A corpus based e-dictionary of terminology as a body of knowledge

Nataša Logar Berginc

A corpus based e-dictionary of terminology as a body of knowledge

Nataša Logar Berginc

Faculty of Social Sciences, University of Ljubljana
Slovenia

Correspondence to: natasa.logar@fdv.uni-lj.si

Abstract. The paper describes the process of compiling an on-line terminological database within the TERMIS project. The compilation began from an LSP corpus (i.e. KoRP, a corpus of public relations texts) and involved automatic term recognition performed for single- and multi-word terms and the automatic extraction of lexical information from the corpus. Entries in the terminological database contain English translations of headwords, explanations, syntactic and collocational information, and corpus examples. The database comprises 2000 entries that also offer links to the KoRP corpus and Gigafida, a reference corpus of Slovene.

Keywords. Automatic term recognition, corpus terminography, extraction of lexical information, public relations.

1. The project

An applied research project titled *Terminology data banks as the bodies of knowledge: The model for the systematization of terminologies* (TERMIS, http://www.termis.fdv.uni-lj.si/) 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, with the intention of making it a model (a) that could be adopted by other disciplines in Slovenia, and (b) that could enable the compilation of terminological databases in a relatively short amount of time.

2. The corpus

The basis of the project was KoRP, a corpus of public relations texts (Logar 2013). The corpus contains 1.8 million words and is a monolingual and synchronous specialised corpus. The corpus has been made freely accessible online (Fig. 1) almost immediately after its completion, i.e. in July 2007. It is lemmatised and morphosyntactically tagged (Grčar, Krek & Dobrovoljc 2012). The texts in the corpus were selected according to carefully designed criteria (Logar 2013: 46-91), which make it representative of a public relations field in Slovenia.

Figure 1: Part of concordance for the term *komunikator* (communicator) in the KoRP corpus

G. Budin and V. Lušicky (eds.), Languages for Special Purposes in a Multilingual, Transcultural World. Proceedings of the 19th European Symposium on Languages for Special Puporses, 8-10 July 2013, University of Vienna, Vienna. Vienna: University of Vienna.

386

3. The headword list

One of the basic analyses of every corpus includes producing a list of words, which represents a point of departure for the identification of lexical items to be included in a (terminological) dictionary or database. Thus, a word list for each word class (adjectives, verbs, nouns) was first extracted from the KoRP corpus and items on the lists were examined, considering their terminological nature or their usefulness for the headword lists of public relations terminological database. Our definition of a term, specifically a public relations term, was: "all words or phrases which have special reference, regardless of the subject field to which they belong, and which may also form part of the lexicon of another subject field must be considered to be part of the terminology of that subject field" (Pearson 1998: 13, 87). In the case of a relatively new field that is still defining its scope – which public relations in Slovenia definitely is – one should avoid taking the opposing stance, i.e. to consider as candidates only terms that originate from the field in question (if such origin could even be determined).

The analysis of word lists revealed two things:

- that the list of adjectives is a good basis for identifying multi-word terms with relational adjectives as modifiers (that combined with the head of the phrase nearly always form a term in Slovene),

- that word lists are not sufficient to identify verb and noun terms.

The problem encountered with verb and noun terms has been also mentioned by Pearson (1998) who saw as the only way to distinguish terms from general vocabulary by looking for them in the text, considering different characteristics of communication (author, addressee, text type etc.). As automatic term extraction already uses this approach by combining various statistical methods with linguistic knowledge on terms (for Slovene Vintar 1999; 2003a; 2003b; 2010), we have adopted this procedure for extracting candidate terms.

Using the LUIZ term extraction tool (http://lojze.lugos.si/cgitest/extract.cgi; Vintar 2010) we have extracted from the KoRP corpus:

- single-word term candidates: nouns, verbs, adjectives, and adverbs;

- multi-word term candidates: noun phrases and verb phrases.

