracted sentences (when the content is too general, too specific, etc.).
- *Definiendum related* tags refer to the fact that many definienda are proper names and abbreviations or that the definitions are provided for terms out of the domain.
- Next category, *segmentation related* tags treat issues such as wrong segmentation, multiple definitions in the same sentence or spanning over a few sentences and embedded definitions.

- The last category, *annotation related tags*, points to examples, where either sentences with the same content are extracted from the corpus twice (and could be omitted) or sentences for which the annotation was changed while double-checking the tags.

Note that the examples frequently contain several tags since the categories are overlapping.

<b>DEFINITION TYPE RELATED TAGS</b>	
<b>Y/N</b>	<p><b>Definition/Non-definition</b></p> <p>The two basic categories denote that a sentence is classified as a definition ("Y") or non-definition ("N"). If no other tag labels are added, the sentences labeled by "Y" denote the most obvious definitions (mainly using the <i>genus</i> and <i>differentia</i> structure or defining by paraphrases). Sentences that are borderline definitions are marked as "Y?", while the "N?", tag is used for borderline sentences containing knowledge-rich contexts but not being definitions.</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Y:</b> <i>Text classification is the task of assigning a text document to one or more categories, based on its contents.</i> ["Y"]</li> </ul> <p>Example non-definition:</p> <ul style="list-style-type: none"> <li>- <b>N:</b> <i>As such, language is a central theme of our research activities.</i> ["N"]</li> </ul>
<b>Ye/Ne</b>	<p><b>Extensional</b></p> <p>Sentences that define a term by providing its constituting parts, all possible realizations or representative examples are tagged as "Ye". Non-definitions in this category ("Ne?") usually denote sentences, in which there are some examples of the class provided, but the examples are not representative enough for the term to be clearly defined (frequently borderline cases).</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Ye:</b> <i>Language resources are corpora and other lexical data: electronic versions of dictionaries for human users and lexicons for language technology applications.</i> ["Ye?"]</li> <li>- <b>Ne:</b> <i>ZV. The following words are adverbs: kam 'where to', kje, kod 'where', kako 'how', kolikokrat 'how many times', kdaj 'when', zakaj 'why', koliko 'how much; how many', doklej 'till when'.</i> ["Ne?"]</li> </ul>
<b>Yn/Nn</b>	<p><b>No hypernym</b></p> <p>Definitions that explore other possibilities than defining the term by its hypernym and the <i>differentia</i> or paraphrases, but are not extensional definitions. The most common types are functional or typifying definitions. For non-definitions, especially borderline cases marked by "Nn?" are interesting.</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Yn:</b> <i>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.</i> ["Yvny"]</li> <li>- <b>Nn:</b> <i>The answer is: 'correctness' is defined by what the annotation scheme allows or disallows — and this is an added reason why the annotation scheme has to be specific in detail, and has to correspond as closely as possible with linguistic realities recognized as such.</i> ["Nsn?"]</li> </ul>
<b>Yp/Np</b>	<p><b>Incomplete (only hypernym)</b></p> <p>Sentences (classified as definitions or non-definitions) that are usually borderline cases, since they do not sufficiently define the definiendum, but provide only its</p>

	<p>hypernym (without providing enough specification elements).</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Yp:</b> <i>Google Similarity Distance (GSD) is a word/phrase semantic similarity distance metric developed by Rudi Cilibrasi and Paul Vitanyi proposed in (Cilibrasi &amp; Vitanyi, 2007).</i> ["Yp"]</li> <li>- <b>Np:</b> <i>As a consequence, lemmatization is an indispensable preprocessing step for most language processing methods including term extraction.</i> ["Npv"]</li> </ul>
<b>Yy/Ny</b>	<p><b>Relational (synonym, antonym or sibling concept)</b></p> <p>Sentences that instead of hypernym and <i>differentia</i> use other concepts such as synonyms, antonyms or sibling concepts (possibly followed by <i>differentia</i>); frequently knowledge-rich context sentences evaluated as non-definitions are in this category ("Ny?").</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Yy:</b> <i>Text mining (Feldman and Sanger, 2007) is a variant of data mining in which models and patterns are extracted from unstructured natural language text.</i> ["Yy"]</li> <li>- <b>Ny:</b> <i>Other terms used for their denomination are: thematic roles, semantic cases, thematic relations, semantic arguments, etc.</i> ["Nay"]</li> </ul>
<b>DEFINITION CONTENT RELATED TAGS</b>	
<b>Yg/Ng</b>	<p><b>Too general</b></p> <p>Often borderline cases (denoted by "?"), where the way of defining a term is too general (without providing enough specific elements), but all in all either still providing enough information for considering a term well defined ("Yg") or not ("Ng"). In <i>genus et differentia</i> definition types, the tag can also denote that the <i>genus</i> is too general, but the definition can be a very informative lexical definition and not a borderline case.</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Yg:</b> <i>A translation memory system is a tool that is designed for helping human translators during translation.</i> ["Yg"]</li> <li>- <b>Ng:</b> <i>Homogeneity is a useful practical notion in corpus building, but since it is superficially like a bundle of internal criteria we must tread very carefully to avoid the danger of vicious circles.</i> ["Ng"]</li> </ul>
<b>Ys/Ns</b>	<p><b>Too specific</b></p> <p>Sentences that are in the majority of cases borderline cases (marked by "?" after the definition/non-definition tag) because they are considered to be too specific to be significant for a general definition of the definiendum; in the example below, the context of WordNet is not mentioned and therefore the sentence is somewhat too specific (and human knowledge is needed to add a valuable context for this definition).</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Ys:</b> <i>A synset is a group of data elements (synonyms) that are considered semantically equivalent [1].</i> ["Ys"]</li> <li>- <b>Ns:</b> <i>Recall is the extent to which all correct annotations are found in the output of the tagger.</i> ["Ns?"]</li> </ul>
<b>Yc/Nc</b>	<p><b>Cyclic</b></p> <p>Definition candidates that are borderline cases since they are cyclic, e.g., using in the definiens part the definiendum itself or a word with the same etymology.</p> <p>Example:</p> <ul style="list-style-type: none"> <li>- <b>Yc:</b> <i>The Naive Bayes (NB) classifier is a probabilistic classifier based on</i></li> </ul>

	<i>Bayes' theorem.</i> ["Ylgc?"]
<b>Yo/No</b>	<p><b>Outdated</b></p> <p>In some cases the definiendum itself or the information provided in a definition is completely or partly outdated, being not anymore valid or relevant. (Often some projects, organizations, etc., are not active anymore and a completely valid definition should at least specify the dates of its activity).</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Yo:</b> <i>The TELRI Association is the independent pan-European voice of the multilingual research and development community, a devoted though impartial partner of the European language industry, and a respected consultant of the European Commission for the Multilingual Information Society.</i> ["Ykogz?"]</li> <li>- <b>No:</b> <i>There is a movement being spearheaded by a special interest group of the Localization Industry Standards Association (LISA) known as OSCAR (Open Standards for Container/Content Allowing Re-use)</i> ["Nkwog?"].</li> </ul>
<b>Yz/Nz</b>	<p><b>Metaphorical</b></p> <p>Sentences in which the hypernym or the sentence itself is used slightly metaphorically.</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Yz:</b> <i>The Basic Multilingual Plane (BMP) is the official name of the "heart and soul of Unicode" (Gillam 2003), which contains the majority of the encoded characters from most of the modern writing systems (with the exception of the Han ideographs used in Chinese , Japanese and Korea ).</i> ["Yzy?3"]</li> <li>- <b>Nz:</b> <i>The World Wide Web is a marvelous place, with a vast range of languages, content domains and media formats.</i> ["Nz?"]</li> </ul>
<b>Yf/Nf</b>	<p><b>Including formulas</b></p> <p>When there is a (part of) mathematical formula in the extracted sentence, we added a special tag, since these sentences should better be discarded because of their noisy nature (errors are due specially to the segmentation). On the small corpus the formulas were in the majority discarded in manual preprocessing, but not in the entire corpus. However, in few cases formulas are mainly explained in textual form and if there are only several symbols missing, we treat them as borderline definitions, since only minimal manual refinement is needed (see the underlined part in the first example below).</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Yf:</b> <i>The precision is defined as the ratio number of correctly classified instances of class c number of instances classified as class c and the recall is defined as the ratio number of correctly classified instances of class c number of instances of class c <u>The trade-off between precision and recall is measured by the value of F1-measure defined as <math>2 * \frac{precision * recall}{precision + recall}</math>.</u></i> ["Yfwm?"]</li> <li>- <b>Nf:</b> <i>A useful example is Riley's entropy semiring : <math>K=R_0 \times R_0</math> hp , <math>hi_ = h1/( 1 - p )</math> , <math>h/( 1 - p )</math> <math>2i</math> hp , <math>hi_{hp0}</math> , <math>h0i = hp + p0</math> , <math>h + h0i - 0 = h0</math> , <math>0i</math> hp , <math>hi_{hp0}</math> , <math>h0i = hp0</math> p , <math>p0</math> h + p <math>h0i - 1 = h1</math> , <math>0i</math> 184 where hp , hi is undefined for p_1.</i> ["Nf"]</li> </ul>
<b>DEFINIENDUM RELATED TAGS</b>	
<b>YI/NI</b>	<p><b>Proper names</b></p> <p>The definitions of proper name definienda (used for named entities) have a</p>

	<p>special tag, since there is no unanimous agreement whether named entities are domain terms that should figure in terminological resources or not. In our case, when a named entity was treated as a term and its definition was provided (tagged as "YI"), we included it in the final glossary. "NI" in contrast indicates that either a named entity is not a relevant domain term—and therefore its definition is not relevant—or that a relevant domain term that is a named entity is not well defined.</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>YI:</b> <i>An initiative of general interest is European Language Resources Association – ELRA [8], – an infrastructure for identifying, collecting, classifying, validating, distributing, and exploiting language and speech resources, such as basic data (corpora, recordings, terminology), linguistic models (grammars, lexica, HMM) and software tools. ["YIk"]</i></li> <li>- <b>NI:</b> <i>The Russian National Corpus (<a href="http://ruscorpora.ru">http://ruscorpora.ru</a>) is an ongoing project that began in 2000. ["NIg"]</i></li> </ul>
<b>Yk/Nk</b>	<p><b>Abbreviations</b></p> <p>A special case of named entities are abbreviations, for which it is even more delicate whether we consider them terms or not. If considered as terms ("Yk") the abbreviation should be explained in the scope of the same terminological resource, etc. in order to make the definition valid. When the problem is not only to expand the abbreviation, but that the concept denoted by the abbreviation is not even well defined then the sentences are marked with "Nk". In most cases these definitions are borderline cases.</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Yk:</b> <i>LMF (ISO 24613) is a model that provides a common standardized framework for the construction of NLP lexicons. ["Yk?"]</i></li> <li>- <b>Nk:</b> <i>The DWDS is a venture which can be developed, expanded and detailed in multiple ways, but one with a practical and academic benefit right from the outset. ["Nkt"]</i></li> </ul>
<b>Yt/Nt</b>	<p><b>Borderline term or not a term</b></p> <p>Sentences that have a definition structure, but in which at the place of definiendum we have a borderline term (mainly borderline cases), are marked with "Yt", while if sentences look like definitions but do not contain a real definiendum (e.g., it is not a domain term) the tag is "Nt".</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Yt:</b> <i>The Cluetrain Manifesto is a " movement" which examines the phenomenon that is the Internet and the substantial changes that we must all implement in our businesses to be successful in the global village. ["Yltd?"]</i></li> <li>- <b>Nt:</b> <i>'Second position' is usually defined as the position after the first syntactic constituent or the first prosodic word. ["Nts"]</i></li> </ul>
<b>Yd/Nd</b>	<p><b>Out-of-domain</b></p> <p>"Yd" denotes definitions that are not fully domain specific, but still relevant for the domain. On the other hand if sentences, even though possibly being definitions, are completely irrelevant for the domain, the tag is "Nd".</p> <p>Ex.:</p> <ul style="list-style-type: none"> <li>- <b>Yd:</b> <i>Serbian language is an Indo-European, South-Slavic language, with 10 million speakers in Serbia (11 million world-wide) (Grimes, 1996). ["Yd?"]</i></li> <li>- <b>Nd:</b> <i>Jakarta is a tropical, humid city, with annual temperatures ranging between the extremes of 75 and 93 degree F (24 and 34 degree C) and a</i></li> </ul>

	<i>relative humidity between 75 and 85 percent.</i> ["Nlds"]
<b>SEGMENTATION RELATED TAGS</b>	
<b>Yw/Nw</b>	<p><b>Wrong segmentation</b></p> <p>"Yw" is used for definitions, for which minimal manual refinement should be performed, since the sentence is wrongly separated from the rest of the text (usually due to wrong title/first-sentence segmentation or unrecognized abbreviations). Also non-definitions can be wrongly segmented ("Nw") but since non-definitions are not relevant for the discussion, it is more the category of borderline cases that is relevant; in the second given example, the sentence is not finished and therefore does not provide a definition.</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Yw:</b> <i>1 INTRODUCTION Lemmatization is the process of determining the canonical form of a word, called lemma, from its inflectional variants.</i> ["Yw"]</li> <li>- <b>Nw:</b> <i>Word sense disambiguation is the problem of assigning which of several possible meanings of a word a certain</i> ["Nwa?"]</li> </ul>
<b>Yv/Nv</b>	<p><b>Embedded</b></p> <p>If a definition is embedded in a sentence that is as a whole not a definition the tag is "Yv" (in the first example below we underline the embedded definition), sometimes it can be a single additional word added to a definitions, while in other cases the definition can be provided in parentheses of a longer sentence. Sentences that have an embedded knowledge-rich context but are not definitions (or contain too much irrelevant information) are marked by "Nv" (also often followed by "?").</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Yv:</b> <i>For POS tagging, the first thing to list is <u>the tagset—i.e., the list of symbols used for representing different POS categories.</u></i> ["Yv?"]</li> <li>- <b>Nv:</b> <i>The next larger level at which errors are tallied in speech recognition is the <u>sentence or utterance, i. e., a sequence of words.</u></i> ["Nvg?"]</li> </ul>
<b>Ym/Nm</b>	<p><b>Defining several terms</b></p> <p>If in a sentence several concepts are defined, the tag is "Ym" (indicating also that the real number of extracted definition is higher than the one stated). In the example below we underline the two definienda. Similar cases, where definition candidates are providing a knowledge-rich context but not being tagged as definitions, are annotated as "Nm".</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Ym:</b> <i>Computational linguistics has theoretical and applied components, where <u>theoretical computational linguistics</u> takes up issues in theoretical linguistics and cognitive science, and <u>applied computational linguistics</u> focuses on the practical outcome of modelling human language use.</i> ["Ymn?"]</li> <li>- <b>Nm:</b> <i>Izraz <u>termin</u> se nanaša na individualno enoto, po drugi strani pa se poimenovanje <u>terminologija</u> nanaša na kolektivni objekt (Kageura, 2002).</i> ["Nmg?"]</li> </ul> <p>Translation: <i>Expression term refers to the individual unit, while on the other hand the naming terminology refers to the collective object (Kageura, 2002).</i> ["Nmg?"]</p>
<b>Ya/Na</b>	<p><b>Spanning across two sentences</b></p> <p>Sometimes definitions span across two sentences. If the evaluated sentence</p>

	<p>provides satisfactory definition context but the sentence suggests that in a sentence before/after useful defining context could be found to provide a more complete definition the sentence was tagged as "Ya". If the isolated sentence is not enough for defining a term, but its structure suggests that a complete definition could be found in the sentences before/after tag "Na" is used.</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Ya:</b> <i>An example of such data collection is the WordNet: a lexical database for the English language (WordNet, 2002; Lexical FreeNet, 2002). ["Ypal?"]</i></li> <li>- <b>Na:</b> <i>Other terms used for their denomination are: thematic roles, semantic cases, thematic relations, semantic arguments, etc). ["Nay?"]</i></li> </ul>
<b>ANNOTATION RELATED TAGS</b>	
<b>Yb/Nb</b>	<p><b>Doubles</b></p> <p>Definitions/non-definitions that appear more than once in the text: sometimes in longer articles the author repeats a sentence more than once; in other cases it is the same sentence that appears in different articles written by the same author.</p> <p>Example:</p> <ul style="list-style-type: none"> <li>- <b>Yb:</b> <i>Word sketches are one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour. ["Yb"]</i></li> </ul>
<b>Yr/Nr</b>	<p><b>Revise annotation</b></p> <p>When checking the evaluations, in several examples we believe that the evaluation into Y/N should be revised. In the majority of cases borderline candidates are concerned, where the limit between definition and non-definition is fuzzy. In other examples it was simply a mistake made during the evaluation. "Yr" therefore means that after reconsideration we evaluated the sentence as non-definition (but initially tagged as definition) and "Nr" that a sentence tagged as non-definition should better be considered a definition. These sentences should preferably be re-evaluated by another annotator.</p> <p>Examples:</p> <ul style="list-style-type: none"> <li>- <b>Yb:</b> <i>Basque is a free-constituent order language where PPs in a multiple-verb sentence can be attached to any of the verbs. ["Yrpsd"]</i></li> <li>- <b>Nb:</b> <i>In single-link clustering, distance between two clusters is the distance between the nearest neighbors in those clusters. ["Nnr?"]</i></li> </ul>

Table 22. Tags and examples of different types of definition candidates.

In summary, we have shown that different combinations of methods lead to higher precision or recall depending on the user's preferences. For Slovene, the setting with a good precision-recall compromise reaches approx. 20% precision and 61% recall, which is a comparable result with more complex machine learning system for Polish (reported by Degórski et al., 2008b) and better than grammar-based systems for Czech and Bulgarian (Przeziórkowski et al., 2007).

The corpus used in this thesis contains complex sentences for which even for human evaluators the decision on tagging them as definition/non-definition is not easy (0.36 kappa). Moreover, there are relatively few definitions in the corpus, which can be seen from a small experiment in which we evaluated 1,000 randomly selected sentences, out of which only 17 were definitions. We repeated the experiment three times, once on the English subcorpus and twice on the Slovene subcorpus, and evaluated positively 17, 20 and 24 sentences. This shows that the improved precision is relatively high and of high importance for the user (from approx. 0.02 to more than 0.2 depending on the selected

settings). For English, several authors report on better (perfectly) performing systems, but as shown in Section 5.3.3, the quantitative evaluation results are rather subjective, as a 10% higher precision can be reported by loosening the criteria of what is a definition. For instance, Reiplinger et al. (2012) model the ACL domain, which is the same type of highly specialized corpus. They propose a five scale point scoring system and if only sentences providing precise and concise descriptions of the concept are considered (score 5), their method has between 10% and 15% precision for the definitions of a selected set of 20 terms, while if upper three levels are considered the precision is about 60%, which is comparable to our results if biased to improved precision.

We believe that the qualitative analysis of Section 5.3.4—complementing the quantitative results—is of major importance for increased understanding of the domain, as well as of the complexity of the definition extraction task.

In further work, we plan to include the recently developed tagger and lemmatizer for Slovene (Grčar et al., 2012) and Tree Tagger (Schmid, 1994) for English and check how they affect the results, especially the pattern-based definition extraction. The integration of a chunker—for Slovene, we could use the information from the recently developed dependency parser (Dobrovoljc et al., 2012)—would help in the noun phrase detection, useful for all the three methods. Another research direction will be that, instead of searching for all definitions sentences as we do now, we limit the search only to definitions of a list of previously selected terms.



