Automation of Lexicographic Work Using General and Specialized Corpora: Two Case Studies

Iztok Kosem[1], Polona Gantar[2], Nataša Logar[3], Simon Krek[4]
[1]Trojina, Institute for Applied Slovene Studies
[2]Fran Ramovš Institute for the Slovene Language, ZRC SAZU, Ljubljana, Slovenia
[3]Faculty of Social Sciences, University of Ljubljana
[4]Jožef Stefan Institute, Ljubljana
iztok.kosem@trojina.si, apolonija.gantar@guest.arnes.si,
natasa.logar@fdv.uni-lj.si, simon.krek@guest.arnes.si

Abstract

Due to increasingly large amounts of authentic data to analyse, lexicographers are nowadays looking to language technologies to provide them with not only the tools to analyse the data, but also with tools and methods that ease and speed up the data analysis. One of the most promising avenues of research has been the automation of early stages of the corpus data analysis, with the aim to summarize, and consequently reduce, the amount of corpus data that the lexicographers need to examine. However, most of this research deals with general lexicography; terminology is yet to extensively test these methods. This paper attempts to address this gap by presenting two separate Slovene research projects, one lexicographic (Slovene Lexical Database) and the other terminological (Termis), that used the same method of automatic extraction of corpus data (presented in Kosem et al. 2013). After describing the projects and the corpora use, similarities and differences in the parameter settings and the quality of extracted data in the two projects are presented. We conclude with discussing the further potential of automation in both general and specialised lexicography.
Keywords: data extraction; terminology; general language; collocations; dictionary; GDEX

1 Introduction

In recent years, lexicography has witnessed several projects where automation of different aspects of lexicographer's work has been successfully implemented, such as detection of new words or meanings (Cook et al. 2013) or initial data extraction (Kosem et al. 2013). This trend of increasing the role of a computer in the dictionary-making process follows Rundell and Kilgarriff's (2011) vision of focussing lexicographer's tasks towards validating and completing the data extracted by a computer.

The calls for automation originate mainly from general lexicography where lexicographers are faced with increasingly larger corpora that they need to analyze. But what about using automation in the making of dictionaries, such as terminological dictionaries, where much smaller and more specialized corpora are used? To what extent can automation methods used in general lexicography be trans-

ferred to specialized lexicography or terminology? This paper attempts to provide some answers to these questions by describing and discussing two Slovenian projects, one lexicographic (Slovene Lexical Database) and the other terminological (Termis), that tested the use of automation in database compilation.

We first briefly describe both projects, and the corpora used for automatic extraction of data. This is followed by the description of the automatic process, and an overview of the settings used in the two projects. Then, the findings are presented, focusing on the differences as well as similarities identified between the results of automatic data extraction in the two projects. We conclude with some thoughts on the further potential of automation in both general and specialised lexicography, and outline our plans for the future.

2 Slovene Lexical Database

The Slovene Lexical Database (SLD; Gantar & Krek 2011) is one of the results of the Communication in Slovene[1] project that has developed language data resources, natural language processing tools and resources, and language description resources for Slovene. The Slovene Lexical Database has a twofold goal: it is intended as the basis for the future compilation of different dictionaries of Slovene, both monolingual and bilingual, and as such its concept is biased towards lexicography. Secondly, it will be used for the enhancement of natural language processing tools for Slovene. The database is conceptualized as a network of interrelated lexico-grammatical information on six hierarchical levels with the semantic level functioning as the organizing level for the subordinate ones. The six levels are:

- lemma or the headword,
- senses and subsenses (labelled with semantic indicators and in many cases described with semantic frames),
- multi-word expressions,
- syntactic structures (representing a formalization of typical patterns on the clause and phrasal level),
- collocations, and
- corpus examples.

1 The operation is partly financed by the European Union, the European Social Fund, and the Ministry of Education and Sport of the Republic of Slovenia. The operation is being carried out within the operational program Human Resources Development for the period 2007–2013, developmental priorities: improvement of the quality and efficiency of educational and training systems 2007–2013. Project web page: http://eng.slovenscina.eu/.

3 Terminological database Termis

Applied research project Termis[2] took place between 2011 and 2013. The aim of the project was the compilation of an online dictionary-like terminological database for the discipline of public relations. The basis of the project was KoRP,[3] a corpus of public relations texts (Logar 2013). It has been envisaged from the beginning that the entries in the terminological database would contain English translations of headwords, explanations, syntactic and collocational information, and corpus examples. The database in now completed and is freely accessible online at http://www.termania.net. It comprises 2000 entries that also offer links to the KoRP corpus and the Gigafida corpus, a reference corpus of Slovene.

