Terminology Extraction for Academic Slovene Using Sketch Engine

Darja Fišer[1,2], Vít Suchomel[3,4], Miloš Jakubíček[3,4]

[1] Dept. of Translation, Faculty of Arts, University of Ljubljana,
Aškerčeva cesta 2, SI-1000 Ljubljana, Slovenia
[2] Department of Knowledge Technologies, Jožef Stefan Institute,
Jamova cesta 3, SI-1000 Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

[3] Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{xsuchom2,jak}@fi.muni.cz

[4] Lexical Computing
Brighton, United Kingdom and Brno, Czech Republic
{vit.suchomel,milos.jakubicek}@sketchengine.co.uk

Abstract. In this paper we present the development of the terminology extraction module for Slovene which was framed within the Sketch Engine corpus management system and motivated by the KAS research project on resources and tools for analysing academic Slovene. We describe the formalism used for defining the grammaticality of terms as well as the calculation of the score of individual terms, give an overview of the definition of the term grammar for Slovene and evaluate it on a Slovene KAS corpus of academic Slovene.

Key words: terminology, term extraction, Sketch Engine, academic discourse, Slovene

1 Introduction

The development of the academic part of any language is an important indicator of its vitality [1,2,3]. A key component of any scientific communication is terminology which needs to be analysed by linguistcs and terminologists but also has to be made easily accessible to domain experts, such as researchers, lecturers and practicioners, as well as to students, translators and editors [4,5]. Since authoritative terminological dictionaries are always lagging behind the fast-paced development of terminology in the scientific domain, corpus-driven and collaborative approaches to terminology management have become an attractive alternative in the past decades [6,7], also for Slovene [8].

To achieve this for Slovene, we have compiled a large KAS corpus of 50,000 scientific texts with over one billion tokens [9] and are now in the process of developing term extraction for it, which is the subject of this paper.

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016, pp. 135–141, 2016. © Tribun EU 2016

The paper is organised as follows: in the next section we present an adaptation of the terminology extraction module in Sketch Engine, a leading corpus management tool [10] and its terminology extraction methodology. Next, we outline the term grammar for Slovene and finally conclude with an evaluation of the terms extracted from the KAS corpus of academic Slovene that will provide useful insight for future refinement of the term extraction tool and the term grammar.

2 The Sketch Engine Environment

Sketch Engine is an online corpus management system providing access to hundreds of text corpora which can be searched and analyzed. It received its name after one of its key features—word sketches, one page summaries of a word's collocational behaviour in particular grammatical relations.

As of 2016, Sketch Engine hosts preloaded corpora for 85 languages and allows users to create new ones, either by uploading their own texts or by building the corpus semi-automatically from the web according to the keywords given by the user. The latter approach is particularly useful for fast development of domain corpora from online resources.

2.1 The Terminology Extraction Module

Sketch Engine contains a terminology extraction module [11] using a contrastive approach for finding term candidates, in a similar manner to other such systems [12,13]. Two corpora are given as input to the term extraction: a focus corpus consisting from texts in the target domain, and a (ideally very big) reference corpus against which the focus corpus is then compared. To improve the accuracy of this process, Sketch Engine selects only grammatically valid phrases.

Therefore the whole extraction is a two-step process:

1. **unithood**: the first step is rule based and language dependent. We assess the grammatical validity of a phrase (*unit hood*) using a so called *term grammar*. A term grammar describes grammatically plausible terms using regular expressions over available annotation in the corpus, such as morphosyntactic tags and lemmas.
2. **termhood**: candidate phrases generated in the first step are then contrasted with the reference corpus by using the "simplemath" statistic [14] which compares their normalized frequencies focusing either on less frequent or more frequent phrases.

The result of the term extraction is a list of (multi-word) term candidates sorted according to their simplemath score.

2.2 Term Grammar for Slovene

The Slovene term grammar v1.0 is based on the original term definition for Russian by Maria Khokhlova and the transformed term definition for Czech by Vít Suchomel. From these sources it has then been considerably extended with additional term patterns needed for a comprehensive terminological analysis of the KAS corpus of academic Slovene.

Default Attrubutes. The grammar defines 12 default attributes, 10 of which are MSD-based to ensure high-accuracy term candidate identification and 2 serve to achieve agreement in gender, number and/or case across the multi-word term for improved accuracy of term identification.

Term Patterns. The main part of the term grammar are term patterns that use combinations of the defined default attributes to identify and render the extracted term candidates. The following parts of speech were considered as possible elements of term patterns: noun, adjective, preposition, conjunction, adverb and verb. v1.0 of the Slovene term grammar enables term extraction of single as well as multi-word terms consisting of noun as well as verb phrases up to length 4.