## 6 Workflow implementation in ClowdFlows

This section describes the details of the online NLP workflow for definition extraction from Slovene and English text corpora, implemented in the ClowdFlows workflow construction and execution environment (Kranjc et al., 2012). The ClowdFlows environment has already been presented in the related work section (Section 4.3.4). In this section we describe the constituent parts of the workflow. Since we use Patterns (Pa), Terms (Te) and Wordnet (W) it is called the *PaTeW workflow*.

The implementation of the definition extraction methodology into the workflow was conceived together with the co-authors of papers Pollak et al. (2012a, 2012c), who had a major role in the actual incorporation of the definition extraction modules into the workflow execution engine. While an early implementation of the definition extraction workflow is described in Pollak et al. (2012a), Pollak et al. (2012c) focuses on the implementation of the ToTrTaLe annotation workflow.

A workflow in ClowdFlows is a set of widgets and connections. A widget is a single workflow processing unit with inputs, outputs and parameters. Connections are used to transfer data between two widgets and may exist only between an output of a widget and an input of another widget. Parameters are set manually by the user. In Figure 12 the definition extraction workflow is shown. Besides the main terminology and definition extraction widgets, several other new auxiliary text processing and file manipulation widgets were developed and incorporated to enable seamless workflow execution.

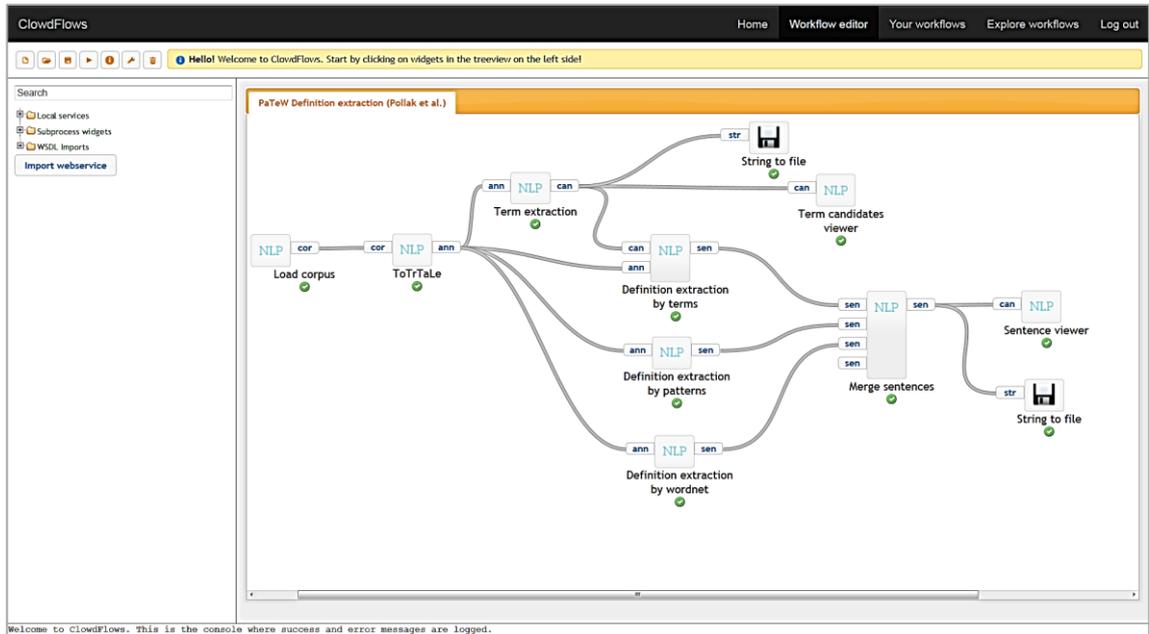


Figure 12: Definition extraction workflow in ClowdFlows, available online at <http://clowdflows.org/workflow/1380>

The workflow contains different types of widgets: preprocessing steps are covered by the *Load corpus* widget and the *ToTrTaLe* widget, the term extraction is implemented in the *Term extraction* widget and the main definition extraction step is implemented by three definition extraction widgets (*Definition Extraction By Patterns*, *Definition Extraction By Terms*, *Definition Extraction By WNet*) followed by the *Merge sentences* widget used for combining the results. The output visualization is provided by the *Term candidates viewer* and the *Sentence viewer* widgets.

The ontology construction phase described in Section 3.3 is currently not seen as part of the definition extraction methodology, but could in the future, if made accessible as a web service, be incorporated in the process and serve in the corpus inspection phase as well as in the final glossary construction phase.

While the term and definition extraction algorithms were implemented in Perl, the web services were implemented in the Python programming language (additionally, some freeware software packages were used, as explained in Section 6.1). The services are currently adapted to run on Unix-like operation systems, but are easily transferable to other operation systems.

The two main contributions of this workflow implementation are the *ToTrTaLe* web service<sup>54</sup> and the *Definition extraction* web service<sup>55</sup>. The *ToTrTaLe* web service has two main functionalities: converting different input files to plain text format (see the *Load corpus* widget described in Section 6.1), while the second one—the *ToTrTaLe* widget presented in Section 6.2—uses the (already existing) *ToTrTaLe* morphosyntactic tagger and lemmatizer to annotate the corpus. The two functionalities correspond to two operations described in one WSDL (Web Service Description Language) file.

On the other hand, the *Definition extraction* web service is available through three different widgets, one for each definition extraction technique (pattern-, term- and wordnet-based definition extraction).

CloudFlows can automatically construct widgets for web services, where each operation maps into one widget (and one web service can have several operations). They identify the inputs and the outputs of the web service's operations from the WSDL description. In addition to implementing the web services mentioned above, additional functionalities were required to adequately support the user in using these web services and some additional platform specific widgets were implemented accordingly. These widgets, not exposed as web services, are run on the server hosting the CloudFlows application.

In the following subsections we present the widgets that constitute the definition extraction workflow.

## 6.1 Load corpus widget

The load corpus widget allows the user to conveniently upload his corpus in various formats, either as a single file or as several files compressed in a single ZIP file. The supported formats are PDF, DOC, DOCX, TXT and HTML, the latter being passed to the service in the form of an URL as a document. Before being transferred, the actual files are encoded in the Base64 representation, since some files might be binary files. So

---

<sup>54</sup> This work was realized together with co-authors (Pollak et al., 2012c). The implementation was mainly done by Nejc Trdin.

<sup>55</sup> The implementation was done in collaboration with Anže Vavpetič (cf. Pollak et al., 2012a).

the first step is to decode the Base64 representation of the document. Based on the file extension, the program chooses the correct converter:

- If the file extension is HTML, we assume that an URL address is passed and that it is written in the document variable. It is also assumed that the document contains only plain text. The web service then downloads the document via the given URL in plain text.
- DOCX Microsoft Word documents are essentially compressed ZIP files containing the parts of the document in XML. The content of the file is first unzipped, and then all the plain text is extracted.
- DOC Microsoft Word files are converted using an external tool, `wvText` (Lachowicz and McNamara, 2006), which transforms the file into plain text. The tool is needed because the whole file is a compiled binary file and it is hard to manually extract the contents without appropriate tools.
- PDF files are converted with the Python `pdfminer` library (Shinyama, 2010). The library is a very good implementation for reading PDF files, with which one can extract the text, images, tables, etc., from a PDF file.
- If the file name ends with TXT, then the file is assumed to be already in plain UTF-8 text format. The file is only read and sent to the output.
- ZIP files are extracted into a flat directory and converted appropriately—as above—based on the file extension. Note that ZIP files inside ZIP files are not permitted.

The resulting text representation is then sent through several regular expression filters, in order to further normalize the text. For instance, white space characters are merged into one character.

The final step involves sending the data. But before that, the files have their unique identifiers added to the beginning of the single plain text file. The following steps leave these identifiers untouched, so the analysis can be traced through the whole workflow. At each step of the web service process, errors are accumulated in the error output variable.

## 6.2 ToTrTaLe widget

The second operation of the ToTrTaLe web service, available through the ToTrTaLe widget, uses the ToTrTaLe text processing tool (Erjavec, 2011) that was in detail described in Section 4.3.1, to annotate the input texts. On the input text tokenization, part-of-speech tagging and lemmatization are performed. The output of these three steps is a string of text tokens, where each word token is annotated with its context disambiguated part-of-speech tag and the base form of the word, i.e., lemma, thus abstracting away from the variability of word-forms. For the ToTrTaLe annotation web service, the mandatory parameters are: the document in plain text format and the language of the text. Additionally, the input parameter for post-processing defines if the post-processing scripts are run on the text. Before implementing it as a web service, ToTrTaLe was available online only as a web application for small parts of files in raw text format only.

The post-processing scripts are Perl implementations of corrections for some tagging mistakes described in Section 4.4.1. In the current post-processing implementation we

added a list of previously unrecognized abbreviations (such as *et al.*, *in sod.*, *cca.*) to avoid incorrect redundant splitting of the sentence. We corrected the wrongly merged sentences by splitting them into two different sentences if certain abbreviations (such as *etc.*) are followed by an upper-case letter in the word following the abbreviation. Other post-processing corrections include the correction of adjective-noun agreement, where we assume that the noun has the correct tag and the preceding adjective takes its properties. Some other individual mistakes are treated in the post-processing script, but not all the mistakes have been addressed. Even if the majority of the described mistakes are currently handled in this optional post-processing step, it should be taken into consideration in future versions of ToTrTaLe, by improving tokenization rules or changing the tokenizer, re-training the tagger with larger and better corpora and lexica, and improving the lemmatization models or learner. We did not perform a proper evaluation of the influence of the post-processing step on the definition extraction, but we can say that already on the Slovene part of the LTC proceedings corpus, there were approximately 350 substitutions only on wrongly segmented sentences due to the *et al.* abbreviation, which is the one used in a big majority of the defining sentences.

Since the web service is useful also on its own not just as part of the definition extraction workflow, a separate ToTrTaLe workflow is also proposed and available at <http://clowdflows.org/workflow/228/> (cf. Figure 13). The accepted languages are English, Slovene and even historical Slovene. If used as a separate preprocessing step, the user can also select whether the output should be in the XML format (default) or in the plain text format. An example of XML output file is given in Table 4.

The data and the processing request are sent by ClowdFlows to the web service ToTrTaLe annotation operation, which is run on a remote server. The output is written into the output variable, and the possible errors are passed to the error variable. The output string variable and the accumulated errors are passed on to the output of the web service, which is then sent back to the client.

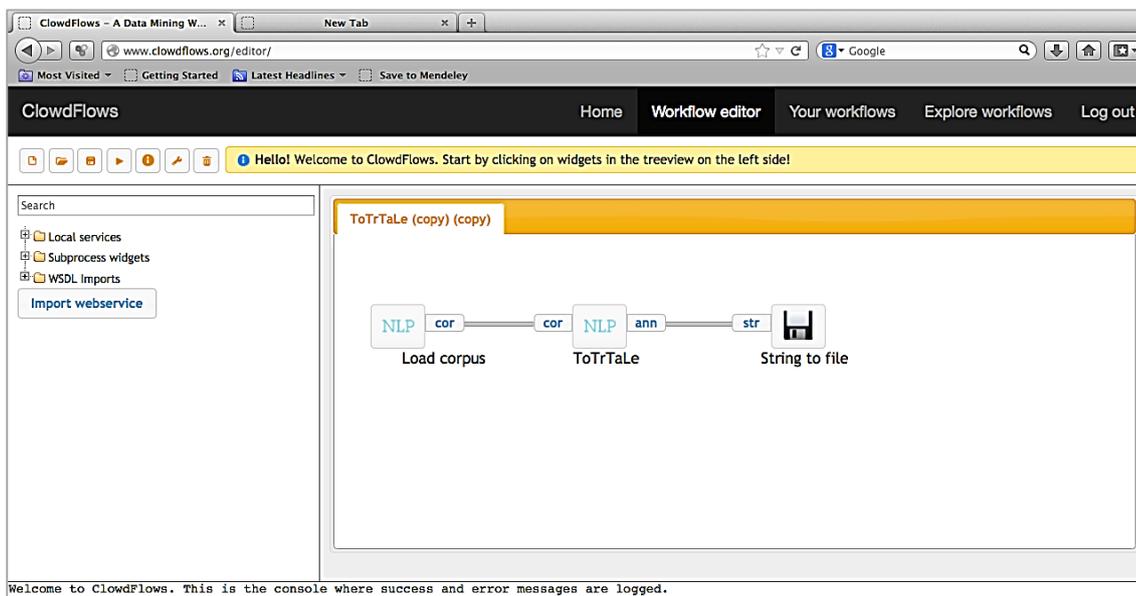


Figure 13. A screenshot of the ToTrTaLe workflow in ClowdFlows, available online at <http://clowdflows.org/workflow/228/>.

## 6.3 LUIZ widget

The term extraction widget implements monolingual term extraction (for Slovene and English) of the LUIZ term recognition tool (Vintar, 2010). LUIZ is in detail presented in the background technologies Section 4.3.2. The term extraction consists of two steps: extracting the noun phrase candidates based on morphosyntactic patterns, followed by weighting and ranking of the candidates based on their ‘termhood’ value, for single word and multi-word terms separately. Before our implementation, the LUIZ term extractor was available online only as a demo for Slovene terminology extraction.

We implemented LUIZ as a web service—more precisely as one of the operations of the definition extraction web service—available as a workflow widget. It is composed of lexicon extraction, keyness ranking (lemmas ranked by relative frequency compared to the reference corpora), candidate noun phrase extraction, and terminological candidates ranking.

Compared to the original system, in our implementation several details were changed, such as filtering the URL tags and punctuation marks, uncapitalizing lemmas from FIDA, and perhaps the most important part, instead of two lists, we propose the unique list of ranked terms, where the termhood values of single word terms and multi-word terminological expressions are ranked on a common scale. First a separate list is made for one-word terms and for multi-word terms, which is then normalized in a way that the top ranked terms of each list have value 1 and the others are proportionally distributed between 0 and 1.

The top term candidates—single- and multi-word terms in common list—extracted from the Slovene part of our Language Technologies Corpus in Table 5 and Table 7.

This term extraction web service operation can be used either separately—if we aim only at extracting the terms from a domain corpus—or as a necessary step for the second definition extraction method, implemented by the term-based definition extraction widget (cf. Section 6.4.2). The list of extracted terms can also be proposed for manual inspection by the user.

## 6.4 Definition extraction widgets

### 6.4.1 Pattern-based definition extraction widget

Pattern-based definition extraction is the first of the three definition-extraction operations of the developed web service. The pattern-based definition extractor seeks for sentences corresponding to predefined lexico-syntactic patterns. The user can upload his own pattern list or use the lists (one for each language) proposed in this thesis. The methodology is presented in detail in Sections 5.1.1 and 5.2.1. Patterns are composed of word forms or lemmas, part-of-speech information as well as more detailed morphosyntactic descriptions, such as case information for Slovene nouns, person and tense information for verbs, etc. The basic pattern for Slovene is for instance “NP-nom Va-r3[psd]-n NP-nom” where “NP-nom” denotes a noun phrase in the nominative case and the “Va-r3[psd]-n” matches the auxiliary verb in the present tense of the third person singular, dual or plural and the form is not negative, in other words it corresponds to *je/sta/so* [*is/are*] forms of the verb *biti* [*be*]. As—at the moment of these experiments—there is no chunker available for Slovene, the basic part-of-speech annotation provided

by ToTrTaLe was needed for determining the possible noun phrase structures and the positions of their head nouns. In further versions of the system a chunker output could be used at least for English.

In the English version, the cases are not expressed in the same manner and therefore the nominative case used in the majority of Slovene patterns cannot be applied. For that reason, the patterns for English are looser and less precise. Therefore an optional *beginning of a sentence* parameter can be applied, restricting the number of proposed candidates. The parameter can have three values *nobeg.* meaning that a pattern can be found anywhere in a sentence, *beg.-novar.* denotes the most restrictive setting in which a pattern must occur at the beginning of a sentence, and *beg.-allvar.* in which different variations of sentence beginnings are permitted before the pattern.

### 6.4.2 Term-based definition extraction widget

The second definition extraction operation of the web service is implemented in the term-based definition extraction widget that is primarily tailored to extract knowledge-rich contexts as it focuses on sentences that contain at least  $n$  domain-specific single or multi-word terminological expressions (terms). The term-based definition extraction uses the results of the term extraction web service. The parameters of this module are: the number of terms, the termhood threshold (defined as the percentage of terms, number of terms or termhood value itself), the number of terms in the nominative case (for Slovene), the constraints that a verb should figure between two terms, that the first term should be a multi-word term, and that the sentence should begin with a term. The evaluation of different parameter settings is given in Sections 5.1.2 and 5.2.2 for Slovene and English, respectively.

### 6.4.3 Wordnet-based definition extraction widget

The third approach, implemented by the *Wordnet-based definition extraction* widget, seeks for sentences where a wordnet term occurs together with its direct hypernym. For English we use the Princeton WordNet (PWN, 2010; Fellbaum, 1998), whereas for Slovene we use sloWNet (Fišer and Sagot, 2008), a Slovene counterpart of WordNet. The approach is evaluated and described in more detail in Sections 5.1.3 and 5.2.3.

## 6.5 Auxiliary widgets

### 6.5.1 Merge sentences widget

Sentence merger widget, allows the user to combine the results of several definition extraction methods. *Intersection* outputs the sentences that were extracted by at least two out of three methods, while *Union* takes the sentences extracted by any of the methods.

### 6.5.2 String to file widget

This widget available in ClowdFlows is used for saving the output of the file.

### 6.5.3 Term viewer widget

Term candidate viewer widget formats and displays the terms (in lemmatized and canonical form) and their scores returned by the term extractor widget.

Score	Term	Canonical
1.000000	language	language
1.000000	corpus linguistics	corpus linguistics
0.356373	corpus	corpus
0.325931	parallel corpus	parallel corpus
0.284885	language resource	language resources
0.283579	translation	translation
0.272218	language data	language data
0.251837	translation unit	translation units
0.191547	machine translation	machine translation
0.175470	foreign language	foreign language
0.149283	natural language	natural language
0.119362	bilingual dictionary	
0.118063	lexical semantics	lexical semantics
0.105274	translation equivalent	translation equivalents

Figure 14. Illustrating the term candidate viewer widget functionality.

### 6.5.4 Sentence viewer widget

Sentence viewer widget (cf. Figure 15) similarly to the term candidate viewer widget, formats and displays the candidate definition sentences returned by the corresponding methods.

Sentence ID	Value
103	Language resources are corpora and other lexical data : electronic versions of dictionaries for human users and lexicons for language technology applications .
121	TELRI is a pan-European network promoting the creation of language resources and the research on the methodology for extracting language data from corpora and for processing them to be used in multilingual language applications .
154	The TELRI Association is the independent pan-European voice of the multilingual research and development community , a devoted though impartial partner of the European language industry , and a respected consultant of the European Commission for the Multilingual Information Society .
157	The latter are language resources in the full sense only if they are derived from corpora , using methods developed in corpus linguistics .
169	Corpus linguistics is the sub-discipline of linguistics that deals with extracting language data from corpora and processing them for various applications such as grammars for human users and for computers , dictionaries , and lexicons .
178	In the beginning , corpora were mostly analyzed in view of grammatical phenomena and usage aspects of very frequent words , like function words .
217	In the European tradition , one focus had always been on the single word , that is , the character strings between blanks , and this concept of the nature of words has become the basis of most dictionary making .
243	Parallel corpora are texts and their translations , human translations , we should add , or at least translations supervised and post-edited by translators who understood both the source and the target text .

Figure 15. Illustrating the definition candidate viewer widget functionality.