4 Using automatic data extraction in the two projects

The decision to use automatic extraction of lexical information from the corpus in both projects comes from the need to reduce time and cost connected with the production of dictionaries, by utilizing new possibilities offered by state-of-the-art tools for corpus analysis. The main idea behind using automatic extraction of corpus data is to reduce the amount of time spent by lexicographers on examining corpus data, especially on browsing through plethora of corpus examples. Lexicographic analysis remains corpus-based (or driven); however, the initial selection of corpus data to be analysed is left to the computer. The lexicographer then examines, validates, and completes the information and shapes it into the final dictionary entry.

The automatic method used in the two projects relies heavily on Word Sketch (Kilgarriff & Tugwell 2002) and GDEX (Good Dictionary Examples; Kilgarriff et al. 2008), two functions that are part of the Sketch Engine corpus tool. The method requires a lemma list, sketch grammar for the building of word sketches, GDEX configuration(s), and settings that set thresholds for data extraction. An API script is then used to extract from the corpus collocates under grammatical relations, defined in the sketch grammar, and examples of their use. The method is described in more detail in Kosem et al. (2013), thus the next sections focus on the main differences in the automatic method used by the two projects.

4.1 Corpora

The basis for the extraction of lexical information for the Slovene Lexical Database was the Gigafida corpus[4] (Logar Berginc et al. 2012), containing 1.18 billion words or 39,427 texts created between 1990

2 http://www.termis.fdv.uni-lj.si/
3 http://nl.ijs.si/noske/sl-spec.cgi/first_form?corpname=korp_sl
4 http://www.gigafida.net/

and 2011 with printed texts representing 84.35% and internet texts 15.65%. Printed part contains fiction (2%), non-fiction and textbooks (4%), and periodicals such as daily newspapers (56%) and magazines (21%). Text originating from the web were published on news portals, pages of large Slovene companies and more important governmental, educational, research, cultural and similar institutions.
Automatic extraction of lexical information for the Termis project was conducted on a much smaller, specialised corpus – the KoRP corpus – containing 1.8 million words. The texts in the KoRP corpus were selected according to carefully designed criteria (Logar 2007) that make the corpus representative of a public relations field in Slovenia. It is important to note that the two corpora were lemmatised and morphosyntactically tagged with the same statistical tagger (Grčar, Krek & Dobrovoljc 2012), enabling comparisons of extracted data.

4.2 List of lemmas

The two projects used completely different approaches to devising a list of lemmas for automatic extraction. For the Slovene Lexical Database, a more homogenous group of lemmas was used, mainly comprising of not too frequent lemmas that were either monosemous of less polysemous according to sloWNet, a Slovene version of Wordnet (Fišer, 2009). Less polysemous nature of lemmas also enabled a better comparison of data extraction with the Termis project, given that the terms in Termis were mainly monosemous. An additional criterion for selection, which was preferred but not mandatory, was the absence of the lemma in the Dictionary of Standard Slovenian (SSKJ). The final selection included 515 nouns, 260 verbs, 275 adjectives and 117 adverbs and was dominated by lemmas with frequency between 1000 (0.85 per million words) and 10,000 (8.5 per million words).
In the Termis project, the lemma list was in fact a headword list and was built using a term extraction tool (Vintar 2010)[5]. The list contained 2127 items: 941 nouns, 199 verbs and 987 multi-word terms. Single- and multi-word term candidates have been extracted using morphosyntactic patterns and term weights, calculated by comparing their frequencies in the KoRP corpus and in a reference corpus of Slovene called FidaPLUS (Arhar Holdt & Gorjanc 2007), as well as phraseological stability of the extracted terminological unit. Each term candidate was carefully examined in its natural environment – the texts in the KoRP corpus – by a terminologist and experts in the field of public relations.

