NLP Workflow for On-line Definition Extraction from English and Slovene Text Corpora

Senja Pollak[1,2]
Anže Vavpetič[1,3]
Janez Kranjc[1,3]
[1]Jožef Stefan Institute
[3]International Postgraduate School Jožef Stefan
Jamova 39, 1000 Ljubljana, Slovenia
senja.pollak@ijs.si

Nada Lavrač[1,3,4]
Špela Vintar[2]

[2]Faculty of Arts, Aškerčeva 2, 1000 Ljubljana, Slovenia
[4]University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia

Abstract

Definition extraction is an emerging field of NLP research. This paper presents an innovative information extraction workflow aimed to extract definition candidates from domain-specific corpora, using morphosyntactic patterns, automatic terminology recognition and semantic tagging with wordnet senses. The workflow, implemented in a novel service-oriented workflow environment ClowdFlows, was applied to the task of definition extraction from two corpora of academic papers in the domain of Computational Linguistics, one in Slovene and another in English. The definition extraction workflow is available online, therefore it can be reused for definition extraction from other corpora and is easily adaptable to other languages provided that the needed language specific workflow components were accessible as public services on the web.

1 Introduction

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. 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. For instance, Navigli and Velardi (2010), Borg et al. (2010) and Westerhout (2010) have reported very good results with nearly fully automated systems applied to English or Dutch texts. While most of the approaches follow the Aristotelian view of what constitutes a definition *(X is_a Y which ...)*, the concept of definition itself is rarely discussed in detail or given enough attention in the results interpretation. A popular way to circumvent the fuzziness of the "definition of definitions" is to label all non-ideal candidates as defining or knowledge-rich contexts to be validated by the user. In line with this philosophy, the definition extraction approach proposed in this work can be tuned in a way to ensure higher recall at a cost of lower precision.

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 (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, lack appropriate preprocessing tools, such as 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 main challenge addressed in this paper 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. This 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 from textbook texts). In addition to definition candidate extraction we used a classifier trained on Wikipedia definitions to help distinguishing between good and bad definition candidates. When analyzing the results we 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, only rarely fully complying with the classical Aristotelian *per genus et differentiam* formula.

While this work inherits the basic methodology from Fišer et al. (2010), again focusing on Slovene definition extraction, incorporating the same basic methods of extracting definition candidates, this paper extends our previous work in many ways. First, the modules initially developed for Slovene have now been extended to enable the extraction of definition candidates also from English corpora. Second, the modules have been refined and implemented as web services, enabling their inspection and reuse by other NLP researchers. Next, the modules have been composed into an innovative definition extraction workflow. Moreover, this completely reimplemented approach has been evaluated on an different corpus, both regarding its genre and size. The corpus is from a very specific domain, which is a much more realistic scenario when developing specialized terminological dictionaries.

The developed workflow was applied to definition extraction from two corpora of academic papers in the area of Computational Linguistics, one in Slovene and another in English. The developed workflow has been implemented in our recently developed service-oriented workflow construction and management environment ClowdFlows[1] (Kranjc et al., 2012). The definition extraction workflow is available on-line[2], therefore it can be reused for definition extraction from other corpora and is easily adaptable to other languages provided that the needed language specific workflow components were accessible as public services on the web.

The paper is structured as follows. Section 2 summarizes the main definition extraction methods incorporated into the NLP definition extraction workflow, followed by the actual definition extraction workflow description in Section 3. Experimental evaluation of the workflow on the Slovene and English Computational Linguistics corpora is presented in Section 4. Section 5 concludes with a discussion, conclusions and plans for further work.

2 Summary of main definition extraction methods

Like in Fišer et al. (2010), we employ three basic methods to extract definition candidates from text. The approach postulates that a sentence is a definition candidate if one of the following conditions is satisfied:

- It conforms to a predefined lexico-syntactic pattern (e.g., NP [nominative] is_a NP [nominative]),
- It contains at least two domain-specific terms identified through automatic term recognition,
- It contains a wordnet term and its hypernym.