In further work, we foresee to test the influence of preprocessing by comparing the ToTrTaLe widget with other preprocessing tools, such as Tree Tagger for English (Schmid, 1994) or recently developed Obeliks for Slovene (Grčar et al., 2012) or alternatively to retrain ToTrTaLe with larger and more recent corpora. For the term extraction, we could compare the English part of LUIZ with comparable systems, such as the one by Sclano and Velardi (2007) or Macken et al. (2013), if they were made available as web services. In the current LUIZ implementation FidaPLUS (Arhar Holdt and Gorjanc, 2007) is used as a reference corpus, but we could easily update it with the recently developed Gigafida (Logar Berginc et al., 2012). For the wordnet-based definition extraction using sloWNet, the widget can be updated by replacing the current sloWNet entries by the updated version.

To the best of our knowledge, the developed workflow is the only publicly available terminology and definition extraction workflow available online, which can be applied to new corpora, enhanced with new modules and adapted to new languages by other researchers. While this workflow includes some tools which were previously developed by other authors (ToTrTaLe by Erjavec, 2011; LUIZ by Vintar, 2010) these modules have been refined and implemented as web services (Pollak et al., 2012c), enabling their inspection and reuse by other NLP researchers.

## 7 Conclusions and further work

The present dissertation addresses the problem of domain modeling from multilingual corpora, focusing on the task of definition extraction, the main added value being the extraction of definitions from Slovene texts.

We addressed a real-life setting of modeling domain knowledge from existing corpora. For Slovene, there are many domains for which domain knowledge is not yet structured, and for which there are no terminological dictionaries or ontologies available. However, small domain corpora can be collected and used as a basis for automatic or semi-automatic domain modeling. In our case, we decided to focus on the domain of Language Technologies. We first constructed the Language Technologies Corpus, containing scientific articles, Bachelor's, Master's or PhD theses, as well as few book chapters and Wikipedia articles. The Slovene domain corpus of approx. 1 million word tokens was collected, preprocessed and annotated. Subsequently, we constructed a comparable corpus in English language, meaning that the corpus covers the same domain and is approximately the same size.

On this corpus (consisting of a Slovene and English subcorpus) we first performed semi-automatic topic ontology construction by using the OntoGen (Fortuna et al., 2007) semi-automatic and data-driven topic ontology editor. These topic ontologies (built for each language separately) were used for obtaining insight into the corpus, semi-automatically splitting the language technologies domain into two bigger subtopics (*speech technologies* and *language technologies*) and more specific subtopics for each field, e.g., in the Slovene subcorpus the *language technologies* domain was split into subdomains, such as *information extraction*, *computer-assisted translation*, *corpora* and *computational semantics*, each of these containing further subdomains. Having constructed this initial domain model, we were interested in extracting more specific domain knowledge, i.e., the domain terminology and definitions. For terminology extraction, we re-implemented the monolingual terminology extraction approach of the LUIZ system (Vintar, 2010), whereas the main focus and the main contribution of this dissertation is on semi-automatic definition extraction methodology, its implementation and the analysis of its results.

Several distinguishing aspects are characteristic of our work. Firstly, our definition extraction methodology is the only methodology available for Slovene. Secondly, since we focus on Slovene, we do not perform complex text preprocessing steps, and rely on simple PoS annotation only, due to the unavailability of a more sophisticated annotation software for Slovene when developing the methodology (in further versions the recently developed parser (Dobrovoljc et al. 2012) could be included in the methodology). Far from being interested only in the quantitative evaluation of the proposed definition extraction approach, one of the core points of the dissertation is also to extract a pilot language technologies glossary as well as to show and analyze a variety of definition types, and related problems. Finally, we implemented our definition extraction methodology—from the initial corpus preprocessing to the final inspection of domain terms and definitions—as a publically available workflow, where a user can, without

any installation, try and use the workflow for modeling his own corpora in Slovene or English or use specific workflow components in building new workflows.

The main definition extraction methodology that we have developed consists of three definition extraction modules for each language, i.e., the pattern-based, the term-based and the wordnet-based definition extraction module, while they can also be combined in a number of ways. The first—pattern-based—approach seeks out sentences corresponding to predefined lexico-syntactic patterns. Based on a preliminary analysis of some examples, a list of patterns was proposed, covering a large variety of definitions besides the standard “is\_a” definition type. We have also evaluated the performance of different patterns. For Slovene, the case information is often used, while for English, we also evaluated various settings concerning the position of the pattern (including an extra condition that the pattern we search for starts the sentence).

The term-based approach was designed primarily for extracting knowledge-rich contexts. A starting point is that a good definition candidate should contain at least two domain terms (possibly the definiendum and its hypernym, but also relational, extensional and other definition types are targeted). Next, other conditions—limiting the number of extracted sentences and mostly leading to increased precision—are added. These include the nominative condition demanding the domain terms to be in the nominative case (for Slovene only), the verb condition (verb should occur between two domain terms), the beginning of the sentence condition (one term should be the first or the second word in a sentence), the condition of the first term being a multi-word expression. Also the main parameter settings, i.e., the threshold parameter and the number of terms (and the number of multi-word terms) can be set in different ways in order to optimize the precision or recall.

The last approach to definition extraction is wordnet-based. Using the Princeton WordNet (PWN, 2010; Fellbaum, 1998) for English and sloWNet for Slovene (Fišer and Sagot, 2008), this approach aims to extract sentences that contain a term present in a sloWNet together with its direct hypernym. One of the problems observed with this approach is a low coverage of terms specific to the language technologies domain.

The three methods were combined in different ways, leading to a relatively low precision and recall compared to some state-of-the-art systems for English (e.g., Navigli and Velardi, 2010 extracting definitions for a list of terms specially from the Web), but comparable to related systems for Slavic languages (e.g., Kobylinski and Przepiórkowski, 2008 as part of the LT4eL project (2008) including Polish, Bulgarian and Czech language).

The focus of the dissertation was on in-depth analysis and discussion of different definition candidates, showing that the decision to classify a sentence as a definition or non-definition is a difficult task in itself, and that a vast majority of the examples dealt with are borderline cases. The sentences extracted from scientific articles or theses are often very complicated. During the thesis elaboration, we evaluated approx. 19,300 Slovene and 14,000 English sentences as definitions or non-definitions (of which more than 1,500 sentences (approx. 1,050 for Slovene and more than 500 for English) were classified as definitions. Moreover, more than 3,400 sentences (including 486 Slovene and 230 English definitions) were analyzed in more detail. This subset of 3,400 Slovene and English sentences was annotated with different tags, showing the complexity of the problem on such a difficult corpus. There are five different groups of tags. The first is the *definition form* category, which denotes the tags related to the analysis of the definition type. Within this category we observed that besides the main definition types

(e.g., *genus* and *differentia* definitions or paraphrases that do not have a special tag), other types of definitions occur, such as extensional definitions, definitions without hypernym (e.g., typifying or functional definitions), incomplete definitions where only a hypernym is provided or relational definitions (using *definiendum*'s synonyms, antonyms or sibling concepts). The next tag group analyzes *definition content* by discussing problems such as too general or too specific definition content, outdated or metaphorical definitions, etc. *Definiendum related* tags were used to add the information that a definiendum is e.g., a named entity, abbreviation or that the terms to be defined are terms out of the domain. The rest of the analysis concerns *segmentation related* issues (pointing to definitions that are e.g., spanning across several sentences or sentences containing several definitions) and *annotation related* issues (identifying doubles or sentences for which the label definition/non-definition should be reconsidered). The annotated dataset is valuable from the linguistic perspective, as well as a potential resource for a machine learning approach to be used in further work. The main definition/non-definition labels can be used for setting a simple classification task, while more fine-grained labels could help in setting up a system for ranking the extracted definitions, or in the development of systems for extracting semantic relations. Some labels could have specific use, such as segmentation labels, that can be used for further improving segmentation errors.

The presented qualitative analysis complements well the quantitative evaluation. With some exceptions (e.g., Westerhout, 2010 working on glossary creation for Dutch), the qualitative analysis of results is often ignored. As it has been shown by the examples of definition candidates extracted from our corpus, the corpus is far from containing simple sentences and represents difficult material for automatic definition extraction. An inter-annotator agreement (IAA) experiment showed that even human evaluators do not easily agree on what a definition is. In our case, the fixed marginal kappa is only 0.33 (we foresee to repeat the experiments by defining the criteria more strictly and check if the IAA score improves). However, this situation is very realistic: for Slovene, new (or constantly growing) domains only rarely have well structured resources, such as Wikipedia entries or textbooks, or large amount of text available on the Web (since in some domains authors publish their work mainly in English). Therefore, a limited amount of academic papers and similar works can be used as material for definition extraction.

Finally, an important contribution of the present dissertation is also the implementation of the definition extraction pipeline in the ClowdFlows workflow construction environment, meaning that the proposed definition extraction workflows, as well as their constituting parts can be used online, with no prior installation required. This is an important benefit for the Slovene language technologies community (even a very simple tool, such as the ToTrTaLe web service for corpus annotation, is an important contribution for any further Slovene NLP workflow). On the other hand, the terminology and definition extraction available in a workflow is, to the best of our knowledge, the only system available in an online workflow, and can therefore be very easily combined with other NLP components, even for English.

In future research, we will act in several directions. Since we have a modular workflow, an obvious step to take is to add to the current implementation other preprocessing tools—such as Tree Tagger for English (Schmid, 1994) or recently developed Obeliks for Slovene (Grčar et al., 2012)—as well as other term extraction systems (e.g., Sclano and Velardi, 2007), if made available as web services. In addition, we foresee to continue the research in the following lines. First, machine learning

methods will be used trying to improve the results (our preliminary experiments were reported in Fišer et al. (2010)). In new experiments we will use the insights (transformed into features) from the experiments presented in this thesis, and implement the system ranking the extracted definition candidates. Next, the methodology for automatic term-alignment from comparable corpora will be implemented in the workflow environment and tested on our corpus (our initial work is presented in Ljubešić et al., 2011 and Fišer et al., 2011). Moreover, as shown in a quick experiment, the quantitative results are very subjective and depend greatly on evaluation criteria. We will re-evaluate a part of our dataset by the five scores scale as proposed in Reiplinger et al. (2012). Our major focus will be on performing new experiments on different (possibly less complex) corpora and explore the influence of text type on definition extraction, as well as in comparison of our approach with other systems, where we will consider training the word-class lattices with Wikipedia as proposed in Faralli and Navigli (2013). In addition, we will consider other possible applications of our methodology, e.g., as a possible preprocessing step for knowledge discovery tasks, since—if the parameters are set adequately—one can filter the text by preserving only knowledge-rich sentences.

#### **List of URLs of developed tools and resources:**

- The developed *PaTeW definition extraction workflow*:  
<http://clowdflows.org/workflow/1380>
- The implemented *ToTrTaLe workflow*: <http://clowdflows.org/workflow/228>
- The pilot *Language technologies glossary*: [http://kt.ijs.si/senja\\_pollak/jt\\_glosar/](http://kt.ijs.si/senja_pollak/jt_glosar/)
- Part of the *corpus* available through a concordancer: [http://nl.ijs.si/cuwi/sdjt\\_sl/](http://nl.ijs.si/cuwi/sdjt_sl/)
- The reimplemented LUIZ *terminology* extractor as part of the PaTeW workflow.

## 8 References

- Ahmad, K., L. Gillam, and L. Tostevin. 2007. "University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Rerieval (WILDER)." In *Proceedings of the Eight Text REtrieval Conference (TREC-8)*, 717–724.
- Anderson, Ashton, Dan Jurafsky, and Daniel A. McFarland. 2012. "Towards a Computational History of the ACL: 1980-2008." In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, 13–21. Jeju Island, Korea: Association for Computational Linguistics.
- Arhar Holdt, Špela, and Vojko Gorjanc. 2007. "Korpus FidaPLUS: Nova generacija slovenskega referenčnega korpusa." *Jezik in slovstvo* 52 (2): 95–110.
- Atkins, Sue, B. T., and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.
- Ayto, John. 1983. "On Specifying Meaning: Semantic Analysis and Dictionary Definitions." In *Lexicography: Principles and Practice*, edited by Reinhard Hartmann, 89–98. London: Academic Press.
- Banchs, Rafael E., ed. 2012. "Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries." In Jeju Island, Korea: Association for Computational Linguistics.
- Baker, Mona, and Gabriela Saldanha. 2009. *Routledge Encyclopedia of Translation Studies*. 2nd ed. Routledge (Taylor and Francis Group).
- Barnbrook, Geoff, and John Sinclair. 1994. "Parsing Cobuild Entries." In *The Languages of Definition: The Formalisation of Dictionary Definitions for Natural Language Processing*, edited by John Sinclair, Martin Hoelter, and Carol Peters, 13–58. Luxembourg: European Commission.
- Béjoint, Henri. 2000. *Modern Lexicography: An Introduction*. Oxford: Oxford University Press.
- Bergenholtz, Henning. 1995. "Wodurch Unterscheidet Sich Fachlexikographie von Terminologie." *Lexicographica* 11: 50–59.
- Berland, Matthew, and Eugene Charniak. 1999. "Finding Parts in Very Large Corpora." In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, 1910:57–64.

- Berthold, Michael R, Nicolas Cebron, Fabian Dill, Thomas R Gabriel, Tobias Kötter, Thorsten Meinel, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. 2008. "KNIME: The Konstanz Information Miner." *Data Analysis Machine Learning and Applications* 11: 319–326.
- Bessé, Bruno de, Blaise Nkwenti-Azeh, and Juan C. Sager. 1997. "Glossary of Terms Used in Terminology." *Terminology* 4 (1): 117–156.
- Biemann, Chris. 2005. "Ontology Learning from Text – a Survey of Methods." *LDV-Forum* 20 (2): 75–93.
- Bird, Steven, R. Dale, B. Dorr, B. Gibson, M. Joseph, M.-Y. Kan, D. Lee, B. Powley, D. Radev, and Y.-F. Tan. 2008. "The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics." In *Proceedings of the 6th International Language Resources and Evaluation Conference (LREC 2008)*. Marrakech, Morocco.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.
- "BNC." 2001. *The British National Corpus, Version 2 (BNC World)*. Mike Scott's list available at: [http://www.lexically.net/downloads/BNC\\_wordlists/downloading\\_BNC.htm](http://www.lexically.net/downloads/BNC_wordlists/downloading_BNC.htm). Last accessed January 15, 2012.
- Borg, Claudia. 2009. "Automatic Definition Extraction Using Evolutionary Algorithms". Master's thesis. University of Malta.
- Borg, Claudia, Michael Rosner, and Pace Gordon. 2009. "Evolutionary Algorithms for Definition Extraction." In *Proceedings of the 1st Workshop in Definition Extraction*. Borovets, Bulgaria.
- Borg, Claudia, Mike Rosner, and Gordon J Pace. 2010. "Automatic Grammar Rule Extraction and Ranking for Definitions." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, edited by Khalid Calzolari, Nicoletta Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, 2577–2584. Valletta, Malta: European Language Resources Association (ELRA).
- Borsodi, R. 1967. *The Definition of Definition*. Porter Sargent Publisher.
- Brants, Thorsten. 2000. "TnT Speech Tagger." In *Proceedings of the 6th Natural Language Processing Conference*, 224–231. Seattle, WA.
- Buitelaar, P., P. Cimiano, and B. Magnini. 2005. *Ontology Learning from Text: Methods, Evaluation and Applications*. (123 of Fr. IOS Press.
- Cabré Castellví, M. Teresa. 2003. "Theories of Terminology: Their Description, Prescription and Explanation." *Terminology* 9 (2): 163–199.

- Cabré, Maria Teresa. 1999. *Terminology: Theory, Methods, and Applications*. Edited by Juan C. Sager. Amsterdam-Philadelphia: John Benjamins Publishing.
- Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20: 37–46.
- Cohen, S. Marc. 2012. "Aristotle's Metaphysics." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2011 ed. Available at: <http://plato.stanford.edu/entries/aristotle-metaphysics/>. Last accessed in February 2013.
- Cohen, Trevor, and Dominic Widdows. 2009. "Empirical Distributional Semantics: Methods and Biomedical Applications." *Journal of Biomedical Informatics* 42 (2): 390–405.
- Copi, Irving M., and Carl Cohen. 2009. *Introduction to Logic*. 13th ed. Upper Saddle River, New Jersey: Pearson/Prentice Hall.
- Cui, Hang, Min-Yen Kan, and Tat-Seng Chua. 2007. "No Soft Pattern Matching Models for Definitional Question Answering." *ACM Transactions on Information Systems (TOIS)* 25 (2).
- De Bessé, Bruno. 1994. "Contribution à La Définition de La Terminologie." In *Langues et Sociétés En Contact. Mélanges Offerts à Jean-Claude Corbeil. (Candiana Romanica, Vol 8.)*, edited by Pierre Martel and Jacques Maurais, 135–138. Tübingen: Max Niemeyer.
- Degórski, Lukasz, Łukasz Kobylński, and Adam Przepiórkowski. 2008a. "Definition Extraction: Improving Balanced Random Forests." In *Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT~2008): Computational Linguistics -- Applications (CLA'08)*, 353–357. Wisła, Poland: IEEE.
- Degórski, Lukasz, Michał Marcińczuk, and Adam Przepiórkowski. 2008b. "Definition Extraction Using a Sequential Combination of Baseline Grammars and Machine Learning Classifiers." In *Proceedings of the 6th International Language Resources and Evaluation Conference (LREC 2008)*, 837–841. Marrakech, Morocco: European Language Resources Association (ELRA).
- Del Gaudio, Rosa, Gustavo Batista, and António Branco. 2013. "Coping with Highly Imbalanced Datasets: A Case Study with Definition Extraction in a Multilingual Setting." *Natural Language Engineering (FirstView)*: 1–33.
- Del Gaudio, Rosa, and António Branco. 2007. "Automatic Extraction of Definitions in Portuguese: A Rule-Based Approach." In *Proceedings of the 13th Portuguese Conference on Artificial Intelligence (EPIA2007)*. LNAI 4874., edited by José Neves, Manuel Filipe Santos, and José Manuel Machado, 4874:659 – 670. Springer Berlin Heidelberg.

- Demšar, Janez, Blaž Zupan, Gregor Leban, and Tomaz Curk. 2004. "Orange: From Experimental Machine Learning to Interactive Data Mining." In *Knowledge Discovery in Databases: PKDD 2004. LNCS Vol. 3202.*, edited by Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, 3202:537–539. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Dobrovoljc, Kaja, Simon Krek, and Jan Rupnik. 2012. "Skladenjski Razčlenjevalnik Za Slovenščino." In *Proceedings of the Eighth Language Technologies Conference*, edited by Tomaž Erjavec and Jerneja Žganec Gros. Ljubljana: Institut Jožef Stefan.
- Dubois, J., and C. Dubois. 1971. *Introduction à La Lexicographie: Le Dictionnaire*. Paris: Larousse.
- Dolezal, Frederic. 1992. "The Meaning of Definition." *Lexicographica*: 1–9.
- Erjavec, Tomaž. 2011. "Automatic Linguistic Annotation of Historical Language: ToTrTaLe and XIX Century Slovene." In *Proceedings of the ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (ACL 2011).
- Erjavec, Tomaž. 2012a. "The Goo300k Corpus of Historical Slovene." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC~2012)*, 2257–2260. Istanbul, Turkey.
- Erjavec, Tomaž. 2012b. "MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages." *Language Resources and Evaluation* 46 (1): 131–142.
- Erjavec, Tomaž, and Sašo Džeroski. 2004. "Machine Learning of Language Structure: Lemmatising Unknown Slovene Words." *Applied Artificial Intelligence* 18 (1): 17–41.
- Erjavec, Tomaž, Darja Fišer, Simon Krek, and Nina Ledinek. 2010. "The JOS Linguistically Tagged Corpus of Slovene." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, edited by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, 1806–1809. Valletta, Malta: European Language Resources Association (ELRA).
- Erjavec, Tomaž, Camelia Ignat, Bruno Pouliquen, and Ralf Steinberger. 2005. "Massive Multi-Lingual Corpus Compilation: Acquis Communautaire and ToTaLe." In *Proceedings of the 2nd Language & Technology Conference* (April 21–23, 2005), 32–36. Poznan, Poland.
- Fahmi, Ismail, and Gosse Bouma. 2006. "Learning to Identify Definitions Using Syntactic Features." In *Proceedings of the EACL Workshop on Learning Structured Information in Natural Language Applications*.

