The first Slovene automatically compiled dictionary of abbreviations

Mojca Kompara

Keywords: *computational lexicography*, *dictionary*, *abbreviation*.

Abstract

Abbreviations are difficult to deal with (Gabrovšek, 1994) and represent a growing phenomena present in all languages. The scope of this article is to present the first Slovene automatically compiled dictionary of abbreviations. In the paper we present how we automatically extract abbreviation-expansion pairs out of newspaper texts and obtain genuine pairs, how we cope with the automatic editing phase and add language qualifiers to expansions and transform non-nominative expansions into nominative. The first Slovene automatically compiled dictionary of abbreviations is available online, free of charge, on the web site of Termania. It is the first dictionary produced automatically from newspaper articles with the help of algorithms. Algorithms represent a link between the text and the semi automatic production of a dictionary of abbreviations. That is why the production and further development of algorithms is essential and useful for lexicographers.

1. Introduction

This article presents the first Slovene automatically compiled dictionary of abbreviations. Automatic recognition of abbreviations in electronic texts has been dealt with by Taghva (1998), Larkey et al. (2000), Schwartz and Hearst (2003), Chang et al. (2002), and Zahariev (2004). Zahariev's (2004) approach is considered special because it is not limited to recognizing only one language. Abbreviations are problematic because of their different variants of expansions. Abbreviations are a problem for machine translation because they are difficult to translate and can be incorrectly interpreted as nouns; for example, Slovene *Nato* 'NATO' (cf. *nato* 'then'), *Kad* 'Pension Fund Management' (cf. *kad* 'vat'), and *Sod* 'Slovenian Compensation Fund' (cf. *sod* 'barrel'). The automated approach recognizes abbreviation-expansion pairs [1] and forms a trial database for compiling dictionaries of abbreviations. The automated approach diminishes routine and time-consuming steps in compiling dictionaries of abbreviations, such as lemmatization of expansions and automatic addition of language qualifiers.

2. Methodology

In order to obtain an automatically compiled dictionary of abbreviations, several steps must be taken. The first step involved obtaining the data for compiling such a dictionary. An algorithm for automatic recognition of abbreviations and abbreviations expansions filtered a corpus of 60 million words and yielded 5,820 abbreviation-expansion pairs. Genuine pairs[2] were included in the *Termania*[3] editing software. The following steps focused on the automatic compilation of the dictionary entries. Expansions were automatically converted into the nominative case and language qualifiers were automatically added to the expansions. The final part of the paper shows simple and complex dictionary entries in *Termania*.

3. Automatic recognition of abbreviation-expansion pairs

Words up to ten capital letters written in parentheses were selected as abbreviation candidates; (e.g., *NATO*). Words up to ten letters with only the first letter capitalized (e.g., *Mig*) were also used. Abbreviations such as *itd.* 'etc.', *npr.* 'e.g.', *ipd.* 'etc.', *itn.* 'etc.', and so on were not included because usually they do not appear with an expansion in a text. In order to identify a genuine set of candidates for abbreviations, words that are not abbreviations—such as proper names, names of places, and so on—were excluded, using the *Slovar slovenskega knjižnega jezika* (Standard Slovene Dictionary). Genuine abbreviations are determined by their expansion(s) and are divided into official and/or non-official. An abbreviation can have several expansions and the algorithm took into consideration all the expansions an abbreviation could have. To recognize expansions, four types of abbreviations were used. The first type is *covered abbreviations* in which letters match the words in left or right context (e.g., *FF* with the expansion *Filozofska fakulteta* 'Faculty of Arts'). The second type is *abbreviations with expansions containing prepositions and conjunctions* (e.g., *FDV = Fakulteta za družbene vede* 'Faculty of Social Sciences'). This algorithm also takes into consideration expansions with one additional word (e.g., *za* 'for'). The third type concerns *abbreviations composed of the first two letters of words* (e.g., *NAMA = Narodni magazin*; literally, 'people's store'). The fourth type covers *abbreviations with prepositions* (e.g. *DZU = Družba za upravljanje* 'trust company'), in which prepositions appear in the abbreviation and also in the expansion. In order to obtain the abbreviations' expansions in texts, all four types of patterns—*(abbreviation) expansion, (expansion) abbreviation, abbreviation (expansion), expansion (abbreviation)*—were used. Abbreviations identical to legal words (lexicalized abbreviations; e.g., *Nama, Kad, Sod*, etc.) are very common in Slovene. The word *Nama* may be both an acronym for *Narodni magazin* and a personal pronoun meaning 'to the two of us' at the beginning of a sentence. Such abbreviations are usually well known but problematic and misleading for the algorithm. They were included in the algorithm rules via the dictionary of abbreviations *Slovarček krajšav*. After the newly established rules for recognition, a demo version of the algorithm was prepared. The system called *Mkstrings* (http://mkstrings.farhouse.si/) is freely accessible and is able to filter texts rich in abbreviations.