In total, 44 term patterns have been defined:

- 4-grams: 22 patterns
- 3-grams: 15 patterns
- bigrams: 6 patterns
- unigrams: 1 pattern

An illustrative example of a term pattern is given in Figure 1. The first line contains rendering instructions for the given pattern. According to it, the first word should be rendered as a lowercased lemma while the rest of the elements should be displayed as lowercased word forms. The second line contains the pattern to be identified in the corpus. The rule will identify all the 4-grams in the corpus that start with a noun and are followed by 2 consecutive adjectives and a noun in the genitive case. In addition, the pattern requires that the second and the third element in the term agree with the final noun in gender, number and case. The third line contains the CQL version of the same pattern suitable for corpus querying to facilitate development and editing of the term grammar. With a similar purpose, the last line shows an example of such a pattern from the corpus.

3 Evaluation

3.1 KAS-PhD Corpus

An evaluation of the term extraction module and the term grammar for Slovene was performed on the KAS subcorpus of 700 PhD theses which contains almost

```
*COLLOC "%(1.lemma_lc)_%(2.lc)_%(3.lc)_%(4.lc)-x"
1:noun 2:adj_genitive 3:adj_genitive 4:noun_genitive & agree(2,4) & agree(3,4)
#"Nc.*" "A.*g.*" "A.*g.*" "Nc.*g.*"
#metoda magnetronskega ionskega naprševanja
```

Fig. 1: Example of a term pattern in term grammar.

150,000 pages of text or 53 million tokens published in the period 2000-2015. Most theses in the corpus are from Social and Technical Sciences, some are from Natural Sciences while there are very few from Biomedical Sciences and the Humanities. While terminology is typically extracted for a limited domain, our main goal in this paper was to evaluate the term grammar, which is why we believe using a heterogeneous corpus is more suitable as it will highlight different characteristics and issues across several domains.

3.2 Results

The evaluation was performed on the 1,000 top-ranking 3- and 4-gram term candidates from the KAS-PhD corpus with respect to the reference slTenTen corpus of general Slovene [15]. A large majority of them were bigrams, with only a few 3- and 4-grams:

- 4-grams: 28 (2.8%) term candidates
- 3-grams: 177 (17.7%) term candidates
- bigrams: 795 (79.5%) term candidates

Manual evaluation consisted of three steps. First, pattern productivity was considered in order to determine which patterns in the term grammar have a good yield. Next, term candidates were checked for unithood and structural accuracy so as to identify any remaining bugs in the term grammar. In the end, termhood of the extracted candidates was tested, the goal of which was to suggest further refinements of term ranking and smoothing.

Results for 4-grams. As there were only 28 4-gram candidates, all were manually examined. By far the most productive patterns in this category of the extracted term candidates are noun phrases that contain a preposition (68%, e.g. `družba z omejeno odgovornostjo`, followed by the much less productive combinations of adjectives and nouns (18%, e.g. `zaznana vrednost blagovne znamke`) and patterns that contain a conjunction (14%, e.g. `mera središčnosti in pomembnosti`).

In terms of structural accuracy and unithood, patterns containing prepositions or conjunctions significantly outperform adjective+noun combinations (75% wrt. 40%), indicating that further fine-tuning of term rendering is required (e.g. `vlaknast*ega* beton visoke trdnosti`). The observed problems with unithood are predominantly truncated candidates (e.g. `/?/ proces na` mednarodne trge), candidates spanning across the border of a term have not been observed in the analysed sample.

Finally, the best-performing category regarding termhood of the extracted candidates were those that contain conjunctions (75% wrt. 60% candidates with prepositions and 58% for adjective+noun combinations). False positives are either phrases common in general-language (e.g. pogovor o likovni nalogi) or unusual constructions, specific of a particular thesis in the corpus (e.g. cestni otrok v makejevki).

Results for 3-grams. 3-grams were evaluated by examining 100 random candidates from the list of 1,000 top-ranking extracted term candidates. Adjective+noun combinations and candidates containing prepositions were equally prolific (43% wrt. 42%, e.g. gostota magnetnega pretoka, računalništvo v oblaku). While 15% of the candidates were verb phrases, it turns out they were all noise as they all contained the verb to be, so they were excluded from further analysis (e.g. biti v uporabi, v raziskavi smo). The term grammar needs to be refined accordingly.

Unithood and structural accuracy is better preserved in candidates containing prepositions (83% wrt. 70% in adjective+noun combinations) where the biggest problem seems to be the preservation of the gender and number of the premodifiers (e.g. magnetn*e* poljsk*e* jakost). Manual analysis gives clear indications that term candidates already subsumed in longer phrases should receive special treatment (e.g. sistem za podporo *odločanju* wrt. sistem za podporo).