- Faralli, Stefano, and Roberto Navigli. 2013. "A Java Framework for Multilingual Definition and Hypernym Extraction." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013, Sofia, Bulgaria)*, 103–108. Association for Computational Linguistics
- Feliu, Judit. 2004. "Relacions Conceptuals i Terminologia: Anàlisi i Proposta de Detecció Semiautomàtica". University Pompeu Fabra, Barcelona.
- Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. Edited by Christiane Fellbaum. Cambridge, MA: MIT Press.
- Fišer, Darja. 2009. "Izdelava slovenskega semantičnega leksikona z uporabo eno- in večjezičnih jezikovnih virov". PhD thesis. Faculty of Arts, University of Ljubljana.
- Fišer, Darja, Nikola Ljubešić, Špela Vintar, and Senja Pollak. 2011. "Building and Using Comparable Corpora for Domain-Specific Bilingual Lexicon Extraction." In *Proceedings of the 4th BUCC Workshop: Comparable Corpora and the Web (24 June, 2011)*, 19–26. Portland, Oregon: Association for Computational Linguistics.
- Fišer, Darja, Senja Pollak, and Špela Vintar. 2010. "Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, edited by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias. Valletta, Malta: European Language Resources Association (ELRA).
- Fišer, Darja, and Benoît Sagot. 2008. "Combining Multiple Resources to Build Reliable Wordnets." In *Text, Speech and Dialogue: 11th International Conference, (TSD 2008), September 2008. LNCS.*, edited by Petr Sojka, Ales Horák, Ivan Kopeček, and Karel Pala, 5246:61–68. Brno, Czech Republic: Springer.
- Fortuna, Blaž, Marko Grobelnik, and Dunja Mladenić. 2006a. "Semi-Automatic Construction of Topic Ontologies." Edited by Markus Ackermann, Bettina Berendt, Marko Grobelnik, Andreas Hotho, Dunja Mladenić, Giovanni Semeraro, Myra Spiliopoulou, Gerd Stumme, Vojtěch Svátek, and Maarten Van Someren. *Lecture Notes in Computer Science (LNCS). Semantics, Web and Mining (Revised Selected Papers of Joint International Workshop, EWMF 2005 and KDO 2005 (October 3-7, 2005))* 4289: 121–131.
- Fortuna, Blaž, Marko Grobelnik, and Dunja Mladenić. 2006b. "Semi-Automatic Data-Driven Ontology Construction System." In *Proceedings of the 9th International Multi-Conference Information Society IS-2006*. Ljubljana, Slovenia.
- Fortuna, Blaž, Marko Grobelnik, and Dunja Mladenić. 2007. "OntoGen: Semi-Automatic Ontology Editor." In *Human Interface (Part II) (HCI 2007), LNCS*, edited by M. J. Smith and G. Salvendy, 4558:309–318. Springer.
- Fortuna, Blaž, Dunja Mladenić, and Marko Grobelnik. 2006c. "Visualization of Text Document Corpus." *Informatica* 29: 497–502.

- Frantzi, K.T., and S. Ananiadou. 1999. "The CValue/NCValue Domain Independent Method for Multi-Word Term Extraction." *Journal of Natural Language Processing* 6 (3): 145–179.
- Fuertes-Olivera, Pedro A., and Henning Bergenholtz, ed. 2010. *E-Lexicography. The Internet, Digital Initiatives and Lexicography*. London, New York: Continuum.
- Gantar, Polona, and Simon Krek. 2009. "Drugačen pogled na slovarske definicije: opisati, pojasniti, razložiti?" In *Infrastruktura slovenščine in slovenistike*, edited by Marko Stabej, 151–159. Ljubljana: Znanstvena založba Filozofske fakultete.
- Garcia, Daniela. 1997. "Structuration Du Lexique de La Causalité et Réalisation D'un Outil D'aide Au Repérage de L'action Dans Les Textes." In *Proceeding de "Rencontres Terminologies et Intelligence Artificielle" (TIA 97)*. Toulouse, France.
- Gaussier, Eric. 1998. "Flow Network Models for Word Alignment and Terminology Extraction From Bilingual Corpora." In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (Coling-ACL)*, 444–450.
- Geeraerts, Dirk. 2003. "Meaning and Definition." In *A Practical Guide to Lexicography*, edited by P. G. J. Van Sterkenburg, 83–93. John Benjamins Publishing.
- Geeraerts, Dirj. 2010. *Theories of Lexical Semantics*. Oxford: Oxford University Press
- Gergonne, J. D. 1818. "Essai Sur La Théorie Des Définitions." *Annales de Mathématiques Pures et Appliquées* 9.
- Girju, Roxana, Adriana Badulescu, and Dan Moldovan. 2003. "Learning Semantic Constraints for the Automatic Discovery of Part-Whole Relations." In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, 1–8. Morristown, NJ, USA: Association for Computational Linguistics.
- Girju, Roxana, Adriana Badulescu, and Dan Moldovan.. 2006. "Automatic Discovery of Part-Whole Relations." *Computational Linguistics* 32 (1): 83–135.
- Glanzberg, Michael. 2011. "Meaning, Concepts, and the Lexicon". *Croatian Journal of Philosophy* 31: 1-29.
- Gómez-Pérez, Asuncion, and David Manzano-Macho. 2003. "A Survey of Ontology Learning Methods and Techniques (Deliverable 1.5)."
- Granger, Sylviane. 2012. "Introduction: Electronic Lexicography-from Challenge to Opportunity." In *Electronic Lexicography*, edited by Sylviane Granger and Magali Pacqot, 1–15. Oxford University Press.

- Granger, Sylviane, and Magali Paquot, ed. 2012. *Electronic Lexicography*. Oxford University Press.
- Grčar, Miha, Simon Krek, and Kaja Dobrovoljc. 2012. "Obeliks: Statistični Oblikoskladenjski Označevalnik in Lematizator Za Slovenski Jezik." In *Proceedings of the Eighth Language Technologies Conference*, edited by V T. Erjavec; J. Žganec Gros (ur.). Ljubljana: Institut Jožef Stefan.
- Gruber, Thomas. 1993. "Towards Principles for the Design of Ontologies Used for Knowledge Sharing." Edited by N Guarino and R Poli. *Formal Ontology in Conceptual Analysis and Knowledge Representation* 43 (5-6): 907–928.
- Gupta, Parth, and Paolo Rosso. 2012. "Text Reuse with ACL: (Upward) Trends." In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, 76–82. Jeju Island, Korea: Association for Computational Linguistics.
- Hall, David Leo Wright, Daniel Jurafsky, and Christopher Manning. 2008. "Studying the History of Ideas Using Topic Models." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, 363–371. ACL.
- Hanks, Patrick. 1987. "Definitions and Explanations." In *Looking Up: An Account of the COBUILD Project in Lexical Computing*, edited by John Sinclair, 116–136. London-Glasgow: Collins.
- Harris, Zellig Sabbetai. 1954. "Distributional Structure." *Word* 10 (2/3): 146–162.
- Harris, Zellig Sabbetai. 1968. *Mathematical Structures of Language*. New York: Interscience Publishers John Wiley & Sons.
- Hearst, Marti A. 1992. "Automatic Acquisition of Hyponyms from Large Text Corpora." In *Proceedings of the 14th Conference on Computational Linguistics-Volume 2 (COLING '92)*, 2:539–545. Nantes, France.: Association for Computational Linguistics.
- Heuberger, R. 2000. *Monolingual Dictionaries for Foreign Learners of English: A Constructive Evaluation of the State-of-the-Art. Reference Works in Book Form and on CD-ROM*. Vienna: Wilhelm Braumüller.
- Hjelmslev, Louis. 1975. "Résumé of a Theory of Language." *Travaux Du Cercle Linguistique de Copenhague* 16. Copenhagen: Nordisk Sprogog Kulturforlag.
- Hull, Duncan, Katy Wolstencroft, Robert Stevens, Carole Goble, Mathew R Pocock, Peter Li, and Tom Oinn. 2006. "Taverna: a Tool for Building and Running Workflows of Services." *Nucleic Acids Research* 34: W729–W732.
- Humbley, John. 1997. "Is Terminology Specialized Lexicography? The Experience of French-Speaking Countries." *Hermes* 18: 13–32.

- Iftene, Adrian, Diana Trandaba, and Ionut Pistol. “Natural Language Processing and Knowledge Representation for e-Learning Environments.” In *Proceedings of Applications for Romanian. Proceedings of RANLP 2007 Workshop.*, 19–25.
- Itagaki, Masaki, Takako Aikawa, and Xiaodong He. 2007. “Automatic Validation of Terminology Translation Consistency with Statistical Method.” In *Proceedings of the Machine Translation Summit XI*, 269–274. Copenhagen, Denmark.
- “ISO 1087-1:2000a.” *International Standard: Terminology Work — Vocabulary — Part 1: Theory and Application*. (Standard cited from the Glossary of Terminology Management of DG TRAD – Terminology Coordination Unit of European Parliament). Last accessed September 3, 2013. <http://termcoord.wordpress.com/glossaries/glossary-of-terminology-management/>.
- “ISO 1087-1:2000b.” 2013. *International Standard: Terminology Work — Vocabulary — Part 1: Theory and Application*. (Standard cited from ISOcat Web Interface). Last accessed December 1, 2013. <https://catalog.clarin.eu/isocat/interface/index.html>.
- “ISO 1087-2.” 2000. *International Standard: Terminology Work - Vocabulary - Part 2: Computer Applications* (Standard withdrawn in 2011). (Standard cited from ISOcat Web Interface). Last accessed December 1, 2013. <https://catalog.clarin.eu/isocat/interface/index.html>.
- “ISO 1087:1990.” 2013. *International Standard: Terminology - Vocabulary*. (Standard withdrawn and replaced by ISO 1087-2000.) (Standard cited from Ken Sall’s Consulting XML TEchnology Webpage and from Strehlow and Wrigh, 1993). Accessed December 1, 2013. <http://kensall.com/gov/glossary/glossary.dtd.txt>.
- “ISO 12620:2009.” *International Standard. Terminology and Other Language and Content Resources — Specification of Data Categories and Management of a Data Category Registry for Language Resources*. (Standard cited from ISOcat Web Interface). Last accessed December 1, 2013. <https://catalog.clarin.eu/isocat/interface/index.html>.
- “ISO 5127:2001.” *International Standard: Information and Documentation — Vocabulary*. (Standard cited from the Glossary of Terminology Management of DG TRAD – Terminology Coordination Unit of European Parliament). Last accessed September 3, 2013. <http://termcoord.wordpress.com/glossaries/glossary-of-terminology-management/>.
- Ittoo, Ashwin, and Gosse Bouma. 2009. “Semantic Selectional Restrictions for Disambiguating Meronymy Relations.” In *Proceedings of the 19th Meeting of Computational Linguistics in the Netherlands*, edited by Barbara Plank, Erik Tjong Kim Sang, and Tim Van de Cruys.
- Jackson, Howard. 2002. *Lexicography: An Introduction*. Routledge.

- Kageura, Kyo. 2002. *The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth*. John Benjamins Publishing.
- Kan, Min-Yen, and Steven Bird. 2013. "ACL Anthology." *ACL Anthology. A Digital Archive of Research Papers in Computational Linguistics*. (Kan, Min-Yen editor, 2008-; Bird, Steven editor, 2001-2007)). Last accessed November 28. <http://aclweb.org/anthology/>.
- Kobyliński, Łukasz, and Adam Przepiórkowski. 2008. "Definition Extraction with Balanced Random Forests." In *Advances in Natural Language Processing: Proceedings of the 6th International Conference on Natural Language Processing, GoTAL~2008*, edited by Bengt Nordström and Aarne Ranta, 5221:237–247. Gothenburg, Sweden: Springer Berlin Heidelberg, LNCS.
- Kosem, Iztok. 2006. "Definicijski Jezik v Slovarju Slovenskega Knjižnega Jezika s Stališča Sodobnih Leksikografskih Načel." *Jezik in Slovnstvo* 51 (5): 25–45.
- Kozakov, L., Y. Park, T. Fin, Y. Drissi, Y. Doganata, and T. Cofino. 2004. "Glossary Extraction and Utilization in the Information Search and Delivery System for IBM Technical Support." *IBM Systems Journal* 43 (3): 546–563.
- Kozareva, Zornitsa, and Eduard Hovy. 2010. "A Semi-Supervised Method to Learn and Construct Taxonomies Using the Web." In *Proceedings of RANLP*, 1110–1118. Cambridge, MA.
- Kranjc, Janez, Vid Podpečan, and Nada Lavrač. 2012. "ClowdFlows: A Cloud Based Scientific Workflow Platform." In *Proceedings of ECML/PKDD-2012 (2)*, edited by Peter A Flach, Tijl De Bie, and Nello Cristianini, 7524:816–819. Springer LNCS.
- Krek, Simon. 2004. "Slovarji Serije COBUILD in Formalizacija Definicijskega Jezika." *Jezik in Slovnstvo* 49 (2): 3–16.
- Krek, Simon, Iztok Kosem, and Krek Gantar. 2013. "Predlog Za Izdelavo Slovarja Sodobnega Slovenskega Jezika." Version 1 (May 20, 2013) [http://www.sssj.si/datoteke/Predlog\\_SSSJ.pdf](http://www.sssj.si/datoteke/Predlog_SSSJ.pdf).
- Kupiec, Julian. 1993. "An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora." In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL)*. Columbus, Ohia, United States.
- Logar Berginc, Nataša, Miha Grčar, Marko Brakus, Tomaž Erjavec, Špela Arhar Holdt, and Simon Krek. *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES : gradnja, vsebina, uporaba*. Ljubljana: Trojina.
- L'Homme, Marie-Claude. 2004. *La Terminologie: Principes et Techniques*. Presses de l'Université de Montréal.

- L'Homme, Marie-Claude, and Elizabeth Marshman. 2006. "Extracting Terminological Relationships from Specialized Corpora." In *Lexicography, Terminology, Translation: Text-Based Studies in Honour of Ingrid Meyer*, edited by L. Bowker, 67–80. Ottawa: University of Ottawa Press.
- Lachowicz, Dom, and Caolán McNamara. 2006. "wvWare, Library for Converting Word Document."
- Lapata, Mirella, and Sebastian Padó. 2007. "Dependency-Based Construction of Semantic Space Models." *Computational Linguistics* 33: 161–199.
- Lefever, Els, Lieve Macken, and Veronique Hoste. 2009. "Language-Independent Bilingual Terminology Extraction from a Multilingual Parallel Corpus." In *Proceedings of the 12th Conference of the European Chapter of the ACL*, 496–504.
- Lewis, C. I. 1929. *Mind and the World-Order*. Chicago, IL: Charles Scribner's Sons.
- Ljubešić, Nikola, Darja Fišer, Špela Vintar, and Senja Pollak. 2011. "Bilingual Lexicon Extraction from Comparable Corpora: a Comparative Study." In *Proceedings of the 1st International Workshop on Lexical Resources (An ESSLLI 2011 Workshop)*. Ljubljana, Slovenia.
- Lowry, Richard. 2013. "Kappa as a Measure of Concordance in Categorical Sorting." <http://vassarstats.net/kappa.html>. Last accessed: December 3, 2013.
- Lyons, John. 1977. *Semantics*. Cambridge: Cambridge University Press.
- LT4eL (2008). Language Technology for eLearning project: [www.lt4el.eu](http://www.lt4el.eu). Last accessed: June 3, 2013.
- Macken, Lieve, Els Lefever, and Veronique Hoste. 2013. "TEXSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-Based Alignment." *Terminology* 19 (1): 1–30.
- Maedche, Alexander, and Steffen Staab. 2009. "Ontology Learning." In *Handbook on Ontologies*, 245–268. Springer.
- Malaisé, Véronique, Pierre Zweigenbaum, and Bruno Bachimont. 2004. "Detecting Semantic Relations Between Terms In Definitions." In *COLING 2004 CompuTerm 2004: 3rd International Workshop on Computational Terminology*, edited by Sophia Ananadiou and Pierre Zweigenbaum, 55–62.
- Marshman, Elizabeth, Tricia Morgan, and Ingrid Meyer. 2002. "French Patterns for Expressing Concept Relations." *Terminology* 8 (1): 1–29.
- May, Kerstin. 2005. "English Monolingual Advanced Learners Dictionaries on CD-ROM: A Comparative Evaluation." Master's thesis. Technische Universität Chemnitz.

- Meyer, Ingrid. 1994. "Linguistic Strategies and Computer Aids for Knowledge Engineering in Terminology." *L'actualité terminologique/Terminology Update* 27 (4): 6–10. Ottawa: Public Works and Government Services Canada.
- Meyer, Ingrid. 2001. "Extracting Knowledge-Rich Contexts for Terminography: A Conceptual and Methodological Framework." In *Recent Advances in Computational Terminology*, edited by Didier Bourigault, Christian Jacquemin, and Marie-Claude L'Homme, 279–302.
- Meyer, Ingrid, Kristen Mackintosh, Caroline Barrière, and Tricia Morgan. 1999. "Conceptual Sampling for Terminographical Corpus Analysis." In *Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE '99)*, 256–267. Innsbruck, Austria.
- Mierswa, Ingo, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. 2006. "YALE: Rapid Prototyping for Complex Data Mining Tasks." In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, edited by T. Eliassi-Rad, L. H. Ungar, M. Craven, and D. Gunopulos, 2006:935–940.
- Muresan, Smaranda, and Judith Klavans. 2002. "A Method for Automatically Building and Evaluating Dictionary Resources." In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain)*. European Language Resources Association.
- Müller, Jakob. 2008. "Kritične Misli in Zamisli o SSKJ." In *Strokovni Posvet o Novem Slovarju Slovenskega Jezika*, edited by Andrej Perdih, 17–25. Ljubljana: ZRC SAZU.
- Murphy, Lynne M. 2010. *Lexical Meaning*. Cambridge: Cambridge University Press.
- Myking, Johan. 2007. "No Fixed Boundaries." In *Indeterminacy in Terminology and LSP: Studies in Honour of Heribert Picht*, edited by Antia Bassegy, 73–91. Amsterdam (The Netherlands) and Philadelphia, USA: John Benjamins Publishing.
- Nakamoto, Kyohei. 1998. "From Which Perspective Does the Definer Define the Definiendum: Anthropocentric or Referent-Based?" *International Journal of Lexicography* 11 (3): 205–218.
- Navigli, Roberto, and Paola Velardi. 2010. "Learning Word-Class Lattices for Definition and Hypernym Extraction." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010, Uppsala, Sweden)*, 1318–1327.
- Navigli, Roberto, Paola Velardi, and Stefano Faralli. 2011. "A Graph-Based Algorithm for Inducing Lexical Taxonomies from Scratch." In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, 1872–1877. Barcelona, Spain.