3.1. *Results*

A large corpus composed of 60 million words (from the newspaper *Delo* from 2005 to 2009) was used. The algorithm filtered the corpus in 30 minutes and yielded 5,820 abbreviation-expansion pairs. The pairs obtained were manually checked and verified using *Google*. To determine whether an abbreviation-expansion pair is genuine it has to be manually checked. In the final list containing abbreviation-expansion pairs, several problems were observed, such as the occurrence of oblique cases, multiple occurrences of the same expansion, and abbreviations without expansions. Abbreviations without a matching expansion in the text were automatically deleted. In the manual revision that followed, the most neutral case was preserved and all identical pairs appearing more than once were deleted. The precision of the algorithm is 96% and the recall is 82%. Among the good expansions many occurred more than

once and/or with tiny modifications; for example, usage of various cases or spellings, as seen in Table 1. In Table 1 only three expansions out of six are genuine, but the genuine ones are also not lemmatized and cannot be included in a dictionary as such.

Table 1. Abbreviation-expansion pair for *MNZ*.

MNZ	MNZ
1 ministrstva za notranje zadeve	1 ministrstva za notranje zadeve
2 medobčinskih nogometnih zvez	2 medobčinskih nogometnih zvez
3 ministrstvom za notranje zadeve	3 Muzej novejše zgodovine
4 Medobčinske nogometne zveze	
5 Muzeja novejše zgodovine	
6 Muzej novejše zgodovine	

After excluding false pairs[4] (4%), verification, and checking good pairs, 2,665 genuine abbreviations-expansion pairs occurred (in various cases).

3.2. *Purpose of the data obtained*

Genuine abbreviation-expansion pairs obtained from the corpus using the algorithm were automatically included in the *Termania* editing software.

4. Conversion of expansions into nominal form

The main problem in the automatic production of simple and complex dictionary entries is abbreviation expansions that appear in non-nominative cases. Slovene has six cases, but for dictionary purposes only the nominative case is used.[5] Another problem is number, which may be singular, plural, or dual, as well as the occurrence of other languages. *Presis*,[6] a machine translation program developed by Amebis, was used. It supports Slovene-to-English, English-to-Slovene, and German-to-Slovene translations and it is part of the *iTranslate4.eu*[7] project. It is a rule-based system consisting of analyzers and generators. Analyzers translate text in Slovene, English, or German to *Presis Interlingua*, and then generators translate *Presis Interlingua* into Slovene or English. The idea is to use the *Presis* Slovene analyzer to translate the expansion in question into *Presis Interlingua*. A special version of the analyzer, which allows only subjects and objects in various cases, is used. If the result is a subject, it is already in the nominative case and no further work is needed. If the result is an object, the *Interlingua* translation is changed and the object becomes the subject. The resulting Slovene "translation" is the same noun phrase in the nominative case. An important part of this procedure is the meaning information that must be removed from the *Interlingua* translation before sending it to the generator, otherwise some words in the generated nominative form may be replaced by synonyms. Some problems appear, such as the nominative plural form instead of some non-nominative singular form, definite forms of adjectives, capitalization, disambiguation in considering the first word to be a (proper) noun instead of an adjective, and the occurrence of doubled expansions.