The first approach is the traditional pattern-based approach. In Fišer et al. (2010), we use a single, relatively non-restrictive *is_a* pattern, which yields useful candidates if applied to structured texts such as textbooks or encyclopaediae. However, if used on less structured authentic specialized texts, such as scientific papers or books used in the experiments described in this paper, a larger range of patterns yields better results. For the described experiments, we used eleven different patterns for Slovene and four different patterns for English.

The second approach 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 term recognition module[3] identifies potentially relevant termi-

[1] http://clowdflows.org

[2] http://clowdflows.org/workflow/76/

[3] The term extraction methodology is described in detail in Vintar (2010).

nological phrases on the basis of predefined morphosyntactic patterns (Noun + Noun; Adjective + Noun, etc.). These noun phrases are then filtered according to a weighting measure that compares normalized relative frequencies of single words in a domain-specific corpus with those in a general reference corpus. As a reference corpus we used FidaPlus[4] for Slovene and BNC[5] for English. The largest coverage is achieved under the condition that the sentence contains at least two domain terms (term pair). Additional conditions are that the first term should be a multi-word term at the beginning of a sentence, and that there is a verb between a term pair (a detailed comparison of results obtained with different settings is beyond the scope of this paper).

The third approach exploits the *per genus et differentiam* characteristic of definitions and therefore seeks for sentences where a wordnet term occurs together with its direct hypernym. For Slovene, we use the recently developed sloWNet (Fišer and Sagot, 2008) which is considerably smaller than the Princeton WordNet (PWN) (Fellbaum, 1998) and suffers from low coverage of terms specific to our domain.

3 NLP workflow for on-line definition extraction

This section describes the NLP workflow, implemented in the ClowdFlows workflow construction and execution environment. We first present the underlying principles of workflow composition and execution. We then present a technical description of the ClowdFlows environment, followed by a detailed presentation of the individual steps of the definition extraction workflow.

3.1 Basics of workflow composition and execution

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., 2007) and RapidMiner (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 of workflows has previously been implemented at 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) implements the described features with a distinct advantage. ClowdFlows 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.

3.2 The ClowdFlows environment illustrated by a simplified NLP workflow

ClowdFlows consists of the workflow editor (the graphical user interface, as shown in Figure 1) and the server-side application, which handles the execution of the workflows and hosts a number of publicly available workflows.

The workflow editor consists of a workflow canvas and a widget repository, where widgets represent embedded chunks of software code, representing downloadable stand-alone applica-

[4]FidaPlus is a 619-million word reference corpus of Slovene (http://www.fidaplus.net).

[5]Mike Scotts wordlist from the BNC World corpus (http://www.lexically.net/downloads/version4/downloading%20BNC.htm) was used.

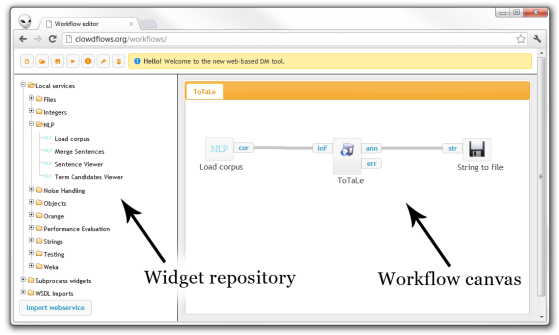

Figure 1: A screenshot of the workflow editor in the Google Chrome browser.

tions which look and act like traditional applications but are implemented using web technologies and can therefore be easily embedded into third party software. All our 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 our workflow public, so that anyone can execute it. The workflow is simply exposed by a unique address which can be accessed from any modern web browser. Whenever the user opens a public workflow, a copy of the workflow appears in her private workflow repository in ClowdFlows. The user may execute the workflow and view its results or expand it by adding or removing widgets.

3.3 A detailed description of the definition extraction workflow and its components

The entire definition extraction workflow implemented in ClowdFlows is shown in Figure 2.