- Pantel, Patrick, and Marco Pennacchiotti. 2006. “Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations.” In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL-06)*:113–120. Sydney, Australia.
- Paradis, Carita. 2012. “Lexical Semantics” *The Encyclopedia of Applied Linguistics*, ed. Chapelle, C.A. Oxford, UK: Wiley-Blackwell
- Parry, William Thomas, and Edward A. Hacker. 1991. *Aristotelian Logic*. Albany, NY: State University of New York Press.
- Paul, Michael, and Roxana Girju. “Topic Modeling of Research Fields: An Interdisciplinary Perspective.” In *Recent Advances in Natural Language Processing (RANLP 2009)*, Borovets, Bulgaria.
- Pearson, Jennifer. 1996. “The Expression of Definitions in Specialised Texts: A Corpus-Based Analysis.” In *Euralex 96 Proceedings*, 2:817–824.
- Pearson, Jennifer. 1998. *Terms in Context*. SCL Series, Vol. 1. Edited by Elena Tognini-Bonelli and Wolfgang Teubert. Amsterdam, The Netherlands and Philadelphia, USA: John Benjamins Publishing.
- Planas, Emmanuel. 2005. “SIMILIS. Second-Generation Translation Memory Software.” In *Proceedings of the 27th International Conference on Translating and the Computer (TC27)*. London, UK.
- Podpečan, Vid, Monika Zemenova, and Nada Lavrač. 2012. “Orange4WS Environment for Service-Oriented Data Mining.” *The Computer Journal* 55 (1): 82/98.
- Pollak, Senja, Anže Vavpetič, Janez Kranjc, Nada Lavrač, and Špela Vintar. 2012a. “NLP Workflow for online Definition Extraction from English and Slovene Text Corpora.” In *Proceedings of the 11th Conference on Natural Language Processing (KONVENS 2012, Vienna, Austria, September 19-21, 2012)*, edited by E. Jancsary, 53–60.
- Pollak, Senja, Nejc Trdin, Anže Vavpetič, and Tomaž Erjavec. 2012b. “A Web Service Implementation of Linguistic Annotation for Slovene and English.” In *Proceedings of the 8th Language Technologies Conference, Proceedings of the 15th International Multiconference Information Society (IS 2012)*, 157–162.
- Pollak, Senja, Nejc Trdin, Anže Vavpetič, and Tomaž Erjavec. 2012c. “NLP Web Services for Slovene and English”: Morphosyntactic Tagging, Lemmatisation and Definition Extraction.” *Informatica* 36: 441–449.
- Przepiórkowski, Adam. 2007. “Slavonic Information Extraction and Partial Parsing.” In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing at ACL 2007*, edited by Jakub Piskorski, Bruno Pouliquen, Ralf Steinberger, and Hristo Tanev, 1–10. Prague.

- Przepiórkowski, Adam, Lukasz Degórski, Miroslav Spousta, Kiril Simov, Petya Osenova, Lothar Lemnitzer, Vladislav Kuboň, and Beata Wójtowicz. 2007. "Towards the Automatic Extraction of Definitions in Slavic." In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing at ACL 2007*, edited by Jakub Piskorski, Bruno Pouliquen, Ralf Steinberger, and Hristo Tanev, 43–50. Prague.
- "PWN." 2010. *Princeton University "About WordNet."* *WordNet. Princeton University.* <http://wordnet.princeton.edu>. Last accessed: November 3, 2011.
- Radev, Dragomir R., Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. "The ACL Anthology Network Corpus." In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, 54–61. Association for Computational Linguistics.
- Radev, Dragomir R., Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. "The ACL Anthology Network Corpus." *Language Resources and Evaluation* 47: 919–944.
- Randolph, J. Justus. 2005. "Free-Marginal Multirater Kappa: An Alternative to Fleiss' Fixed-Marginal Multirater Kappa." Paper Presented at the Joensuu University Learning and Instruction Symposium 2005, (October 14-15th, 2005, Oensuu, Finland). ERIC Document Reproduction.
- Randolph, J. Justus. 2008. "Online Kappa Calculator." [justusrandolph.net/kappa/](http://justusrandolph.net/kappa/) Last accessed: September 2, 2013.
- Reiplinger, Melanie, Ulrich Schäfer, and Magdalena Wolska. 2012. "Extracting Glossary Sentences from Scholarly Articles: A Comparative Evaluation of Pattern Bootstrapping and Deep Analysis." In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, 55–65. Jeju Island, Korea: Association for Computational Linguistics.
- Robinson, Richard. 1954. *Definition*. Clarendon Press.
- Robinson, Richard. 1963. *Definition*. Clarendon Press.
- Robinson, Richard. 1972. *Definitions*. Oxford: Oxford University Press.
- Rozman, Tadeja. 2010. "Vloga Enojezičnega Razlagalnega Slovarja Slovenščine Pri Razvoju Jezikovne Zmožnosti". PhD thesis. University of Ljubljana.
- Saussure, Ferdinand De. 1997. *Predavanja iz splošnega jezikoslovja*. Ljubljana: Studia Humanitatis.
- Schmid, Helmut. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees." In *Proceedings of the International Conference on New Methods in Language Processing.*, 44–49. Manchester, United Kingdom.

- Schmied, Josef. 2007. "The Chemnitz Corpus of Specialised and Popular Academic English." *Studies in Variation, Contacts and Change in English 2*.
- Sclano, F., and Paola Velardi. 2007. "TermExtractor: a Web Application to Learn the Common Terminology of Interest Groups and Research Communities." In *Proceedings of the 9th Conf on Terminology and Artificial Intelligence TIA 2007*: 8–9.
- Shinyama, Yusuke. 2010. "PDFMiner." <http://www.unixuser.org/~euske/python/pdf>.
- Sierra, Gerardo, Rodrigo Alarcón, César Aguilar, and Carme Bach. 2008. "Definitional Verbal Patterns for Semantic Relation Extraction." Edited by Alain Auger and Caroline Barrière. *Terminology (Special Issue) 14 (1)*: 74–98.
- Sinclair, John. 1987a. *Collins COBUILD English Language Dictionary*. 1st ed. London-Glasgow: Collins ELT.
- Sinclair, John.. 1987b. *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*. London-Glasgow: Collins.
- Smailović, Jasmina, and Senja Pollak. 2011. "Semi-Automated Construction of a Topic Ontology from Research Papers in the Domain of Language Technologies." In *Proceedings of the 5th Language & Technology Conference*, 121–125. Poznan, Poland.
- Smailović, Jasmina, and Senja Pollak.. 2012. "Topic Ontology Construction from English and Slovene Language Technologies Corpora." In *Proceedings of the 8th Language Technologies Conference, Proceedings of the 15th International Multiconference Information Society (IS 2012)*. Ljubljana, Slovenia.
- Snow, Rion, Dan Jurafsky, and Andrew Y. Ng. 2004. "Learning Syntactic Patterns for Automatic Hypernym Discovery." In *Proceedings of Advances in Neural Information Processing Systems*, 1297–1304.
- Snow, Rion, Daniel Jurafsky, and Andrew Y. Ng. 2006. "Semantic Taxonomy Induction from Heterogenous Evidence." In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 801–808.
- Stabej, Marko, Vojko Gorjanc, Simon Krek, Špela Arhar Holdt, Jasna Hočevar, Urška Jarnovič, Amanda Saksida, et al. 2006. "FidaPLUS: Korpus Slovenskega Jezika." <http://www.fidaplus.net/>. Last accessed January 15, 2012.
- Storrer, Angelika Wellinghoff, Sandra. 2006. "Automated Detection and Annotation of Term Definitions in German Text Corpora." In *Proceedings of the Fifth International Language Resources and Evaluation Conference (LREC 2006)*. Genoa, Italy.

- Strehlow, R A, and Sue Ellen Wright. 1993. *Standardizing Terminology for Better Communication: Practice, Applied Theory, and Results*.
- Svensen, Bo. 1993. *Practical Lexicography: Principles and Methods Of Dictionary Making*. Oxford University Press.
- Swartz, Norman. 2010. "Definitions, Dictionaries, and Meanings." *Lectures and Class Notes of N. Swartz*. Department of Philosophy, Simon Fraser University.
- TEI P5. 2007. "TEI P5: Guidelines for Electronic Text Encoding and Interchange." *Text Encoding Initiative Consortium*. <http://www.tei-c.org/Guidelines/P5/>.
- TEI P5. 2013. "TEI P5: Guidelines for Electronic Text Encoding and Interchange (edited by Lou Burnard and Syd Bauman)." *Text Encoding Initiative Consortium (version 2.5.0, Last Updated on 26th July 2013)*. <http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>.
- Velardi, Paola, Stefano Faralli, and Roberto Navigli. 2013. "OntoLearn Reloaded" A Graph-Based Algorithm for Taxonomy Induction." *Computational Linguistics* 39 (3): 665–707.
- Velardi, Paola, Roberto Navigli, and Pierluigi D'Amadio. 2008. "Mining the Web to Create Specialized Glossaries." *IEEE Intelligent Systems* 23 (5): 18–25.
- Viera, Anthony J., and Joanne M. Garrett. 2005. "Understanding Interobserver Agreement: The Kappa Statistic." *Family Medicine* 37 (5): 360–363.
- Vintar, Špela. 2002. "Avtomatsko Luščenje Izrazja Iz Slovensko-Angleških Vzporednih Besedil." In *Jezikovne Tehnologije*: Zbornik Konference: Proceedings of the Conference., edited by Tomaž Erjavec and Jerneja Žganec, Gros, 78–85. Ljubljana: Institut "Jožef Stefan."
- Vintar, Špela. 2008. *Terminologija: Terminološka Veda in Računalniško Podprta Terminografija*. Ljubljana: Filozofska fakulteta.
- Vintar, Špela. 2010. "Bilingual Term Recognition Revisited. The Bag-of-Equivalents Term Alignment Approach and Its Evaluation." *Terminology* 16 (2): 141–158.
- Vintar, Špela, and Darja Fišer. 2009. "Adding multi-word expressions to sloWNet." In *Proceedings of the 12th International Multiconference Information Society 2009, (Mondilex Fifth Open Workshop)*. Ljubljana, Slovenia.
- Vintar, Špela, and Darja Fišer. 2013. "Enriching Slovene WordNet with Domain-Specific Terms." *Translation: Computation, Corpora, Cognition* 1 (1): 29–44.
- Vogel, Adam, and Dan Jurafsky. 2012. "He Said, She Said: Gender in the ACL Anthology." In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, 33–41. Jeju Island, Korea: Association for Computational Linguistics.

- Walter, Stephan. 2008. "Linguistic Description and Automatic Extraction of Definitions from German Court Decisions." In *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC 2008)*, 2926–2932. Marrakech, Morocco.
- Walter, Stephan, and Manfred Pinkal. 2006. "Automatic Extraction of Definitions from German Court Decisions." In *Proceedings of the ACL'06 Workshop on Information Extraction Beyond the Document*, 20–26. Sydney, Australia.
- Warrens, Matthijs J. 2010. "Inequalities Between Multi-Rater Kappas." *Advances in Data Analysis and Classification* 4 (4): 271–286.
- Weinreich, U. 1967. "Lexicographic Definition in Descriptive Semantics." In *Problems in Lexicography*, edited by Householder and Saporta, 2nd ed., 25–43. Bloomington, IN: Indiana University Press.
- Westerhout, Eline. 2009. "Definition Extraction Using Linguistic and Structural Features." In *Proceedings of the 1st International Workshop on Definition Extraction (RANLP-09)*, 61–67. Borovets, Bulgaria.
- Westerhout, Eline. 2010. "Definition Extraction for Glossary Creation□: a Study on Extracting Definitions for Semi-Automatic Glossary Creation in Dutch." Lot Dissertation Series 252.
- Westerhout, Eline, and Paola Monachesi. 2007. "Extraction of Dutch Definitory Contexts for e-Learning Purposes." In *Proceedings of the Computational Linguistics in the Netherlands*.
- Wikipedia. 2013. "Enumerative definition." [http://en.wikipedia.org/wiki/Enumerative\\_definition](http://en.wikipedia.org/wiki/Enumerative_definition). Last accessed: May 27, 2013.
- Wright, Sue Ellen. 2011. "Terminography (explanation Note 2.2.4)." *ISO 1087-1:2000, 3.6.2*. <http://www.isocat.org/rest/dc/4092> available through Clarin ISocat web interface (<https://catalog.clarin.eu/isocat/interface/index.html>).
- Wüster, Eugen. 1979. *Einführung in Die Allgemeine Terminologielehre Und Terminologische Lexikographie*. UNESCO ALSED LSP Network.
- Yang, Hui, and Jamie Callan. 2009. "A Metric-Based Framework for Automatic Taxonomy Induction." *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*: 271–279.
- Zgusta, Ladislav. 1971. *Manual of Lexicography*. Berlin, New York: De Gruyter Mouton.
- Zhang, Chunxia, and Jiang Peng. 2009. "Automatic Extraction of Definitions." In *2nd IEEE International Conference on Computer Science and Information Technology (ICCSIT 2009)*, 364–368. Beijing, China: I

## 9 List of figures

Figure 1. Slovene part of the main Language technologies corpus by text type. ....	36
Figure 2. English part of the main Language technologies corpus by text type.....	36
Figure 3. English topic ontology (LTC proceedings corpus) without cleaning the documents and without renaming concepts. ....	39
Figure 4. English topic ontology (LTC proceedings corpus) after manually renaming the concepts, using active learning for adding concepts and manually moving some documents from one concept to another. ....	39
Figure 5. Slovene topic ontology (LTC proceedings corpus) after manually renaming the concepts, using active learning for adding concepts and manually moving some documents from one concept to another. ....	40
Figure 6. Concept visualization of English text documents (LTC proceedings corpus). The automatic splitting into two main topics, which we label computational linguistics and speech technologies, can be observed. ....	41
Figure 7. Concept visualization from Slovene text documents (LTC proceedings corpus). The automatic splitting into two main topics that we label Računalniško jezikoslovje and Govorne tehnologije can be observed. ....	41
Figure 8. Topic ontology constructed from the entire Slovene LT corpus.....	42
Figure 9. Topic ontology constructed from the entire English LT corpus. ....	43
Figure 10. Definition extraction methodology overview.....	58
Figure 11. A screenshot of the ClowdFlows workflow editor in the Google Chrome browser. ....	64
Figure 12: Definition extraction workflow in ClowdFlows, available online at <a href="http://clowdflows.org/workflow/1380">http://clowdflows.org/workflow/1380</a> .....	133
Figure 13. A screenshot of the ToTrTaLe workflow in ClowdFlows, available online at <a href="http://clowdflows.org/workflow/228/">http://clowdflows.org/workflow/228/</a> .....	136
Figure 14. Illustrating the term candidate viewer widget functionality. ....	139
Figure 15. Illustrating the definition candidate viewer widget functionality. ....	139



## 10 List of tables

Table 1. Counts of the Language Technologies Corpus in term of sentences and word tokens. ....	33
Table 2. Counts (# of articles) of the Slovene small corpus covering the Proceedings of the Language Technologies conference. For each year the information about the number of articles and other included units is provided. ....	35
Table 3. Counts (# of articles) of the English small corpus covering the Proceedings of the Language Technologies conference. For each year the information about the number of articles and other included units is provided. ....	35
Table 4. A sample output of ToTrTaLe, annotating sentences and tokens, with lemmas and MSD tags on words.....	62
Table 5. Top ranked 20 terms from the Slovene Language Technologies Corpus.....	67
Table 6. Top ranked 20 terms from the English Language Technologies Corpus. ....	67
Table 7. Precision of the reimplemented LUIZ term extraction method (on Slovene and English Language Technologies Corpus), evaluated by two annotators (the average denotes that the arithmetic mean of the two scores is above 2 in the first row and 5 in the second). The IAA scores for overall agreement and unweighted kappa are given in the last two rows and show low IAA agreement. ....	68
Table 8. Recall of terminological candidates extracted from the LT Corpus.....	69
Table 9. Evaluation of precision and recall for different variations of “X_is_Y” type of patterns on the Slovene corpus. ....	73
Table 10. Precision and recall of individual patterns on the Slovene data set. Second column contains the translated pattern in English, the third column indicates the number of extracted sentences with a number of true definitions between the parentheses. The fourth column gives the precision on all the extracted sentences and the last one the recall estimate, calculated on the 150 definitions dataset with the number of elements extracted from this dataset between the parentheses. ....	78
Table 11. Selection of settings of term-based methods from Table 23 and Table 24. (A: basic; B: highest recall; ee: highest precision; R: best precision-recall tradeoff (settings without nominative); w: best precision-recall tradeoff (settings with nominative); R&w: union of w and R, set as a suggested combination. (For the less restrictive settings, we evaluated the precision on 1,000 randomly selected definition candidates (sign ♣), the others were evaluated in totality.).....	87
Table 12. sloWNet-based definition extraction results.....	96

Table 13. Pattern-based definition extraction for English with the beginning of the sentence condition. Precision is evaluated on all the extracted sentences, while we used 150 definition test set for evaluating the recall. ....	101
Table 14. Pattern-based definition extraction for English without the beginning of the sentence condition.....	101
Table 15. Pattern-based definition extraction for English with the beginning of the sentence condition – extended starting patterns. Precision is evaluated on all the extracted sentences, while we used 150 definition test set for evaluating the recall.....	103
Table 16. Term-based definition extraction on the English part of the corpus. ....	110
Table 17. WordNet-based definition extraction candidates .....	114
Table 18. Summary of definition extraction methods on the Slovene subcorpus. For the term-based approach a combined setting of two methods (R) of Table 23 and (w) of Table 24 is taken, the same as selected as the most appropriate term-based setting shown in Table 11. <i>Union</i> denotes the sentences extracted by any of the three methods, while <i>Intersection</i> denotes the sentences extracted by at least two out of three methods.....	119
Table 19. Summary of definition extraction methods on the Slovene subcorpus, presenting the results of a selection of combined methods.....	120
Table 20. Summary of results of three definition extraction methods on the English subcorpus, as well as <i>Union</i> (i.e., sentences extracted by any of the methods) and <i>Intersection</i> (i.e., sentences extracted by at least two out of three methods).....	121
Table 21. Summary of different combinations of definition extraction methods on the English subcorpus. As the basis three variants of the pattern-based approach are taken that are filtered or enlarged in different ways by term-based or WordNet-based definition extraction.....	122
Table 22. Tags and examples of different types of definition candidates.....	130
Table 23. Evaluation of the term-based definition extraction approach – settings without the nominative condition. For the less restrictive settings, we evaluated the precision on 1,000 randomly selected definition candidates (sign ♣), the others were evaluated in totality. The recall was evaluated on the preselected 150 definitions dataset.....	168
Table 24. Evaluation of the term-based definition extraction approach – settings with the nominative condition. For the less restrictive settings, we evaluated the precision on 1,000 randomly selected definition candidates (sign ♣), the others were evaluated in totality. The recall was evaluated on the preselected 150 definitions dataset .....	169

## APPENDIX A: Term-based definition extraction experiments

To check the combinations of parameters that perform well, we evaluated the precision and recall for each parameter setting. The evaluation is performed by inspecting the results of experiments presented in Table 23 and Table 24 and when in the text we refer to different settings, we refer to experiments (A) to (FF) from Table 23 for experiments without the nominative conditions, and to experiments (a) to (ff) from Table 24 when the nominative condition is applied. To evaluate the precision, in the majority of experiments all the sentences were inspected for correctness, enabling us to analyze a large number of different definitions and detect preliminary entries for the glossary. For the less restrictive settings, we provide only an estimate of precision by evaluating a subset of 1,000 randomly selected elements (precision scores marked with ♣). The estimated recall was evaluated on the 150 definitions test set, while another estimation of the recall can be seen from the number of actually extracted definitions (in *precision* column): the more definitions extracted, the better the recall. In the rest of this section, we verify the seven hypotheses from Section 5.1.2 examining the influence of each parameter on the precision/recall of the system.