4.1. *Results*

After all changes are applied for conversion into the nominative form, the result was 2,661 correct expansions and 433 mistakes. Conversion solved 70.9% of the cases. The majority of the remaining mistakes concerned capitalization and number, but there were also some problems with unknown words.

5. Language identification

All expansions were automatically given language qualifiers. Statistical methods are often used for determining the language of a text (Dunning, 1994). However, these methods do not work well on very short texts, such as expansions. In the first stage the expansion is sent to the Slovene analyzer. If the analyzer is successful, the language code "sl" is assigned; if not, the language code "sl-x" is assigned (i.e., Slovene – to be manually checked). In cases where there is still no code, the English analyzer is applied and, if it is successful, the language code "en" is assigned; if not, the language code "en-x" is assigned (i.e., English – to be manually checked). The procedure continues for every language. If no analyzer is successful and they all consider a word unknown, then the language code "xx" (i.e., unknown – to be checked and assigned manually) is assigned. Out of 3,094 language tags, 126 were assigned incorrectly; 95.9% of the selected tags were correct.

6. Entries in Termania

After applying these changes, the first Slovene online dictionary of abbreviations was available online free of charge at the *Termania* website. Entries in *Termania* are either simple or complex. Simple entries are composed of a Slovene abbreviation-expansion pair; for example, *FF, Filozofska fakulteta* 'Faculty of Arts' or *AB, Alzheimerjeva bolezen* 'Alzheimer's disease'. As seen from Figure 1, the language qualifier (sl) is provided and the expansion is in the nominative case (*Alzheimerjeva bolezen*).

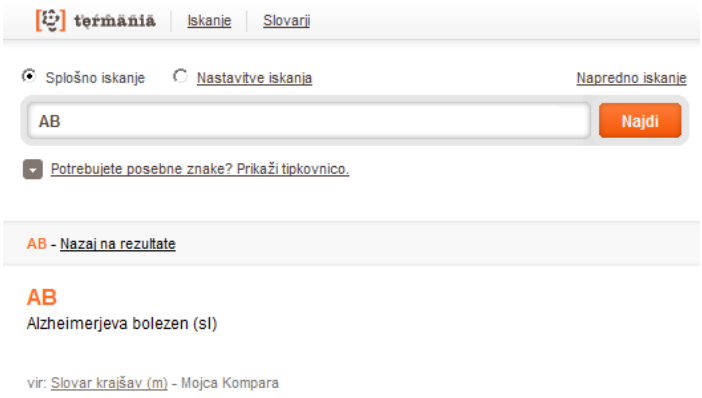

Figure 1. Simple dictionary entry for AB.

At present the algorithm provides simple entries entirely automatically. Such entries work

perfectly well in a Slovene dictionary of abbreviations, and the only embellishment that could be added is encyclopedic data or additional explanation of the term. Not all abbreviations need such data, and so a manual selection of items should be made. Encyclopedic data must be short and clear. A possible solution is encyclopedias such as *Wikipedia*. The term *Alzheimerjeva bolezen* exists in *Wikipedia*, and so the explanation could be included directly into the dictionary entry *AB* or only as a hyperlink. Complex entries contain foreign abbreviations in which the language is provided and the expansion is checked, but the Slovene translation is missing and there is no encyclopedic data, as seen from Figure 2.

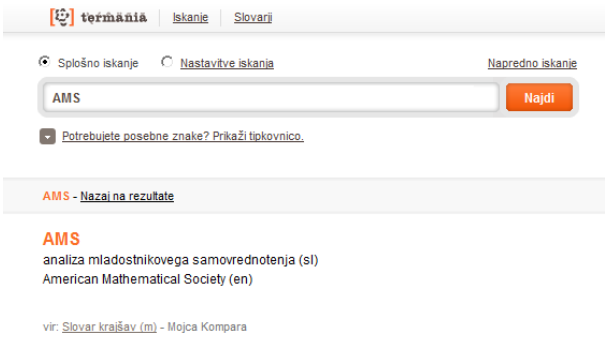

Figure 2. Complex dictionary entry for AMS.