The widgets implementing the existing software components include:

- ToTaLe tokenization, morphosyntactic annotation and lemmatization tool (Erjavec et al., 2010) for Slovene and English[6].
- LUIZ term recognition tool (Vintar, 2010) for Slovene and English, with a new implementation of scoring and ranking of term candidates.

The core definition extraction widgets include:

- Pattern-based definition extractor,
- Term recognition-based definition extractor,
- WordNet- and sloWNet-based definition extractor.

Numerous other new auxiliary text processing and file manipulation widgets were developed and incorporated to enable a seamless workflow execution. These include:

- Load corpus widget, which allows the user to conveniently upload her corpus in various formats (PDF, txt, doc, docx) either as single files or several files together in one flat ZIP file,
- Term candidate viewer widget, which formats and displays the terms (and their scores) returned by the term extractor widget (a subset of the extracted term candidates is illustrated in Figure 3),
- Sentence merger widget, which allows the user to join (through intersection or union) the results of several definition extraction methods,
- Definition candidate viewer widget, which, similarly to the term candidate viewer widget, formats and displays the candidate definition sentences returned by the corresponding methods (Figure 4 illustrates the widget's output, listing the extracted definition candidates to be inspected by the user).

4 Experimental evaluation on the Language Technologies corpus

This section describes the corpus, the experimental results achieved and the quantitative and qualitative evaluation of results.

[6]In future versions of the workflow, we plan to replace the ToTaLe web service with ToTrTaLe which handles also ancient Slovene and produces XML output.

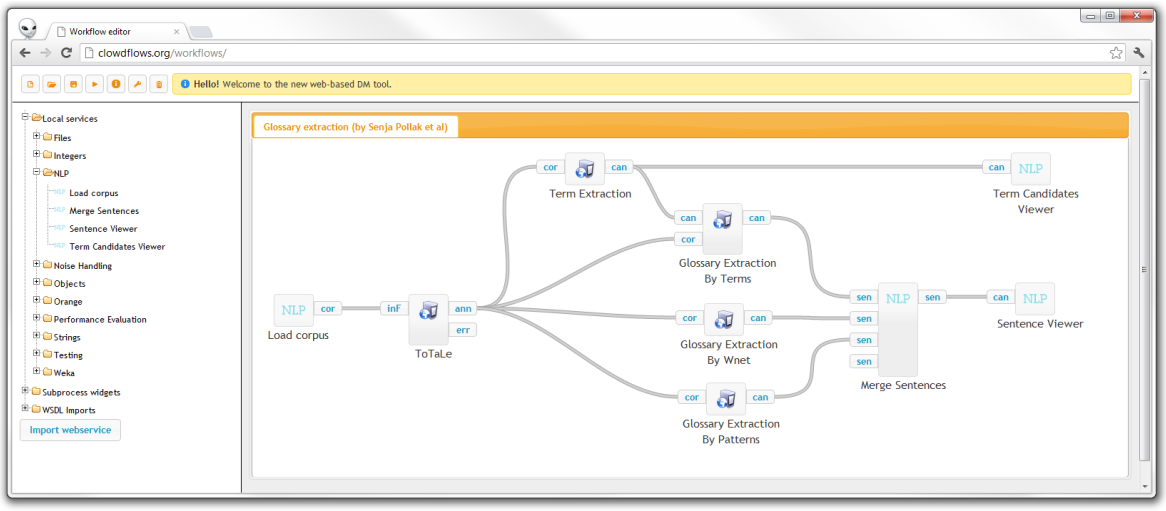

Figure 2: A screenshot of the entire definition extraction workflow.