### Verifying Hypothesis 1

Firstly, we can see that the precision is higher when the term threshold parameter is set higher (i.e., selecting only top 1% of extracted terms ranked by termhood value vs. larger selection of top 10%). The first four experiments with basic settings, i.e., using only the condition that a sentence should contain two terms (A) and (B) and the second two where also the verb condition is applied (C) and (D), clearly shows the influence of the threshold value on precision and recall. Note that when identifying the terms in a sentence, nested terms are not counted separately but only the longest term above the threshold is considered, e.g., in term *computational linguistics*, *computational linguistics* and *linguistics* are both terms, but only *computational linguistics* is counted as a term. In (A) and (B) the estimated precision is 0.052 with 1% of terms, compared to 0.030 with 10% of terms. In (C) and (D), the estimated precision increases from 0.039 to 0.047 when only 1% instead of 10% of terms is used. However in both cases, recall decreases with the higher number of terms taken into consideration. Similarly, the percentage of terms influences the results in all other settings. Examine for example the results in (H) to (T) and (d) to (p) where we can see that changing the threshold and keeping all other settings the same has a big influence on precision. For example, have a look at Experiments (S) and (T) where the precision without applying nominative condition reaches up to 0.172 threshold, when there are at least 5 domain terms in a sentence candidate and we take into account top 1% of domain terms, while the same setting with 10% of domain terms extracts the candidates with precision 0.1101, but higher precision means less definitions extracted (75 for 1% and 222 for 10%). When the nominative condition is applied (see e.g., (o) and (p)) we get 0.202 precision (40 definitions) for 1% of terms and 0.1292 (160 definitions) for 10%, with 5 domain terms in a sentence. The hypothesis is confirmed also in all other experiments. An additional

test was performed with 5% of terms and as expected the precision results are higher than for 10% and lower than for 1% of terms, and the opposite is true for the number of extracted definitions and estimated recall (see for instance (CC), (DD), (EE), as well as (aa), (cc), (dd)).

### **Verifying Hypothesis 2**

Experiments were made for checking whether setting the number of terms in the sentence to more than 2 (the default) could increase the performance. This can be seen in the evaluation of precision in Experiments (H), (K), (O) and (S), in which for 1% threshold the precision constantly improves with the higher number of terms set from  $n=2$  (precision 0.12 in (H) to  $n=5$  (precision 0.172 in (S)). It is also true for 10% in the experiments with the settings from  $n=3$  to  $n=5$  (i.e., Experiments (M), (P), (T) but not for  $n=2$ , which has the same precision as  $n=4$ , but one can observe that the evaluation of the precision for this setting is only an estimate on 1,000 randomly selected sentences. When using the nominative condition, the hypothesis that the precision increases with a higher number of domain terms holds true for the setting with ‘verb between any terms’ (VA)—explained below in verifying hypothesis 3—(Experiments (d), (f), (j), (n)), but when the verb condition is used in the ‘verb-first’ (VF) variant it is true for 10% ((e), (h), (l), (p)) but not in all of the 1% threshold experiments. Even if there are few exceptions to the main observation that the precision increases with more domain terms in a sentence, in all the experiments precision results of the lowest setting ( $n=2$ ) is worse than the results of the highest setting ( $n=5$ ).

### **Verifying Hypothesis 3**

Next, the simple condition that a verb should appear between two domain terms is tested in several experiments. In the first two settings, basic ones ((A) and (B)) and the one with the verb condition ((C) and (D)), we can see that the experiments considering 10% of terms confirm the hypothesis that the verb condition improves the precision (precision in Experiment (D), outperforms the basic settings in (B) (precision 0.039 and 0.030, respectively) but not the 1% setting (results of Experiment (C) compared to basic settings in (A)). Note that these are estimates only (sign ♣), evaluated on 1,000 randomly selected sentences form the entire set of definition candidates extracted by the method. Therefore we tested the hypothesis also in other settings, i.e., (G) and (H); (Q) and (R); (Y) and (Z); (c) and (d); (m) and (n); (u) and (v), where all the examples confirm the hypothesis of verb condition providing better precision results. Note that in the experiments mentioned just before, the verb condition does not affect the recall.

Another verb-related hypothesis was tested: if we consider more than two terms, how does the settings where a verb occurs between the first two terms (VF) and the one where a verb appears between any terms (VA) influence the results in terms of precision and recall. In (J) and (K) we get higher precision (0.1381) when a verb condition relates to a verb between the first two terms vs. 0.1444 when the broader condition is considered. The estimated recall on the recall test set does not show any difference, but the number of extracted definitions shows that fewer definitions are extracted with a stricter condition (157 when a verb is between the first two terms compared to 166 when a verb occurs between any terms). Similar tendencies can be observed in Experiments (L) and (M); (N) and (O); (Z) and (AA); (f) and (g); (h) and (i); (j) and (k); (v) and (x); (aa) and (ee). On the other hand, limiting the verb to the place between the first two terms only does not improve the precision in Experiments (R) and (S); (CC)

and (FF); (n) and (o), all three using 5 domain terms condition (explained above), where the precision is higher when a verb can occur between any terms. On the other hand, also Experiments (aa) and (ee) use the 5 domain terms condition and the VF setting performs better in terms of precision than VA. Note also that in the majority of experiments, the increase in precision is quite small.

#### **Verifying Hypotheses 4 and 5**

The constraints of having a domain term at the beginning of the sentence and that the first appearing domain term should be a multi-word expression and not a single-word term, yielded better precision results, especially when applied together. Compared to a simple setting with 2 domain terms and the verb condition (cf. Experiment (E) with 0.047 estimated precision, the precision increases when we add the two conditions ('beginning of a sentence' and 'multi-word term first') and obtain 0.12 precision (cf. Experiment (H)). Experiment (G) compared to (H) confirms once again, that it is better to apply also the 'verb condition'. Precision increases also when the two tested parameter settings are applied individually (cf. (E) for the 'beginning of the sentence') and (F) for 'multi-word first'). However, we can see that the cost for higher precision in a large decrease in (estimated) recall (cf. 0.8933 in (D), compared to 0.08 in (H) meaning that we must know which setting to use when we want to optimize precision and which one for a better recall. If we take a look at examples with nominative constraints, we see similar results, i.e., in terms of precision, the best results are obtained if all three constraints are applied, namely verb condition, beginning of a sentence condition and multi-word term preceding other terms condition (0.1858 in (d)) compared to precision between 0.083 and 0.1653 if one of the three conditions is missing ((a)–(c)). On the other hand, the estimated recall and the number of extracted definitions importantly decrease with the beginning of a sentence and multi-word term conditions (e.g., 258 definitions and 0.2133 recall when beginning of a sentence condition is not applied (a), compared to 84 definitions and 0.0333 recall with all three constraints used together with the nominative condition (d).

#### **Verifying Hypothesis 6**

The next hypothesis was that multi-word terms bear higher terminological value than simple words and thus that setting the number of multi-word terms in a sentence higher should yield better precision. The hypothesis is confirmed, since all the experiments prove it (see for example pairs of experiments (H) and (U); (K) and (V); (O) and (AA); (S) and (FF). In more detail, for 2 domain terms with termhood value set at 1% of domain terms and applying verb, beginning and multi-word first conditions, the precision raises from 0.12 (H) to 0.1527 when the extra condition that both terms should be multi-word terms is applied (Experiment (U)). Similar observation can be made with 5 domain terms (cf. Experiments (S) and (R) with precision scores 0.172 and 0.1751, respectively, depending on which type of verb condition is used (VF/VA)), where much higher precision can be reached if at least 3 out of 5 terms are multi-word expressions (precision is 0.2009 with the 'verb-first' condition (Experiment (FF) and 0.2111 with 'verb-anywhere condition' (Experiment (CC)). In experiments with the nominative condition it also holds true that precision increases with the number of multi-word terms. E.g., for minimum number of terms set to 2, precision increases from 0.1858 (Experiment (d)) to 0.2236 (Experiment (q)) when the two terms are multi-word expression; for minimum number of terms set to 5, the precision increases from 0.202

Settings – without nominative constraints								Results		
	Threshold (1)	Terms (#) (2)	Verb (VA-VF) (3)	Beginning sent. (4)	Multi-word first (5)	Multi-word (#) (6)	Nominatives (#) (7)	Extracted sentences (#)	Precision (and # of definitions)	Recall on 150 test set (# of definitions)
Basic										
A.	1%	2	no	no	no	no	no	28,215	0.052♣	0.8533 (128)
B.	10%	2	no	no	no	no	no	35,624	0.030♣	0.9733 (146)
Verb										
C.	1%	2	yes	no	no	no	no	22,176	0.047♣	0.7067 (106)
D.	10%	2	yes	no	no	no	no	29,840	0.039♣	0.8933 (134)
Beginning of the sentence										
E.	1%	2	yes	yes	no	no	no	8,548	0.065♣	0.2733 (41)
First multiterm										
F.	1%	2	yes	no	yes	no	no	7,958	0.075♣	0.34 (51)
Beginning and First multiword term										
G.	1%	2	no	yes	yes	no	no	1,715	0.1044 (179)	0.08 (12)
H.	1%	2	yes	yes	yes	no	no	1,492	0.12 (179)	0.08 (12)
I.	10%	2	yes	yes	yes	no	no	4,486	0.095♣	0.1667 (25)
J.	1%	3	yes-VA	yes	yes	no	no	1,202	0.1381(166)	0.0667 (10)
K.	1%	3	yes-VF	yes	yes	no	no	1,087	0.1444(157)	0.0667 (10)
L.	10%	3	yes-VA	yes	yes	no	no	4,010	0.0828 (332)	0.1667 (25)
M.	10%	3	yes-VF	yes	yes	no	no	3,671	0.088(323)	0.1667 (25)
N.	1%	4	yes-VA	yes	yes	no	no	893	0.1522 (136)	0.0533 (8)
O.	1%	4	yes-VF	yes	yes	no	no	712	0.1587 (113)	0.0533 (8)
P.	10%	4	yes-VF	yes	yes	no	no	2,797	0.095 (266)	0.1467 (22)
Q.	1%	5	no	yes	yes	no	no	619	0.168 (104)	0.0467 (7)
R.	1%	5	yes-VA	yes	yes	no	no	594	0.1751 (104)	0.0467 (7)
S.	1%	5	yes-VF	yes	yes	no	no	436	0.172 (75)	0.0333 (5)
T.	10%	5	yes-VF	yes	yes	no	no	2,016	0.1101 (222)	0.1267 (19)
Number of multiword terms										
U.	1%	2	yes	yes	yes	2	no	825	0.1527(126)	0.0533 (8)
V.	1%	3	yes-VF	yes	yes	2	no	696	0.1667 (116)	0.0467 (7)
W.	10%	3	yes-VF	yes	yes	2	no	2,971	0.0966 (287)	0.16 (24)
X.	10%	3	yes-VF	yes	yes	3	no	1,954	0.111 (217)	0.0933 (14)
Y.	1%	4	no	yes	yes	2	no	670	0.1642 (110)	0.04 (6)
Z.	1%	4	yes-VA	yes	yes	2	no	636	0.1729 (110)	0.04 (6)
AA.	1%	4	yes-VF	yes	yes	2	no	508	0.1772 (90)	0.04 (6)
BB.	10%	4	yes-VF	yes	yes	2	no	2,397	0.1026 (246)	0.14 (21)
CC.	1%	5	yes-VA	yes	yes	3	no	289	0.2111 (61)	0.02 (3)
DD.	5%	5	yes-VA	yes	yes	3	no	1,078	0.1317 (142)	0.06 (9)
EE.	10%	5	yes-VA	yes	yes	3	no	1,707	0.1125 (192)	0.0867(13)
FF.	1%	5	yes-VF	yes	yes	3	no	214	0.2009 (43)	0.0067(1)

Table 23. Evaluation of the term-based definition extraction approach – settings without the nominative condition. For the less restrictive settings, we evaluated the precision on 1,000 randomly selected definition candidates (sign ♣), the others were evaluated in totality. The recall was evaluated on the preselected 150 definitions dataset.

	Settings – with nominative constraints							Results		
	Threshold (1)	Terms (#) (2)	Verb (VA-VF) (3)	Beginning sent. (4)	Multi-word first (5)	Multi-word (#) (6)	Nominatives (#) (7)	Extracted sentences (#)	Precision (and # of definitions)	Recall on 150 test set (# of definitions)
a)	1%	2	yes	no	yes	no	2	2,377	0.1085 (258)	0.2133 (32)
b)	1%	2	yes	yes	no	no	2	2,787	0.083♣	0.1333 (20)
c)	1%	2	no	yes	yes	no	2	508	0.1653 (84)	0.0333 (5)
d)	1%	2	yes	yes	yes	no	2	452	0.1858 (84)	0.0333 (5)
e)	10%	2	yes	yes	yes	no	2	1968	0.1113(219)	0.1133 (17)
f)	1%	3	yes-VA	yes	yes	no	2	438	0.1918 (84)	0.0333(5)
g)	1%	3	yes-VF	yes	yes	no	2	406	0.2020 (82)	0.0333 (5)
h)	10%	3	yes-VA	yes	yes	no	2	1,917	0.1142 (219)	0.1133 (17)
i)	10%	3	yes-VF	yes	yes	no	2	1,820	0.117 (213)	0.1133 (17)
j)	1%	4	yes-VA	yes	yes	no	2	371	0.1995 (74)	0.0267 (4)
k)	1%	4	yes-VF	yes	yes	no	2	287	0.2021 (58)	0.0267 (4)
l)	10%	4	yes-VF	yes	yes	no	2	1544	0.1198 (185)	0.1 (15)
m)	1%	5	no	yes	yes	no	2	299	0.1973 (59)	0.02 (3)
n)	1%	5	yes-VA	yes	yes	no	2	280	0.2107 (59)	0.02 (3)
o)	1%	5	yes-VF	yes	yes	no	2	198	0.202 (40)	0.0133 (2)
p)	10%	5	yes-VF	yes	yes	no	2	1,238	0.1292 (160)	0.0933 (14)
Multiword										
q)	1%	2	yes	yes	yes	2	2	313	0.2236 (70)	0.0267 (4)
r)	1%	3	yes-VF	yes	yes	2	2	287	0.2369 (68)	0.0267 (4)
s)	10%	3	yes-VF	yes	yes	2	2	1,572	0.1259 (198)	0.1067 (16)
t)	10%	3	yes-VF	yes	yes	3	2	1,144	0.1381 (158)	0.06 (9)
u)	1%	4	no	yes	yes	2	2	300	0.2133 (64)	0.02 (3)
v)	1%	4	yes-VA	yes	yes	2	2	279	0.2294 (64)	0.02 (3)
w)	1%	4	yes-VA	yes	yes	2	1	499	0.1944 (97)	0.0333 (5)
x)	1%	4	yes-VF	yes	yes	2	2	216	0.2315 (50)	0.02 (3)
y)	1%	4	yes-VF	yes	yes	2	1	395	0.2051 (81)	0.0333 (5)
z)	10%	4	yes-VF	yes	yes	2	2	1370	0.1285 (176)	0.0933 (14)
aa)	1%	5	yes-VA	yes	yes	3	2	141	<b>0.2624 (37)</b>	0.0133 (2)
bb)	1%	5	yes-VA	yes	yes	3	1	228	0.2281 (52)	0.02 (3)
cc)	5%	5	yes-VA	yes	yes	3	2	672	0.1518 (102)	0.0467 (7)
dd)	10%	5	yes-VA	yes	yes	3	2	1,078	0.1354 (146)	0.06 (9)
ee)	1%	5	yes-VF	yes	yes	3	2	102	<b>0.2647(27)</b>	0.0067 (1)
ff)	1%	5	yes-VF	yes	yes	3	1	169	0.2248 (38)	0.0067 (1)

Table 24. Evaluation of the term-based definition extraction approach – settings with the nominative condition. For the less restrictive settings, we evaluated the precision on 1,000 randomly selected definition candidates (sign ♣), the others were evaluated in totality. The recall was evaluated on the preselected 150 definitions dataset

(Experiment (o)) to 0.2647 (Experiment (ee)) or from 0.2107 Experiment (n) to 0.2624 Experiment (aa) if at least 3 terms are multi-word expressions.

### Verifying Hypothesis 7

In the last set of experiments, we checked the influence of the condition requiring the terms to be in the nominative case. This is applicable only to Slovene. We can see that the nominative case condition leads to higher precision. Take for example experiments with 2 domain terms: Experiment (H) at 1% threshold and Experiment (I) at 10% threshold that yield 0.12 and 0.095 precision scores, respectively. If we compare them to experiments with similar settings but with the nominative condition added (2 terms should be in the nominative case), precision increases importantly: 0.1858 precision in Experiment (d) and 0.1113 in (e). One of the highest precision results in our experiments is obtained with 5 terms and nominative condition (2 terms in nominative case): 0.2107 (Experiment n)) which is higher than when the nominative condition is not applied (0.1751 in (R)), however much smaller number of definitions is extracted.

To conclude this table explanation we provide the overall best precision results, which is achieved if we apply all the above-mentioned constraints (i.e., the verb condition, the beginning of the sentence condition; the multi-word term preceding other terms; the higher number of multi-word terms and the nominative conditions): the precision gets above 26% if 5 domain terms out of which 3 should be multi-word expressions and 2 terms in nominative are used (cf. Experiments (aa) and (ee)), compared to precision between 0.20 and 0.21 in Experiments (CC) and (FF), which are (in terms of precision) the best performing settings without the nominative condition. However, the number of extracted definitions and the estimated recall with these restrictive settings are very low. If we loosen the nominative case condition from two terms in the nominative case to only one nominative term, the precision is lower but the estimated recall and the number of extracted definitions are higher. See Examples (w), (y), (bb) and (ff), based on which this conclusion is made. Compared to the best settings with two nominatives, the precision for best setting decreases from 0.2315 (x) to 0.2051 (y) for 4 terms and from 0.2647 (ee) to 0.2248 for 5 terms (ff). On the other hand more definitions are extracted, i.e., 81 instead of 50 and 38 instead of 27, respectively.

In summary, a general trend is that the higher the termhood value<sup>56</sup> and the number of nominatives in the sentence, the higher the precision and the lower the recall. Moreover, the more terms and multi-word terms in a sentence, the better the precision. In addition, other constraints, such as verb between two terms, having a term at the beginning of a sentence and a multi-word term as the first domain term term improve the results. Based on the objective of the application, the user can choose to tune the approach for higher precision or recall by selecting different parameter settings.

---

<sup>56</sup> Terms extracted by the term extraction method are ranked by their termhood value, meaning that if e.g., 1% of terms are used, the termhood value is higher than if 2% of all extracted terms are used, etc.

## Razširjeni povzetek

Človeško znanje je dostopno v strokovnih besedilih, terminoloških slovarjih in enciklopedijah, v zadnjem času pa tudi v računalniku razumljivih predstavitev 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, ontologij in tematskih ontologij. Luščenji terminologije in definicij sta pomembna koraka modeliranja strokovnega znanja, vendar so razvite metode in orodja večinoma prilagojena za posamezne jezike, a le redko za manj razširjene jezike, kot je slovenščina. Zato je glavni doprinos doktorske disertacije, ki ponuja metodologijo za luščenje definicijskih stavkov iz korpusov v slovenskem in angleškem jeziku, prav luščenje definicij iz slovenskih nestrukturiranih besedil.