4.3 GDEX configurations

The GDEX tool (Kilgarriff et al. 2008) ranks corpus examples according to their quality, using measurable parameters such as example length, whole sentence form, syntax, and presence/absence of rare words, etc. The majority of work associated with devising GDEX configurations for automatic extraction was done during the SLD project; drawing on the experience in developing the first version of

5 http://lojze.lugos.si/cgitest/extract.cgi

GDEX for Slovene (Kosem et al., 2011), four different configurations were designed, one for each word class in the SLD (noun, verb, adjective, adverb), the process involving several iterations of evaluation and comparison of results produced by the last two versions of configuration (Kosem et al. 2013). A good indication of the difference between the first version of GDEX for Slovene and the version for automatic extraction is that the former was designed to provide at least three good examples among the ten offered, while the latter aimed to have the top three examples meet the criteria of a good example. The Termis project's point of departure was using the final GDEX configurations used by the SLD project, evaluating them on a sample of lemmas and making adjustments to the heuristics, which proved to be minor, until the results were satisfactory. In the end, two different GDEX configurations were used, one for nouns and multi-word units, and one for verbs.

4.4 Settings for extraction

This part of the automatic extraction introduced the greatest number of differences between the two projects. Preparation of settings for extraction included providing values for the following six parameters:

- number of examples per collocate
- number of collocates per grammatical relation
- minimum frequency of a collocate
- minimum frequency of a grammatical relation
- minimum salience of a collocate
- minimum salience of a grammatical relation.

For the SLD project, three examples per collocate were extracted, and for the Termis project two examples per collocate. Both projects used a limit of maximum 25 collocates per grammatical relation.
The values of the remaining four parameters had to be obtained with statistical and manual analysis of the word sketches of a sample of lemmas used in automatic extraction. Namely, initial tests during the SLD project showed that the same values could not be used for all grammatical relations and collocates; for example, more salient and frequent relations of word classes (e.g. *adjective + noun* for adjectives) required higher thresholds due to a large number of collocates. Also, corpus frequency of the lemma played a vital role in setting the values; more frequent lemmas had more extensive word sketches and required higher thresholds, whereas rarer lemmas required lower thresholds or no thresholds at all. Consequently, both projects divided lemma lists into different frequency groups, with different settings used for each group. The SLD project used three frequency groups for each word class, with different frequency ranges for different word classes. On the other hand, the Termis project used three frequency groups for verbs, four frequency groups for nouns, and three frequency groups for multi-word units. Each category in the Termis project contained one group, the so-called 0 group, that included low frequency lemmas for which all the data available in the word sketches was extracted.

The only values that were shared by the two projects were values for minimum collocation salience for nouns and values for minimum gramrel salience for verbs; all other values were (much) lower for the Termis project than for the SLD project. This was a direct result of the difference in the sizes of the corpora used for automatic extraction of data.

4.5 Extracted lexical information: general language vs. specialized language

It is worth noting that a term as a name for a concept in a certain discipline is more difficult to specify than it is presented and argumented in the general theory on terminology (Wüster 1931; Felber 1984) – at least if terms are observed and identified in the context (Pearson 1998, as well as other perspectives, e.g. Cabré Castelví 2003). Such complexity of terms has been adequately summarized by Sager (1998/99) who argued that terms are merely words with a specific function, or in other words, terms are formally not very different from other words. This fact causes great difficulties to terminographers during preparatory stages, i.e. while preparing the headword list; on the other hand, this similarity between terms and other words is an advantage during the extraction of lexical context, as terminography can utilize lexicographic knowledge and tools for the analysis and description of a general language.

So far, we have compared grammatical relations/syntactic structures found in both Slovene Lexical Database and Termis, using a smaller number of noun entries that have a higher frequency per million words in the KoRP corpus than in the Gigafida corpus. The analysis showed that a large percentage of words acquire the specialised meaning only at a context level, especially with compounds or when we are dealing with polysemous words that have one of their meanings used also in a specialised domain or have developed their own specialised meaning.

The comparative analysis also focused on identifying syntactic structures common to both the general corpus (Gigafida) and the specialised corpus (KoRP), more specific to one of the corpora, or exclusive to one of the corpora. Similar comparison was made for collocations in both vocabularies. The sketch grammar contains 258 grammatical relations functioning as syntactic structures, and the automatically extracted data for noun entries showed that there were 69 (27%) attested syntactic structures, i.e. structures with identified collocates, in both corpora, 188 (73%) syntactic structures were found only in the Gigafida corpus, while one syntactic structure was found only in the KoRP corpus. These findings confirm that terminology does not differ from general language on a syntactic level, i.e. does not form terminology-specific syntactic structures. There are exceptions, however they are specific to particular lexical items; thus, a syntactic structure can be found in the language, but is not typical for a specific verb, noun, adjective etc. as used in the general language. An example of this is the structure VERB + NOUN4 for the collocation *communicate message*, which is typical for the field of public relations, but not for general Slovene where the pattern *communicate* + *about* + NOUN5 is more commonly used.