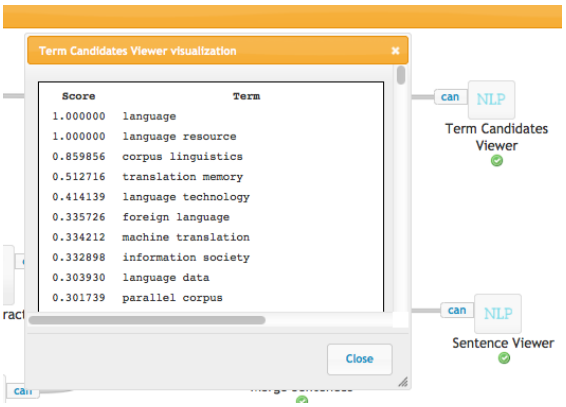

Figure 3: Viewing of selected term candidates.

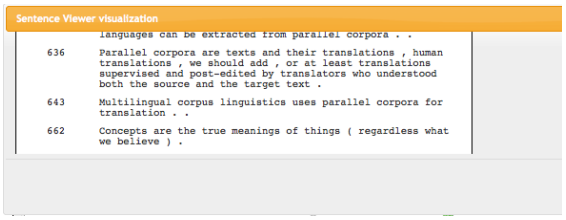

Figure 4: Viewing of selected definition candidates.

4.1 The corpus

For highly specialized domains and for languages other than English, the web may not provide an ideal corpus, especially not for the purpose of terminology extraction where a certain level of representativeness and domain coverage is crucial. Our corpus consists of papers published in the proceedings of the biennial Language Technologies conference (Jezikovne tehnologije) that has been organized in Slovenia since 1998. The articles are in Slovene or English. To improve vocabulary coverage we added other text types from the same domain, including Bachelors, Masters and PhD theses, as well as several book chapters and Wikipedia articles.

The total size of the corpus is 44,750 sentences (903,621 tokens) for Slovene and 43,019 sentences (929,445 tokens) for English.

4.2 Experimental results

In this section we evaluate the term extraction (see Subsection 4.2.1) and the glossary extraction method (in Subsection 4.2.2). More attention is paid to the latter, where not only quantitative results are provided, but we also analyze and discuss the results from the linguistic perspective.

4.2.1 Term extraction results

We evaluated top 200 (single- or multi-word) domain terms for each language (see Table 1). 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 termi-

nological collocation, not fully lexicalized, compositional in meaning; score 3).

Precision	English terms	Slovene terms
Yes (2-5)	0.775	0.845
Yes (5)	0.48	0.55

Table 1: Precision of term extraction method.

The second part of term evaluation involved the assessment of recall. The domain expert annotated a random text sample of the Slovene and English corpus with all terminological expressions (approximately 65 for each language), and the samples were then compared to the lists of terms extracted by the LUIZ system. Table 2 shows the results for both samples using either all term candidates or just the top 10,000/5,000.

	Number of terms	Recall
Slovene	38,523	0.694
	10,000	0.527
	5,000	0.444
English	25,007	0.779
	10,000	0.644
	5,000	0.491

Table 2: Recall of terminological candidates extracted from the Slovene and English Language Technologies corpus.

4.2.2 Definition extraction results

The results of definition extraction methods on the Language Technologies corpus are presented in Table 3, showing the number of candidates extracted with each individual method, as well as the number of candidates obtained with the intersection of at least two methods (*Intersect*) and those extracted by at least one of the three methods (*Union*). The latter shows that by using all the methods of the NLP definition extraction workflow we extracted 4,424 definition candidates for English and 6,638 for Slovene.

The reason for extracting a larger candidate set for Slovene compared to English is that the Slovene corpus is larger in the number of sentences, that the pattern-based approach is more elaborate (containing 11 patterns compared to only 4 patterns for English), and that the number of extracted Slovene terms is larger.

Number of candidates	English def. candidates	Slovene def. candidates
Patterns	474	1,176
Terms	866	1,539
Wordnet	3,278	4,415
Union	4,424	6,638
Intersect	192	472

Table 3: Definition candidates extracted from the Language Technologies corpus.

Precision	English def. candidates	Slovene def. candidates
Patterns	0.44	0.26
Terms	0.08	0.15
Wordnet	0.13	0.05
Union	0.09	0.11
Intersect	0.33	0.25

Table 4: Precision of definition extraction methods.