V **uvodnem poglavju** predstavimo glavne cilje doktorske disertacije in prispevke k znanosti. 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. Osrednji cilj doktorske disertacije je iz razpoložljivih besedil polavtomatsko izluščiti model domene v obliki strokovnega izrazja in definicij. V ta namen predlagamo novo metodologijo luščenja definicij za slovenščino in angleščino in njeno implementacijo v obliki spletno dostopnega delotoka (angl. *workflow*). Kot obravnavano področje smo si izbrali področje jezikovnih tehnologij. Poleg glavnega doprinosa v obliki razvite metodologije luščenja definicij in njene implementacije (pri čemer še posebej poudarimo luščenje definicij iz slovenskih besedil, saj je to za razliko od angleščine še neraziskano področje) je v doktorskem delu predstavljenih še nekaj drugih pomembnih prispevkov k znanosti.

Za namene modeliranja izbranega področja smo zgradili primerljiv slovensko-angleški *Korpus jezikovnih tehnologij*. V slovenskega smo vključili vse članke, predstavljene na konferenci Jezikovne tehnologije do vključno leta 2010, ter ga dopolnili z drugimi tipi besedil, kot so diplomske, magistrske in doktorske naloge, poglavja iz knjig in članki. Za angleščino smo zgradili slovenskemu delu primerljiv korpus. Slovenski del korpusa, ki zajema konferenčne zbornike, je dostopen prek konkordančnika na naslovu [http://nl.ijs.si:3003/cuwi/sdjt\\_sl](http://nl.ijs.si:3003/cuwi/sdjt_sl).

Celotno metodologijo smo strnili v prosto dostopen delotok, implementiran v spletnem okolju za gradnjo delotokov Clowdflovs (Kranjc et al., 2012), ki je dostopen na naslovu: <http://www.clowdflovs.org/workflow/1380/>. 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.

Poleg osrednjega spletnega servisa za luščenje definicijskih stavkov, ki ga v delotoku sestavljajo trije glavni gradniki, sta za nadaljnjo rabo pomembni tudi novi implementaciji že obstoječih orodij v okolju Clowdflovs: implementacija slovenskega

in angleškega jezikoslovnega označevalnika ToTrTaLe (Erjavec et al., 2010), ki smo ga s soavtorji implementirali tudi kot samostojen delotok (Pollak et al., 2012c) in je dostopen na povezavi <http://clowdflows.org/workflow/228/>, ter implementacija slovenskega in angleškega luščilnika terminologije LUIZ (Vintar, 2010), ki je dostopen kot gradnik našega glavnega delotoka (Pollak et al., 2012a), po želji pa ga lahko vključimo tudi v druge delotoke.

Izluščene angleške in slovenske definicijske kandidate smo ocenili in razvrstili s pripisanimi oznakami v dve glavni kategoriji kategoriji – 'definicije' in 'ne-definicije', poleg tega smo podizbor stavkov označili z bolj podrobnimi kategorijami, ki povezujejo leksikografski pogled na definicije (npr. podkategorije, vezane na tip ali vsebino definicije) s problematiko avtomatskega luščenja definicij iz besedil (npr. oznaka za napačno segmentiran stavek). Eden izmed rezultatov doktorske disertacije je tudi prosto dostopni pilotni *Slovarček jezikovnih tehnologij* (dostopen na strani [http://kt.ijs.si/senja\\_pollak/jt\\_glosar/](http://kt.ijs.si/senja_pollak/jt_glosar/)), ki smo ga ročno izdelali na podlagi avtomatsko izluščenih definicijskih kandidatov.

V **drugem poglavju** podamo širši pregled sorodne literature. Področje predstavimo tako z jezikoslovne perspektive, kjer predstavimo predvsem leksikografski pogled na definicije, kot tudi iz jezikovnotehnološke perspektive modeliranja domene in luščenja področnega znanja, še posebej v obliki luščenja definicij.

Najprej se posvetimo vprašanju odnosa med *jezikom* in *pomenom* (oz. v saussurjevski terminologiji med označevalcem in označencem). Ko nekaj *izjavimo*, se nanašamo na konkretne ali abstraktne realnosti. Nekatero skupino označencev so si na podlagi njihovih razlikovalnih lastnosti med seboj zelo podobne (tvorijo isti *koncept*) in so zelo različne od ostalih. V komunikacijskih dejanjih konceptov ne opisujemo z njihovimi razlikovalnimi lastnostmi, temveč za njih uporabljamo označevalce, to so *besede oz. leksikalne enote*. Njihov pomen pa lahko opišemo z *definicijami*, ki so zbrane v *slovarjih* ali slovarjem podobnih zbirkah. Definicije so v leksikografski tradiciji razdeljene v različne definicijske tipe, slovarji pa so zavezani določenim leksikografskim načelom.

Odnos med označevalcem (besedo oz. leksikalno enoto) in označencem lahko razložimo s pojmom *leksikalnega pomena*, ki ga avtorji, kot sta Zgusta (1971) in Svendsen (1993), razlagajo z njegovimi tremi sestavinami: *denotat* zajema objektivni pomen, *konotacija* subjektivni oz. emotivni pomen, *obseg* (angl. *range of application*) pa omejuje veljavnost besede glede na nekatere lastnosti, vezane na slog, pomen ali na slovnično kategorijo. V drugem poglavju uvedemo razliko med *splošnim* in *strokovnim jezikom* ter med področji *leksikologije* in *terminologije* kot tudi *leksikografije* in *terminografije*.

Večji del drugega poglavja posvetimo obravnavi *definicij*. V filozofski literaturi avtorji ločijo več vrst definicij (Copi in Cohen, 2009; Parry in Hacker, 1991). *Leksikalne definicije* (angl. *lexical definitions*) se uporabljajo v slovarjih in razlagajo že uveljavljeni pomen definienduma. 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 (tj. definirani pojem) 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 zožajo, 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.

Nadalje se posvetimo različnim *definičijskim strategijam* in pogledamo, katere tipe leksikalnih definicij navaja obstoječa literatura. 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 *definienduma*, kar pomeni da navajajo vse možne oz. najbolj tipične realizacije definiranega pojma (gre torej za naštevanje vseh oz. tipičnih pripadajočih elementov razreda) (Copi in Cohen, 2009).

V nadaljevanju obravnavamo različne podtipe intenzionalnih in ekstenzionalnih definicij, ki jih omenja leksikografska literatura. Najprej se posvetimo *intenzionalnim definicijam*. Najbolj tipične – in po mnenju nekaterih avtorjev najbolj prestižne – so leksikografske definicije z obliko *genus et differentiae*. To so definicije, kjer je *definiendum* definiran z nadpomenko oz. najbližjim rodnom (*genus*) in vrstnimi razlikami (*differentiae specifica*) oz. vsaj eno bistveno značilnostjo, ki *definiendum* (oz. razred *definienduma*) ločuje od ostalih pripadnikov rodu (Svensen, 1993). Ker se pri definicijah *genus-differentiae* pomen *definiensa* analizira, se imenujejo tudi analitične definicije. 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* je *definiendum* definiran s svojo rabo, namenom oz. funkcijo. Še en podtip je definiranje s pomočjo *tipičnih lastnosti*, kar je običajno uporabljeno v kombinaciji z zgoraj omenjenimi analitičnimi ali funkcijskimi definicijami. *Operacijske definicije* pa definirajo *definiendum* z definiranjem specifičnih testov, ki so ponovljive operacije, ki vodijo vedno do enakih rezultatov.

Druga strategija za definiranje pojmov je z njeno *ekstenzijo*. 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 in Hacker, 1991; Copi in 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 literatura omenja še nekaj tipov ekstenzionalnih definicij, ki se ponavadi (a ne izključno) nanašajo na ekstenzionalne navedbene definicije. *Naštevalne definicije* (angl. *enumerative definitions*) so poseben podtip ekstenzionalnih definicij, v katerih

naštejemo vse predstavnike definirane 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*. Zadnji tip definicij pa so *kontekstualne definicije*, kjer pomen v resnici ni definiran v ožjem pomenu besede, temveč je impliciran in ga je treba razbrati iz konteksta, saj med definiendumom in definiensom ni jasne strukturne ločnice.

Podpoglavje se zaključi s kritično obravnavo leksikografskih principov tvorjenja dobrih definicij. Po teh načelih mora biti definiendum definiran z izrazi, ki so splošnejši od njega, izogibati se je treba krožnosti v definicijah in zbirkah, pri analitičnih definicijah se je potrebno osredotočiti na bistvene značilnosti, glede sloga pa se morajo definicije ogibati dvoumnega in metaforičnega izražanja, ter če je le možno, uporabljati trdilno obliko (prim. Jackson, 2002; Zgusta, 1971; Béjoint, 2000; Svensen, 1993).

V drugem delu drugega poglavja 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 Sciano in Velardi (2007), Ahmad et al. (2007), Frantzi in Ananiadou (1999), Kozakov et al. (2004) ter Vintar (2010) za slovenščino. Za dvojezično luščenje terminologije pa so na voljo komercialna (SDL MultiTerm, Similis) in nekomercialna (npr. Lefever et al., 2009; Macken et al., 2013; Vintar, 2010) orodja.

V specializiranih besedilih se poleg samih terminov skrivajo še drugi dragoceni deli znanja, med njimi tudi definicije, katerih luščenje predstavlja osrednjo temo pričujoče doktorske raziskave. Metode luščenja definicij so bile razvite za več jezikov, kot so angleščina (Navigli in Velardi, 2010; Borg et al., 2010), nizozemščina (Westerhout, 2010), francoščina (Malaisé et al., 2004), nemščina (Fahmi in Bouma, 2006; Storrer in Wellinghoff, 2006; Walter, 2008), kitajščina (Zhang in Jiang, 2009), portugalsščina (Del Gaudio in Branco, 2007; Del Gaudio et al., 2013), romunščina (Iftene et al., 2007), poljščina (Degórski et al., 2008a, 2008b) kot tudi za druge slovanske jezike (Przepiórkowski et al., 2007). Za slovenščino smo začeli razvijati metodologijo v Fišer et al. (2010) in Pollak et al. (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 in 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 (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 skladijskih in leksikalnih značilnosti (vzorcev). Takšno metodo je uporabil že Hearst (1992), a tudi v novejših raziskavah se različice metode z vzorci še vedno pojavljajo (npr. Muresan in Klavans, 2002; Walter in Pinkal 2006; Storrer in Wellinghoff, 2006; Del Gaudio in Branco, 2007). Tudi med pristopi, uporabljenimi v doktorski disertaciji, 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 naučiti pravil za razlikovanje med pravimi in nepravimi definicijami. Z običajnimi klasifikacijskimi algoritmi, kot so naivni Bayes, odločitvena drevesa in metoda podpornih vektorjev (SVM), je različnim avtorjem uspelo razlikovati med dobro in slabo oblikovanimi definicijami (Del Gaudio in Branco, 2009; Chang in Zheng, 2007; Velardi et al., 2008; Fahmi in Bouma, 2006; Westerhout, 2010; Kobyliński in Przepiórkowski, 2008; Del Gaudio et al., 2013), za popolnoma avtomatske pristope pa so uporabljeni tudi genetski algoritmi (Borg et al., 2010) ter mreže besednih vrst (Navigli in Velardi, 2010; Faralli in Navigli, 2013).

Poleg definicij in semantičnih relacij se raziskovalci že nekaj časa posvečajo tudi (pol)avtomatski izgradnji taksonomij in ontologij. Nekateri imajo za cilj razširiti že obstoječo ročno zgrajeno ontologijo, kot sta WordNet<sup>57</sup> (Fellbaum, 1998) ali Open Directory Project,<sup>58</sup> drugi pa želijo ustvariti ontologijo brez predloge. Zanimivi so pristopi Snow et al. (2006) za inkrementalno gradnjo taksonomij ter Yang in Callan (2009), ki z gručenjem v skupine (angl. *clustering*) za vsak par terminov v taksonomiji izračunata semantično razdaljo. Kozareva in Hovy (2010) uporabljata vzorce ter metode grafov. Navigli et al. (2011) in Velardi et al. (2013) najprej uporabijo svojo metodo za luščenje definicij in nadpomenk iz korpusov in spleta (Navigli in Velardi, 2010), nato pa iz grafa, pridobljenega iz vseh nadpomenk, izluščijo taksonomijo.

Na kratko predstavimo tudi modeliranje področja jezikovnih tehnologij. ACL Anthology (Kan in Bird, 2013) je digitalni arhiv, ki do danes zajema nad 24.000 člankov iz revij ter s konferenc s področja računalniškega jezikoslovja. Podmnožica teh člankov sestavlja referenčni korpus ACL ARC (Bird et al., 2008). Radev et al. (2009, 2013) so zgradili tudi mrežo ACL AAN, iz katere je razvidno, kdo citira koga, kateri avtorji sodelujejo itd. Na teh korpusih je bilo izvedenih več raziskav, predvsem na temo odkrivanja raziskovalnih področij (Hall et al., 2008; Paul in Girju, 2009; Anderson et al., 2012), pri čemer vsi avtorji uporabljajo latentno Dirichletovo alokacijo (Blei et al., 2003). Radev in Abu-Jbara (2012) na ACL AAN izpeljeta analizo citatov in pokažeta, da je le-ta uporabna za raziskovanje trendov v računalniškem jezikoslovju, summarizacijo ter vrsto drugih nalog. Podobno nalogo, kot smo si jo zastavili sami, obravnava Reiplinger s soavtorji (2012), ki z leksikoskladenjskimi vzorci ter globoko sintaktično analizo lušči kandidate za slovar iz angleškega korpusa ACL ARC.

V nadaljevanju predstavimo področje gradnje spletnih servisov in delotokov. Okolja za rudarjenje podatkov, ki omogočajo gradnjo in uporabo delotokov, so npr. Weka (Witten et al., 2011), Orange (Demšar et al., 2004), KNIME (Berthold et al., 2008) in Rapid-Miner (Mierswa et al., 2006). Njihova skupna lastnost je kanvas, v katerem uporabnik gradi delotoke s preprostim principom primi-odloži. Porazdeljeno procesiranje je uporabljeno v servisno orientiranih arhitekturah, kot sta Orange4WS (Podpečan et al., 2012) in Taverna (Hull et al., 2006). Orodje Taverna (Hull et al., 2006) omogoča, da so delotoki dostopni vsem, saj jih lahko avtorji naložijo in naredijo dostopne prek spletne povezave. ClowdFlows (Kranjc et al., 2012) je aplikacija v oblaku, ki omogoča, da brez namestitev katerih koli programov dostopamo do že zgrajenih delotokov ali gradimo nove delotoke iz poljubnega brskalnika.

V **tretjem poglavju** si natančneje zastavimo cilj disertacije, ki je iz besedil

---

<sup>57</sup> Besedo wordnet pišemo z malo začetnico, kadar se nanašamo na tip leksikalnih zbirk, ki s svojo strukturo in leksiko sledijo načelom prvega tovrstnega projekta WordNet z Univerze v Princetonu, za katerega uporabljamo veliko začetnico (pri tej odločitvi se zgledujemo po Fišer, 2007).

<sup>58</sup> <http://www.dmoz.org/> (Zadnji dostop: 1. december, 2013)

določenega področja (pol)avtomatsko zgraditi model domene. Model domene lahko razumemo na različne načine. V disertaciji začnemo z luščenjem terminov kot pomembnih nosilcev področnega znanja, težišče pa je na luščenju definicijskih kandidatov, ki omogočajo bolj kompleksno razumevanje domene. Razvijemo metodologijo luščenja definicij iz slovenskih in angleških besedil, predvsem pomemben je slednji, saj je luščenje definicij za slovenščino še neraziskano področje. Metodologijo apliciramo na področje jezikovnih tehnologij. Metodologijo strnemo v obliki spletnega delotoka, orodja, ki je prosto dostopno ter preprosto za uporabo. Če je po eni strani rezultat modeliranja področnega znanja nabor izluščenih terminov in njihovih definicij (ki so predstavljeni v obliki pilotnega *Slovarčka jezikovnih tehnologij*), pa v tretjem poglavju uporabimo tudi alternativni pristop razumevanja domene (oz. korpusa) prek gradnje tematskih ontologij.

V nadaljevanju predstavimo gradnjo *Korpusa jezikovnih tehnologij*. Osnovni (kratki) korpus sestoji iz člankov konference Jezikovne tehnologije, ki v Sloveniji od leta 1998 dalje poteka vsako drugo leto (ta korpus je v angleščini poimenovan *LTC proceedings corpus*). Vsi članki iz konferenčnih zbornikov so glede na jezik razvrščeni v slovenski ali angleški del korpusa. Korpus je bil tudi temeljito prečiščen, iz njega so bile izključene sekcije z referencami, imena avtorjev člankov in institucije. Velikost malega (LTC) korpusa je 545.641 različnic. Manjši korpus smo nato razširili z drugimi tipi člankov in izdelali glavni *Korpus jezikovnih tehnologij* (v angleščini *LT corpus*). Pri gradnji tega referenčnega korpusa področja jezikovnih tehnologij v Sloveniji smo že omenjenim člankom zbornikov konference Jezikovne tehnologije dodali izbor doktorskih, magistrskih in diplomskih nalog ter poglavij iz knjig, člankov iz drugih konferenc ter Wikipedije. Po istem principu smo zgradili angleški del primerljivega korpusa. Velikost slovenskega 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).

Predstavitvi korpusa sledi podpoglavje, v katerem domeno jezikovnih tehnologij modeliramo z uporabo (pol)avtomatskega orodja za izdelavo tematskih ontologij OntoGen. Če velja *ontologija* (prim. npr. Gruber, 1993) za formalno reprezentacijo znanja, v kateri so opisani koncepti domene ter odnosi med njimi, je pri *tematski ontologiji* (Fortuna et al., 2006a) področje oz. domena oz. natančneje korpus dokumentov, ki domeno definirajo, opisan s koncepti v obliki najbolj karakterističnih ključnih besed ter s hierarhičnimi odnosi med njimi (podrejeni in nadrejeni koncept). Polavtomatsko orodje OntoGen (Fortuna et al., 2007), ki ga uporabljamo za gradnjo tematskih ontologij, z gručenjem razdeli set dokumentov na hierarhično organizirane koncepte in podkoncepte (vsebine oz. tematike) ter jih opiše s ključnimi besedami, ki jih uporabnik lahko tudi preimenuje. Orodje OntoGen omogoča tudi vizualizacijo v obliki *atlasa dokumentov*. Na angleškem in slovenskem delu malega (*LTC proceedings*) in velikega (LT) korpusa jezikovnih tehnologij smo zgradili modele domene, v katerih smo poleg avtomatskega gručenja uporabili možnost ročnega poimenovanja konceptov ter možnost *aktivnega učenja* (angl. *active learning*), ki ga ponuja program OntoGen. Uporabnika tako program za mejne primere dokumentov vpraša, v katero kategorijo sodijo, ter na podlagi tega izboljša klasifikacijo oz. išče dokumente za manjkajoče koncepte, ki v osnovi niso bili vključeni v ontologijo. Modeli domen so vizualno predstavljeni, na tem mestu pa lahko poudarimo, da se v obeh jezikih ter na obeh domenah – na korpusu člankov konferenčnih zbornikov (*LTC proceedings corpus*) ter na glavnem *Korpusu jezikovnih tehnologij* področje – področje deli na dve glavni podpodročji *računalniško jezikoslovje* (angl. *computational linguistics*) ter govorne

tehnologije (angl. *speech technologies*). *Korpus jezikovnih tehnologij* je veliko večji in bolj heterogen. Opazili smo, da so avtomatsko kategorije veliko slabše razdeljene, kar pripisujemo predvsem zelo različni dolžini dokumentov (od povzetkov do celih doktorskih disertacij). Zgradili smo osnovne tematske ontologije, kjer vidimo male razlike med slovenskim in angleškim delom ontologije. Npr., v slovenskem delu poimenujemo eno izmed topik *računalniško podprto prevajanje*, ki se nato deli na *pomnilnike prevodov* ter *strojno prevajanje*, medtem ko angleški del korpusa vključuje le strojno prevajanje, ne zajema pa področja *pomnilnikov prevodov* ali širše *strojno podprtega prevajanja*. Na koncu na kratko ovrednotimo zgrajene tematske ontologije.