5 Discussion

The automatic extraction approach proved successful in both projects, in terms of providing good enough data for devising database entries and saving a great deal of lexicographer's/terminologist's time spent on more routine tasks. One of the important findings is that the steps used in the general language project (SLD) could be replicated in the terminological project (Termis), with some elements requiring little change (e.g. GDEX configurations) or no change at all (e.g. sketch grammar). Also, the evaluation of extracted corpus sentences in both projects reported good quality of the examples. The comparison clearly shows that the main work on any future project adopting this methodology would be dedicated to determining the settings for data extraction. Namely, this step exhibited the greatest differences between the projects, mainly on account of a significant difference in the size of the corpora used for automatic extraction.

Different nature of projects also enabled us to evaluate and test the approach on different lemmas in terms of corpus frequency and consequently in the amount of corpus data available. In SLD, the minimum frequency of a lemma was 600 occurrences (0.5 times per million words)[6] in the Gigafida corpus, whereas the threshold in Termis was determined by terminological potential of the word rather than its frequency (for example, some terms had only two or three occurrences in the KoRP corpus[7]). For high frequency lemmas, more work on settings for extraction was required in order to find the right balance between exporting enough data and excluding irrelevant grammatical relations and/or collocates. For very rare lemmas, i.e. for those in groups 0 in the Termis project[8], it was established that the value of the automatic approach is mainly in saving lexicographer's time by directly exporting all the data for each lemma and importing it into the dictionary-writing system, thus changing the lexicographer's task from analysis-selection-copying to validation-deletion.

The automatic extraction of data for multi-word units was conducted only for Termis, as the project was conducted after the conclusion of the SLD project when a new feature called Multi-word links had already been implemented in the Sketch Engine. The automatic extraction of lexical information was only possible for two-word patterns such as *adjective + noun* and *noun + noun*, and not for others (e.g. *noun + preposition + noun*). It is therefore not possible to make comparisons of the projects as far as automatic extraction of data for multi-word units is concerned. Nonetheless, we can report that the data obtained in the Termis project was found to be of similar quality as the data for single-word terms, with the main difference being in the GDEX configuration and settings used.

What is left for lexicographers to do are tasks such as sense division, definition writing, distributing and cleaning the automatically extracted information etc.; and as shown by studies such as Kosem et al. (2013), some of those tasks can be left to non-lexicographers, e.g. by using crowd-sourcing. Further-

6 Majority of lemmas had frequency between 1000 (0.85 per million words) and 10,000 (8.5 per million words) in the Gigafida corpus.

7 This still meant that these terms had a higher frequency per million words (1.1) than the least frequent lemmas in the SLD project.

8 Groups 0 contained 889 terms in total.

more, the feedback from the terminologists devising entries in the Termis project showed that many extracted examples already contained definitions of terms or at least the information needed to devise them, indicating further avenues for the implementation of automation. It is noteworthy that the entries in the Termis database contain (encyclopaedic) definitions that are short (one sentence), medium-length (multi-sentence paragraph) or even longer (several paragraphs); in general they are longer than definitions (or semantic frames) in the Slovene Lexical Database.

6 Conclusion

Technological advances gave rise to corpora, enabling lexicographers to describe language more accurately and in greater detail than ever before, but ironically, corpora have now become a problem for lexicographers due to the increasingly larger amounts of data they contain. Consequently, it seems inevitable that more and more lexicographic tasks will become automated. There is simply too much data to analyse and not enough time to do it in – in addition, users want quick(er) access to up-to-date information. Initial experience on a Slovene lexicographic project has showed promising results, but it is even more encouraging that the automatic approach appears to be suitable also for terminological purposes.

The automatic method by Kosem et al. (2013) has the most potential for projects where a dictionary or a database is devised from scratch,[9] but it can also be useful for existing dictionaries. For example, periodical automatic extraction of regularly updated corpus data could facilitate quicker detection and description of new meanings and usages of the words. This remains one of the avenues of future research; namely, how to automatically extract and include in the database only the new information on the use of a particular word or phrase. By this we do not mean only new words and meanings, but also new uses of existing meanings.