From a set of extracted definition candidates, obtained as outputs of each of the methods, 100 sentences were randomly selected and used for the evaluation of the precision of our workflow (see Table 4).

Precision is better for English than for Slovene. Concerning the patterns, the reason can be in less fixed word order in Slovene, while for the wordnet-based method we observed that the selected wordnet pairs were too general and that many domain specific terms were not found in sloWNet.

To evaluate the recall of our methods, we randomly selected 1,000 sentences for each language. In the Slovene data set there were 21 definitions out of which 10 were extracted by at least one of our methods (0.4762 recall). The English 1,000 sentences random corpus contained 25 definitions, out of which 15 were extracted (0.6 recall). We plan to perform further evaluation to get the results on a larger test set.

To gain a better insight into the types of definition candidates, we reassessed each method and analyzed their output. It is clear from these results that simple patterns still procure best results, while the union of different methods yields a lot of potentially interesting candidates, but much more noise.

When analyzing the evaluation sets we observed that a definition in real text is often not easy to define and evaluate. A lot of sentences in running text can be considered as borderline cases, often without the hypernym and defining the term either through its extension or its purpose; see the examples (i) and (ii) that have not been identified by any method, but can be considered as definitions.

(i) *Z aktivno kamero lahko torej "s pogledom sledimo" obrazu govorca, kadar se ta premika.* [*With an active camera we can "eyetrack" the face of the speaker when he is moving.*]

(ii) *Osembitni kodni nabor ISO 8859-2 je na mestih s kodami od 0 do 127 identičen standardu ISO 646, na preostalih 128 mestih pa kodira vse potrebne znake za pisanje v albanščini, češčini, [...] in slovenščini.* [*The 8-bit ISO 8859-2 codepage is identical to ISO 646 at codes ranging from 0 to 127, while it uses the other 128 codes to encode the characters used for writing Albanian, Czech [...] and Slovene.*]

The analysis of candidate sentences shows that the notion of definition, especially when we attempt to formalize it, needs to be reconsidered. Different definition types found in the evaluation set include: *formal definitions with genus and differentia structure* (X is Y), whereby the definiendum does not necessarily occur at the beginning of the sentence; definitions with genus and differentia structure, where the verb is other than the verb "biti" [to be]; sentences where a term is not defined through its hypernym, but through a *sibling concept and the differentia*; *informal definitions*, subordinated in a sentence and introduced with a relative pronoun; *extensional definitions*, i.e. definition which instead of specifying the hypernym lists all the possible realizations of a concept (X includes Y, Z and Q); *defining by purpose* (hypernym is omitted); *definition as textual formula* used for mathematical concepts.

5 Discussion, conclusions and further work

One of the contributions of this paper is the improvement and an in-depth experimental assessment of the individual methods constituting our NLP definition extraction workflow. The other main contribution is the implementation of the definition extraction workflow, which has been made publicly available within a novel service-oriented workflow composition and management platform ClowdFlows. The contributions and plans for further work are discussed in more detail below.

Based on the qualitative analysis of our methods, we identified a number of definition types not traditionally covered by definition extraction systems. Based on these findings we started to improve our methodology in several ways. Even if compared to previous experiments the patterns were already extended from strict *is_a* pattern to a larger set of patterns, the approach could be further extended to cover all the alternatives listed above (e.g., extensional definitions). The pattern-based method had the highest precision, but should be extended with other methods to ensure better coverage (e.g., by the candidates at the intersection of wordnet- and term-based methods). Concerning the term-based extraction method we are conducting further experiments, based on the threshold for the termhood parameter setting and the additional restriction that the terms identified should be in the nominative case (for Slovene). Finally, the wordnet method was improved by limiting sloWNet nouns in Slovene to nominative case only and using a different setting of the window parameter. Regarding the evaluation of recall, the experiments on a larger test set are being performed.