Termhood, the toughest test for the extracted candidates, shows that 63% of the candidats containing prepositions could be considered terms while the rest are flormulae typical of academic writing (e.g. razlika med anketiranci) or even general-language constructions (e.g. spoznavanje prek spleta). Adjective+noun combinations do better in this respect, achieving 73% accuracy. Again, scholar-specific phrasing is frequent (i.e. teza doktorske disertacije), slightly less so as far as general-language patterns are concerned (e.g. nova finančna storitev). It is interesting to note that term candidates extracted from technical and natural sciences theses typically suffer from unithood issues while lack of termhood is generally observed in the candidates extracted from social sciences and humanities documents.

4 Conclusions

We presented the construction of the first version of the Sketch Engine term grammar for Slovene and its application to term extraction from a corpus of PhD theses from different scientific domains against the slTenTen corpus. In manual evaluation we focused on 4- and 3-grams for which we analysed pattern productivity, unithood and structural accuracy of the extracted candidates as well as their termhood. While substantially fewer 4-grams were extracted, their pattern range was greater then in 3-grams. Eventhough unithood and structural accuracy varied more in 4-grams and was also lower in general than in 3-grams, termhood results were similar in both. This suggests that accuracy can be easily improved by further refining the term grammar.

The presented term grammar for Slovene is applicable to other corpora using compatible morpho-syntactic tagging and will be made freely available on the website of the KAS project: `http://nl.ijs.si/kas/english/`. Apart from term grammar refinement we plan to perform a set of comparative analyses on domain-specific subcorpora as well as extend the performance test to less scientific but more prolific MA and BA theses. For this, we will enhance the term extraction output which will enable the user to switch term ranking by termhood or by term patterns as needed in order to be able to focus on a particular term pattern or term pattern family. A systematic comparison of term extraction recall and precision with the CollTerm tool [16] will also be performed.

Acknowledgments. The work described in this paper was funded by the Slovenian Research Agency within the national basic research project "Slovene Scientific Texts: Resources and Description" (J6-7094, 2014-2017), the COST Action "European Network of e-Lexicography" (IS1305, 2013-2017) and the Ministry of Education of the Czech Republic within the Lindat Clarin Center.

References

1. Kalin Golob, M., Stabej, M., Stritar Kučuk, M., Červ, G., Koprivnik, S.: Genre analysis: Jezikovna politika in jeziki visokega šolstva v Sloveniji. Založba FDV (2014)
2. Swales, J.: Genre analysis: English in academic and research settings. Cambridge University Press (1990)
3. Genre, A.: Language use in professional settings. Applied Linguistics and Language Study.) London: Longman (1993)
4. Logar, N.: Aktualni terminološki opisi in njihova dostopnost. (2013) 58–64
5. Logar, N.: Korpusna terminologija: primer odnosov z javnostmi. Trojina: zavod za uporabno slovenistiko; Založba FDV (2013)
6. Daille, B., Gaussier, É., Langé, J.M.: Towards automatic extraction of monolingual and bilingual terminology. In: Proceedings of the 15th conference on Computational linguistics-Volume 1, Association for Computational Linguistics (1994) 515–521
7. Heylen, K., De Hertog, D.: Automatic Term Extraction. Handbook of Terminology, Volume 1. John Benjamins Publishing Company (2015)
8. Vintar, Š.: Bilingual term recognition revisited the bag-of-equivalents term alignment approach and its evaluation. Terminology **16**(2) (2010) 141–158
9. Erjavec, T., Fišer, D., Ljubešić, N., Logar, N., Ojsteršek, M.: Slovenska znanstvena besedila: prototipni korpus in načrt analiz. In Erjavec, T., Fišer, D., eds.: Proceedings of the Conference on Language Technologies and Digital Humanities, Ljubljana, Slovenia, Academic Publishing Division of the Faculty of Arts (2016) 58–64
10. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. Lexicography **1** (2014)

11. Kilgarriff, A., Jakubíček, M., Kovář, V., Rychlý, P., Suchomel, V.: Finding terms in corpora for many languages with the Sketch Engine. EACL 2014 (2014) 53
12. Dagan, I., Church, K.: Termight: Identifying and translating technical terminology. In: Proceedings of the Fourth Conference on Applied Natural Language Processing. ANLC '94, Stroudsburg, PA, USA, Association for Computational Linguistics (1994) 34–40
13. Logar, N., Špela Vintar, Špela Arhar Holdt: Luščenje terminoloških kandidatov za slovar odnosov z javnostmi. In: Proceedings of the Eighth Language Technologies Conference, Jožef Stefan Institute (2012) 135–140
14. Kilgarriff, A.: Comparing corpora. International journal of corpus linguistics **6**(1) (2001) 97–133
15. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., Suchomel, V., et al.: The tenten corpus family. In: 7th International Corpus Linguistics Conference CL. (2013) 125–127
16. Pinnis, M., Ljubešic, N., Stefanescu, D., Skadina, I., Tadic, M., Gornostay, T.: Term extraction, tagging, and mapping tools for under-resourced languages. In: Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012), June. (2012) 20–21