For now translations, descriptions and encyclopedic data can be included only manually and so complex entries are not produced automatically, like the simple ones are. In the future there will be an attempt to provide automatic translations, descriptions, and encyclopedic data in complex entries.

7. Conclusion

This paper presents how abbreviation-expansion pairs are automatically extracted out of newspaper texts and genuine pairs are obtained, how the automatic editing phase is handled and language qualifiers are added to expansions, and how non-nominative expansions are transformed into nominative ones. The work is placed online, free of charge, at the *Termania* editing software website so it can be shared with users. It is the first dictionary produced automatically[8] from newspaper articles with the help of algorithms. Algorithms make it possible to create a semiautomatic or fully automatic dictionary of abbreviations and such a dictionary represents the future of electronic lexicography. Algorithms represent a link between the text and the semiautomatic production of a dictionary of abbreviations. This is why the production and further development of the algorithm is essential and useful for lexicographers.

Notes

[1] An abbreviation-expansion pair is an abbreviation together with what it stands for; for example, *NATO = North Atlantic Treaty Organization*.

[2] Abbreviation-expansion pairs that result in real abbreviations and not just plain text after manual revision; for example, *MNZ = Muzej novejše zgodovine* 'Museum of Contemporary History'.

[3] http://www.termania.net/. *Termania* is a free on-line dictionary portal with integrated dictionary browsing and editing tools developed by *Amebis* software (Kamnik, Slovenia) in cooperation with Trojina, the Institute for Applied Slovene Studies.

[4] These are abbreviation-expansion pairs that manual checking shows to not be real abbreviations, but simply plain text; for example, *PO = prvih oddelkih* 'first departments'.

[5] Filtering yielded expansions in several cases; for example, *Muzeja novejše zgodovine* (genitive), *Muzeju novejše zgodovine* (dative/locative), and *Muzej novejše zgodovine* (nominative). For dictionary purpose, only the nominative case is applicable; (i.e., *Muzej novejše zgodovine*).

[6] http://presis.amebis.si

[7] http://itranslate4.eu

[8] Some procedures were performed manually, such as selecting genuine pairs, and cannot be automated yet.

References

Amebis Termania. http://www.termania.net.

Chang, J. T. et al. 2008. 'Creating an Online Dictionary of Abbreviations from MEDLINE.' *Journal of American Medical Informatics Association (JAMIA)* IX.VI: 612–620.

Dunning, T. 1994. 'Statistical Identification of Language.' *Technical Report MCCS 94-273*, New Mexico State University.

iTranslate4. http://itranslate4.eu.

Larkey, L. et al. 2000. 'Acrophile: An Automated Acronym Extractor and Server.' In *Proceedings of the fifth ACM conference on Digital libraries 2000, San Antonio, 2-7 June 2000*. San Antonio, Texas, USA.

MKstrings. http://mkstrings.farhouse.si/.

Presis. http://presis.amebis.si.

Schwartz, A. S. and M. A. Hearst 2003. 'A simple algorithm for identifying abbreviation definitions in biomedical texts.' In *Proceedings of the Pacific Symposium on Biocomputing, 3-7 January 2003.* Lihue, Hawaii, USA.

Slovarček krajšav. http://bos.zrc-sazu.si/kratice.html.

Slovar slovenskega knjižnega jezika. http://bos.zrc-sazu.si/sskj.html

Taghva, K. and J. Gilbreth 1998. 'Recognizing acronyms and their definitions.' *IJDAR* I.IV: 191–198.

Zahariev, M. 2004. *A (Acronyms)*. PhD thesis, School of Computing Science, Simon Fraser University, Ottawa, Canada.