Konec tretjega poglavja predstavlja uvod v glavno temo doktorske disertacije, luščenje definicij iz besedilnih korpusov. Na majhnem izseku našega korpusa analiziramo opažene definicije. Ugotovimo (ter ugotovitve navežemo na teoretično poglavje o tipih definicij), da klasična definicija, sestavljena iz *genusa* in *differentiae*, še zdaleč ni edini način definiranja stavkov v znanstvenih besedilih. Poleg kategorije *genus-differentiae* (znotraj katere vidimo podskupine definicij, kot so definiranje z glagolom *biti*, *drugimi glagoli* ali *brez glagolov*), ločimo *kategorijo definicij s pomočjo sinonimov*, *antonimov*, *sestrskih terminov* in *parafraz*, *kategorijo ekstenzionalnih definicij* ter kategorijo, ki zajema ostale tipe, predvsem definiranje termina s pomočjo njegove rabe (funkcijske definicije, angl. *functional definitions*) ali lastnosti (angl. *typifying definitions*).

V **četrtem poglavju** predstavimo metodologijo za doseganje osrednjega cilja disertacije, to je polavtomatskega modeliranja področnega znanja v obliki terminov in definicij. V nadaljevanju predstavimo že obstoječe tehnologije, ki smo jih v našem delu uporabili, ter nekatere od njih tudi ovrednotimo.

Najprej napravimo kratek shematski prikaz metodologije in predstavimo metode za luščenje definicijskih stavkov. 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).

Vzorci prvega pristopa so bili določeni za vsak jezik posebej, na podlagi analize vzorca definicijskih stavkov, uporabljajo pa leme, besedne oblike ter oblikoskladenjske oznake, kot so npr. skloni samostalnikov (za slovenščino), oseba za glagole itd.

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. 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).

Tretja metoda meri na tip definicij *genus et differentia* in lušči stavke, ki vsebujejo dva izraza, od katerih je eden nadpomenka drugega. Za luščenje stavkov s pojmi v hierarhičnem odnosu smo uporabili semantični leksikon WordNet (PWN, 2010) za angleščino ter sloWNet (Fišer in Sagot, 2008) za slovenščino.

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

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

V nadaljevanju predstavimo že obstoječe vire, orodja in programe, ki smo jih uporabili v našem delu. Začnemo z opisom jezikoslovnega označevalnika ToTrTaLe (Erjavec et al., 2011), s katerim angleška in slovenska besedila segmentiramo, lematiziramo in označimo z oblikoskladenjskimi oznakami. Obstoječe orodje smo implementirali v obliki spletnega servisa. Opišemo tudi luščilnik terminologije za slovenščino in angleščino LUIZ (Vintar, 2010), ki deluje na podlagi oblikoskladenjskih vzorcev ter primerjave pogostosti besed v danem korpusu v primerjavi z referenčnim korpusom. Orodje smo implementirali kot gradnik delotoka ter ga uporabili v našem delotoku za luščenje terminologije in pri eni izmed metod za luščenje definicij. Pozornost namenimo tudi leksikalnima bazama WordNet (Fellbaum, 1998) in sloWNet (Fišer in Sagot, 2008), v katerih so besede (literali) združene v skupine sopomenk (sinseti), vsak sinset pa predstavlja svoj concept. 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 eni od metod luščenja definicij.

Podrobneje predstavimo platformo ClowdFlows (Kranjc et al., 2012), ki je bila uporabljena za implementacijo izdelanega delotoka za luščenje terminologije in definicij. ClowdFlows (Kranjc et al., 2012) je sestavljen iz urejevalnika delotokov (grafičnega uporabniškega vmesnika), kjer lahko uporabnik tudi izbira med že obstoječimi gradniki, ter iz uporabniku nevidnega strežniškega dela, ki skrbi za izvajanje delotokov in shranjevanje velikega števila javno dostopnih delotokov.

Na koncu na kratko evalviramo označevalnik ToTrTaLe ter luščilnik terminov LUIZ, saj sta to tehnologiji, ki imata bistven vpliv na rezultate luščenja definicij. Pri označevalniku ToTrTaLe opišemo napake, pri čemer smo se osredotočili predvsem na slovenščino, napake pa izhajajo iz napačne segmentacije stavkov, napačnega pripisovanja oblikoskladenjskih oznak ter napačne lematizacije. 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.

Pri evalvaciji luščilnika terminologije 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 % kandidatov 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. Za natančnost smo izračunali tudi stopnjo strinjanja med dvema ocenjevalcema. Z linearno uteženo mero  $\kappa$  je strinjanje 32 % za slovenščino ter 26 % za angleščino. Kljub rahlim

variacijam v uteževanju kappe rezultati kažejo na rahlo (kar pomeni večje od naključnega) do zmerno strinjanje, nikoli pa strinjanje med njimi ni precejšnje ali skoraj popolno.

**Peto poglavje** opisuje osrednje eksperimente doktorske disertacije. Razdeljeno je na tri podpoglavja. Prvo zajema luščenje definicij iz slovenskega dela korpusa, drugi del na luščenje iz angleškega dela, tretji del pa najprej povzame osrednje rezultate prvih dveh podpoglavij, glavni doprinos pa je v različnih kombinacijah treh metod. V tem poglavju tudi pokažemo na pomanjkljivost kvantitativne evalvacije ter analiziramo različne tipe definicijskih kandidatov.

Luščenje z leksikoskladenjskimi vzorci iz slovenskih besedil začnemo z eksperimentom o najpreprostejšem vzorcu »X je Y«. Preučimo sedem variacij vzorca y, od oblike »samostalnik *je/sta/so* samostalnik«, do »sam. bes. zv. v imenovalniku (angl. X) *je/sta/so* sam. bes. zv. v imenovalniku«, pri čemer preverimo tudi, kako vpliva pogoj, da se vzorec pojavi na začetku stavka, ter pogoj, da sledi drugemu samostalniku oz. samostalniški besedni zvezi še kaj. Za nadaljevanje izberemo drugi zgoraj podani vzorec, brez pogoja o pojavitvi na začetku izluščenega stavka, saj ima veliko boljši priklic na račun malo slabše natančnosti kot nekateri bolj restriktivni vzorci. Poleg že omenjenega vzorca z uporabo tretje osebe glagola *biti* v sedanjiku, ki povezuje dve samostalniški besedni zvezi v imenovalniku, smo na podlagi analiziranega vzorca definicij definirali še 11 drugih vzorcev, ki uporabljajo npr. glagole *definirati*, *opredeliti*, *opisati*, *nanašati se*, *pomeniti*, *imenovati*, *poimenovati*, *govoriti o* v različnih kontekstih. Če uporabimo vseh 12 tipov vzorcev, izluščimo iz korpusa 389 definicij, kar ocenjujemo na nekaj pod 60 %, z 22,5-odstotno natančnostjo. Ocenimo tudi posamezne vzorce, kar nam omogoča, da v nadaljnjem delu, če se rezultati potrdijo tudi na drugih tipih korpusov, ohranimo le boljše delujoče vzorce. Vse vzorce ponazorimo z izluščenimi primeri, analiziramo pa tudi napačno izluščene primere, kot so presplošni ali prespecifični stavki, stavki izven domene ipd. Pokažemo tudi, da nekateri stavki, za katere bi pričakovali, da jih bo sistem izluščil, niso med kandidati, kar pripišemo med drugim napakam sistema ToTrTaLe za oblikoskladenjsko označevanje.

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. Da bi izboljšali natančnost, smo testirali naslednje hipoteze, ki so implementirane kot parametri v delotoku. Prva hipoteza je, da je natančnost večja, če upoštevamo le termine z višjo terminološko vrednostjo. To smo testirali s spreminjanjem vrednosti parametra, ki določa odstotek najvišje uvrščenih terminov, ki jih upoštevamo (npr. v zgornjem 1 % so bolj zanesljivi kandidati kot v zgornjih 10 %). Druga hipoteza je, da so boljši kandidati tisti, ki imajo več terminoloških izrazov (kar smo preizkusili tako, da smo izbrali parameter pogoja vsaj treh terminov v stavku in ne le osnovnega pogoja, ki upošteva stavke z vsaj dvema terminoma). Naslednja hipoteza preverja, ali na natančnost vpliva pogoj, da se glagol nahaja med dvema terminoma. Če upoštevamo več kot dva termina, testiramo dve variaciji, pri prvi damo pogoj glagola med prvima dvema terminoma, pri drugi pa med katerima koli. Naslednji pogoj, za katerega menimo, da bo izboljšal natančnost, je termin na začetku stavka (prva ali druga beseda). Naslednja hipoteza je, da bo natančnost 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. Hipoteze smo podrobno testirali in rezultate posameznih eksperimentov prikazali v prilogi A. Zgoraj našete hipoteze smo v eksperimentih potrdili in v glavnem velja, da višja kot je terminološka vrednost in število terminov v imenovalniku, večja je natančnost in nižji priklic. Enako velja za število terminov in število večbesednih terminov. Dodatne omejitve, kot so glagol med terminološkimi izrazi, termin na začetku stavka ter pozicija večbesednega termina pred enobesednimi, v večini primerov dodatno izboljšajo natančnost. Najvišjo natančnost dosežemo z najstrožjimi pogoji, vendar tako s približno 26-odstotno natančnostjo izluščimo le 27 definicij. Z izbrano zmernejšo kombinacijo pogojev pa izluščimo 126 definicij z natančnostjo okoli 17,5 %. Poleg kvantitativnih rezultatov tudi analiziramo izluščene kandidate. Metodo luščenja z uporabo terminov primerjamo z metodo z uporabo vzorcev ter prikažemo prednosti in pomanjkljivosti metod. 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. Iz istih razlogov 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.

V tretji metodi luščimo definicije s pomočjo semantičnega leksikona tipa wordnet. Za luščenje definicijskih kandidatov iz slovenskih besedil smo uporabili semantični leksikon sloWNet (Fišer in 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 nadpomenka drugega. Ta metoda ima najslabšo natančnost, saj izluščimo 270 definicij s samo 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.

Naslednje podpoglavje obravnava luščenje definicij iz angleških dokumentov. Pri prvi metodi, tj. metodi luščenja z uporabo leksikoskladenjskih vzorcev, smo za angleščino prilagodili vzorce za luščenje definicij iz slovenskih besedil. Preizkusili smo dve različni nastavitvi, in sicer eno, v kateri se vzorec začne na začetku stavka, ter drugo, 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. Začetek stavka se pokaže kot dober pogoj za boljšo natančnost, ki pa gre na račun manjšega 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 % natančnostjo. Testiramo tudi vmesno rešitev, pri kateri dopuščamo bolj raznolike začetke stavkov, s čimer izluščimo 200 definicij z 28,5-odstotno natančnostjo.

Pri luščenju definicij s pomočjo terminoloških kandidatov iz angleških besedil preverjamo enake hipoteze kot v slovenščini z izjemo pogoja terminov v imenovalniku. Slednje je tudi razlog za dosti slabše delovanje sistema v angleškem jeziku (pod 10 %).

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.

Pri vsaki metodi izluščene stavke tudi analiziramo ter komentiramo možne razloge za dobre oz. pomanjkljive rezultate. Rezultate posameznih metod povzamemo, nato pa se

posvetimo različnim kombinacijam metod na obeh jezikih. Različne kombinacije metod so naravnane k boljši natančnosti ali priklicu. Z 20-odstotno natančnostjo iz slovenskega korpusa izluščimo 486 definicij, medtem ko jih z 32-odstotno natančnostjo izluščimo 107. Pri luščenju definicij iz angleških besedil lahko s kombinacijo metod dosežemo tudi 54-odstotno natančnost (a tako izluščimo le 58 definicij), medtem ko z drugo kombinacijo z 22,5-odstotno natančnostjo izluščenih 230 definicij.

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 (Randolph, 2005; Warrens, 2010), 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, ki označuje dober rezultat strinjanja med ocenjevalci. Ponazorili smo tudi razlike med stavki, kjer se skoraj vsi ocenjevalci strinjajo, in tistimi, kjer se mnenja ocenjevalcev najbolj razhajajo.

V zadnjem delu petega poglavja se posvetimo kvalitativni analizi. Analizirali smo 3404 stavke s 716 definicijami (skupno za slovenščino in angleščino). 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 (funkcijske definicije), definicije samo z nadpomenko (razvrstitvene definicije, angl. *classificatory definitions*), oznake, vezane na *vsebino definicije* (npr. presplošne ali prespecifične), *definiendum* (označili smo, ali gre za lastna imena, kratic, 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 so vezane na morebitne napake v postopku *označevanja*. Sem sodijo stavki, ki se v korpusu pojavljajo večkrat, ali pa stavki, za katere je bila osnovna klasifikacija v kategoriji definicija/nedefinicija spremenjena po prvem ocenjevanju. Vsako oznako (ki se lahko med seboj tudi kombinirajo) ponazorimo s stavkom, ki je bil označen kot definicija in nedefinicija.

V **šestem poglavju** predstavimo delotok, ki implementira našo metodologijo in omogoča enostavno uporabo brez potrebnih namestitev programa. S soavtorji (Pollak et al. 2012a, 2012c) smo zgradili delotok v okolju ClowdFlows. 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. Sledi že omenjeni gradnik *ToTrTaLe*, ki kliče spletni servis ToTrTaLe 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: <http://clowdflows.org/workflow/228/>). Naslednji gradnik delotoka *Term extraction* implementira nekoliko prilagojen luščilnik terminologije LUIZ (Vintar, 2010). Uporabnik zopet 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, na podlagi vnaprej določenih leksikoskladenjskih vzorcev pa z njimi izlušči stavke, ki jim ustrezajo. Dodatni parameter omogoča, da uporabnik izbere,

ali se mora vzorec obvezno nahajati na začetku stavka ali kjer koli (trenutno ta parameter uporabljamo za angleščino, za katero ne moremo uporabljati informacije o sklonu). 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, tako da dobimo stavke izluščene z različnimi metodami. 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 lemmami ter v kanonični obliki) ter definicijskih kandidatov, tretji pa se uporablja za shranjevanje rezultatov.

V **zadnjem poglavju** predstavimo zaključke, glavne prispevke disertacije ter načrte za nadaljnje delo. Glavni cilj disertacije je bil razviti postopek, ki uporabniku omogoča (pol)avtomatsko izluščiti model področnega znanja iz nestrukturiranih besedil v obliki osnutka slovarja. 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). Luščenju iz slovenskih besedil posvetimo več pozornosti, saj podobne metode še ne obstajajo. Dodatni prispevek disertacije 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. Na podlagi v ta namen zgrajenega primerljivega *Korpusa jezikovnih tehnologij* v angleškem in slovenskem jeziku smo izluščili in ovrednotili veliko število definicijskih kandidatov, končni izbor pa smo strnili v pilotni *Slovarček jezikovnih tehnologij*. Pomemben prispevek doktorske disertacije je tudi kvalitativna analiza avtomatsko izluščenih definicijskih kandidatov. Poleg osnovne razvrstitve v dve kategoriji (stavek je ali ni definicija) smo končni nabor stavkov analizirali in označili tudi s podrobnejšimi kategorijami. V predlagani analizi so dodatne oznake ločene v kategorije, vezane na obliko definicije, vsebino definicije, definiendum, segmenatacijo ter označevanje. Za razumevanje različnih vsebin, ki jih korpus pokriva, pa ponujamo tudi osnovni model v obliki tematskih ontologij, zgrajen z orodjem OntoGen.

V primerjavi z delom drugih avtorjev ima naše delo kar nekaj prednosti, ki jih izpostavljamo v nadaljevanju. 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 vedenju – edini tovrstni sistem. Osredotočimo se ne le na kvantitativne rezultate, temveč tudi analiziramo in kritično ocenimo probleme, ki se pojavljajo ob luščenju definicij iz nestrukturiranih znanstvenih besedil. Novost je sistem za luščenje definicij za slovenščino. Poleg luščenja s pomočjo leksikoskladenjskih vzorcev, ki ima že kar dolgo tradicijo (prim. Hearst, 1992),

uvedemo tudi bolj ohlapni metodi s pomočjo terminov in wordneta, vse tri metode pa lahko med seboj tudi kombiniramo. Za razliko od nekaterih drugih pristopov z uporabo strojnega učenja (npr. Navigli in 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. Zanimivo bi bilo v delotok vključiti korak tematskih ontologij ter ga povezati z luščenjem definicij. Za jezikoslovno označevanje besedil bomo preizkusili vpliv orodja Tree Tagger (Schmid, 1994) za angleščino ali nedavno razviti Obeliks za slovenščino (Grčar et al., 2012), za luščenje terminologije pa bi želeli preizkusiti sistema avtorjev Sclano in Velardi (2007) ali Macken et al. (2013). Poleg tega bomo luščenje definicij poskusili izboljšati s strojnim učenjem (začetne eksperimente smo predstavili v Fišer et al. (2010), v novih eksperimentih pa bomo značilke gradili tudi z uporabo atributov, ki smo jih predstavili v pričujočem delu). Dodali bomo poravnavo terminov, izluščenih iz primerljivih korpusov z metodo, predstavljeno v Ljubešić et al. (2011) ter Fišer et al. (2011). Nadaljevali bomo s preučevanjem vplivov različnih načinov evalvacije (npr. petstopenjska lestvica avtorjev Macken et al., 2013). Težišče 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 and Navigli (2013). Preučili bomo tudi možnost uporabe predstavljene metode kot koraka predprocesiranja za različne aplikacije odkrivanja znanj iz besedil.



**IZJAVA O AVTORSTVU  
DOKTORSKE DISERTACIJE**

Podpisana Senja Pollak, z vpisno številko 18091231, rojena 2. 10. 1980 v Kopru, sem avtorica doktorske disertacije z naslovom:

**Polavtomatsko modeliranje področnega znanja iz večjezičnih korpusov  
Semi-automatic Domain Modeling from Multilingual Corpora**

S svojim podpisom zagotavljam, da:

- je predložena doktorska disertacija izključno rezultat mojega lastnega raziskovalnega dela;
- sem poskrbela, da so dela in mnenja drugih avtorjev oz. avtoric, ki jih uporabljam v predloženem delu, navedena v seznamu virov in so v delu citirana v skladu z mednarodnimi standardi in veljavno zakonodajo v RS na področju avtorskih in sorodnih pravic;
- je elektronska oblika identična s tiskano obliko doktorske disertacije;
- soglašam z objavo doktorske disertacije na spletnih straneh Filozofske fakultete Univerze v Ljubljani.

V Ljubljani, dne \_\_\_\_\_

Podpis avtorice: \_\_\_\_\_