Concerning the new NLP workflow implementation of our definition extraction modules based on morphosyntactic patterns, automatic terminology recognition and semantic tagging with WordNet/sloWNet senses, its on-line availability and modularity are a great advantage compared to the existing NLP software, including other terminology and definition extraction tools. The workflow implementation within the novel ClowdFlows workflow composition and execution engine enables workflow reuse for definition extraction from other corpora, experiment reproducibility, as well as the ease of workflow refinement by the incorporation of new NLP modules implemented as web services and workflow extensions to other languages.

In future work we plan to refine the definition extraction components to improve the precision

and the recall and to develop new workflow components for on-line natural language processing.

Acknowledgments

We are grateful to Darja Fišer for past joint work on definition extraction from Slovene texts. The presented work was partially supported by the Slovene Research Agency and the FP7 European Commission project "Large scale information extraction and integration infrastructure for supporting financial decision making" (FIRST, grant agreement 257928).

References

Michael R. Berthold, Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Ktter, Thorsten Meinl, Peter Ohl, Kilian Thiel and Bernd Wiswedel. 2007. KNIME: The Konstanz Information Miner. In *GfKl. Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, 319-326.

Claudia Borg, 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 Evaluations*, LREC 2010, Valletta, Malta, 2577-2584.

Łukasz Degórski, Michał Marcinczuk and Adam Przepiórkowski. 2008a. Definition extraction using a sequential combination of baseline grammars and machine learning classifiers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, LREC 2008, Marrakech, Morocco, 837-841.

Łukasz Degórski, Lukasz Kobyliński and Adam Przepiórkowski. 2008b. Definition extraction: Improving balanced random forests. In *Proceedings of the International Multiconference on Computer Science and Information Technology*, 353-357.

Janez Demšar, Blaž Zupan, Gregor Leban and Tomaž Curk. 2004. Orange: From experimental machine learning to interactive data mining. In *Proceedings of ECML/PKDD-2004*, LNCS Volume 3202, 537-539.

Tomaž Erjavec, Darja Fišer, Simon Krek and Nina Ledinek. 2010. The JOS linguistically tagged corpus of Slovene. In *Proceedings of the 7th International Conference on Language Resources and Evaluations*, LREC 2010, Valletta, Malta, 1806-1809.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press. Online version: http://wordnet.princeton.edu

Darja Fišer and Benoît Sagot. 2008. Combining multiple resources to build reliable wordnets. *Text, Speech and Dialogue* (LNCS 2546). Berlin; Heidelberg: Springer, 61-68.

Darja Fišer, 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), 2932-2936.

Duncan Hull, Katy Wolstencroft, Robert Stevens, Carole Goble, Matthew R. Pocock, Peter Li and Thomas M. Oinn. 2006. Taverna: A tool for building and running workflows of services. *Nucleic Acids Research* (Web-Server-Issue) 34: 729-732.

Lukasz Kobyliński and Adam Przepiórkowski. 2008. Definition extraction with balanced random forests. In *Proceedings of GoTAL'2008*, 237-247.

Janez Kranjc, Vid Podpečan and Nada Lavrač. 2012. ClowdFlows: A cloud-based scientific workow platform. In *Proceedings of ECML/PKDD-2012*. Springer LNCS (in press).

Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz and Timm Euler. 2006. YALE: rapid prototyping for complex data mining tasks. In *Proceedings of KDD-2006*, ACM, 935-940.

Roberto Navigli 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.

Vid Podpečan, Monika Zemenova and Nada Lavrač. 2012. Orange4WS environment for service-oriented data mining. *The Computer Journal, 55*(1): 82-98.

Adam Przepiórkowski, Łukasz 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 Balto-Slavonic NLP workshop* at ACL 2007, Prague, 43-50.

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

Ian H. Witten, Eibe Frank and Mark Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques.* Third Edition. Morgan Kaufmann.

Eline Westerhout. 2010. *Definition Extraction for Glossary Creation: A study on extracting definitions for semi-automatic glossary creation in Dutch.* Netherlands Graduate School of Linguistics / Landelijke - LOT Dissertation Series.