Future plans as far as the Slovene Lexical Database is concerned include a more in-depth evaluation of entries devised with automatically extracted data, as well as their comparison with manually devised entries. We also aim to test automatic extraction on more frequent lemmas, where we expect much more work with setting parameters for extraction. Further use of the automatic approach is planned on the terminological side, possibly by testing its usefulness in a few other domains. Finally, we aim to explore automatic extraction of information not covered by the existing automatic method. One of such areas is definition extraction; for example, future plans with the Termis database include conducting an experiment on automatic definition extraction from the KoRP corpus, using the recently-developed methodology, specially adapted for Slovene (Pollak 2014).

9 For example, the automatic data extraction method is an integral part of a proposal for a new dictionary of contemporary Slovene (Krek et al., 2013).

7 References

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

Cabré Castellví, M. T. (2003). Theories of terminology: Their description, prescription and explanation. *Terminology* 9(2), pp. 163–199.

Felber, H. (1984). *Terminology Manual*. Paris: Infoterm.

Fišer, D. (2009). SloWNet – slovenski semantični leksikon. In M. Stabej (ed.) *Infrastruktura slovenščine in slovenistike (Obdobja 28)*. Ljubljana: Znanstvena založba Filozofske fakultete, pp. 145–149.

Cook, P., Lau, J. H., Rundell, M., McCarthy, D., Baldwin, T. (2013) A lexicographic appraisal of an automatic approach for detecting new word senses In Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 49-65.

Gantar, P., Krek, S. (2011). Slovene lexical database. In D. Majchraková, R. Garabík (eds.) Natural language processing, multilinguality: sixth international conference, Modra, Slovaška, 20-21 October 2011, pp. 72-80.

Grčar, M., Krek, S., Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In T. Erjavec, J. Žganec Gros (eds.) *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, pp. 89-94.

Logar, N. (2007). Korpusni pristop k pridobivanju in predstavitvi jezikovnih podatkov v terminoloških slovarjih in terminoloških podatkovnih zbirkah: doktorska disertacija. Ljubljana: Filozofska fakulteta.

Logar, N. (2013). *Korpusna terminografija: primer odnosov z javnostmi*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

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

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of the 13th EURALEX international congress.* Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, pp. 425-432.

Kilgarriff, A., Tugwell, D. (2002). Sketching words. In H. Corréard (ed.) *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins.* Euralex, pp. 125-137.

Kosem, I., Gantar, P., Krek, S. (2013). Automation of lexicographic work: an opportunity for both lexicographers and crowd-sourcing. In Kosem, I., Kallas, J., Gantar, P., Krek, S., Langemets, M., Tuulik, M. (eds.) *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia*. Ljubljana/Tallinn: Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut, pp. 32-48.

Kosem, I., Husák, M., McCarthy, D. (2011). GDEX for Slovene. In I. Kosem, K. Kosem (eds.) Electronic Lexicography in the 21st Century: New applications for new users, Proceedings of eLex 2011, Bled, 10-12 November 2011. Ljubljana: Trojina, Institute for Applied Slovene Studies, pp. 151-159.

Krek, S., Kosem, I. Gantar, P. (2013). *Predlog za izdelavo Slovarja sodobnega slovenskega jezika*, v1.1. Available at: http://www.sssj.si/datoteke/Predlog_SSSJ_v1.1.pdf

Pollak, S. (2014). *Polavtomatsko modeliranje področnega znanja iz večjezičnih korpusov/Semi-automatic domain modeling from multilingual corpora* (Semi-automatic Domain Modeling from Multilingual Corpora). PhD thesis. Ljubljana: University of Ljubljana, Faculty of Arts, Department of Translation. Accessed at: http://kt.ijs.si/theses/phd_senja_pollak.pdf. [25/03/2014]

Pearson, J. (1998). *Terms in context*. Amsterdam, Philadelphia: John Benjamins.

Rundell, M., Kilgarriff, A. (2011). Automating the creation of dictionaries: where will it all end? In F. Meunier, S. De Cock, G. Gilquin, M. Paquot (eds.). *A Taste for Corpora. A tribute to Professor Sylviane Granger*. Amsterdam: Benjamins, pp. 257–281.

Sager, J. C. (1998/99). In Search of a Foundation: Towards the Theory of the Term. *Terminology*, 5(1), pp. 41-57.

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

Wüster, E. (1931). *Internationale Sprachnormung in der Technik, besonders in der Elektrotechnik*. Berlin: VDJ.