Both 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 general corpus, in our case FidaPLUS, a reference corpus of Slovene (http://www.fidaplus.net; Arhar Holdt & Gorjanc 2007) – as well as phraseological stability of the extracted terminological unit. We have identified 39 morphosyntactic patterns in total: 30 with a noun as a headword, 9 with a verb as a headword. The result of the extraction were lists with 47.007 multi-word units (excluding proper nouns) and 16.190 single-word units (excluding proper nouns).

The lists were carefully analysed and evaluated in order to determine the successfulness of the extraction method. When the top part of the list containing extracted term candidates was compared with the top parts of the noun and verb frequency lists in KoRP, we noticed only minor differences; however, they all favoured the lists of extracted terms. In other words, the lists with extracted terms offered better results. Our expectations were thus confirmed, so we decided to use only automatically extracted lists of term candidates for building our headword list. When creating a headword list 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 (see more in Logar Berginc, Vintar, Arhar Holdt 2012; Logar Berginc, Kosem 2013).

4. Entry contents

Entries in terminological dictionaries and databases contain different types of information. A detailed description of a microstructure of an entry in a normative explanatory terminological

G. Budin and V. Lušicky (eds.), Languages for Special Purposes in a Multilingual, Transcultural World. Proceedings of the 19th
European Symposium on Languages for Special Puporses, 8-10 July 2013, University of Vienna, Vienna. Vienna: University of Vienna.

387

dictionary with foreign language equivalents is available in Košmrlj Levačič (2006: 72-84):

> Each term has its own entry, consisting of lemma and explanatory part. Terms are normally shown with accents. Then, grammar information and any homonyms are provided. If a term is an abbreviation, the full form is also provided. Explanatory part can have a label that shows a specific field or subfield to which the term belongs. Then, an explanation or definition is offered. After the definition, the entry can contain other elements, such as synonyms, hypernyms, hyponyms, polysemy and partitiveness. Foreign language equivalents with synonyms are provided at the end of the entry.

The TERMIS project focussed on the parts of an entry in a terminological dictionary or database that can be improved by using a corpus-based approach and state-of-the-art lexicographic tools. This includes mainly contextual information, and to a smaller extent definitions and norm.

4.1. Automatic extraction of lexical information and examples of use

Contextual information is rarely found in terminological dictionaries (even in online dictionaries, e.g. Caruso 2011), and two rubrics are relevant in this case: collocations and examples.

4.1.1. Collocations

"Collocations are lexically and/or pragmatically constrained recurrent cooccurrences of at least two lexical items which are in a direct syntactic relation with each other" (Heid & Gouws 2006: 980). This notion is well-known in English lexicography (e.g. Firth 1957; Halliday 1966; Church, Hanks 1990; Sinclair 1991; Krishnamurthy 2004). Teubert's (2005/1999: 106) refers to collocation by saying that corpus linguistics can contribute particularly to the area of lexico-grammar. The author (ibid.:113-114) says that corpus linguistics, offering statistical information on co-occurrence of words, can provide much better information on semantic cohesion between collocations, as opposed "classic" linguistics. Until large amounts of data could be processed systematically, there was no other possibility to describe co-occurrence of words other than using grammar rules. Key is thus in identification of repetitive segments of text, and the condition for this is a large enough corpus.

Including collocational information on headwords has become an integral element of contemporary corpus based lexicography (e.g. Čermák 2006), whereas terminography is yet to make this information a regular dictionary feature. The TERMIS project aimed to build a body of knowledge, not merely a dictionary, therefore we decided to include lexically and/or pragmatically constrained recurrent co-occurrences of terms, and terms and other lexemes.

Due to the fact that we collaborated on the TERMIS project, as well as the *Communication in Slovene* project (http://www.slovenscina.eu/projekt), where a lexical description of contemporary Slovene has been produced (http://www.slovenscina.eu/spletni-slovar/leksikalna-baza; Gantar, 2009; Gantar & Krek 2011), we used the same method in the TERMIS project for extracting lexical information (syntactic relations, collocations, and examples) for single and multi-word terms from the KoRP corpus. The method uses the Sketch Engine tool and its Word sketch function (http://www.sketchengine.co.uk/; Kilgarriff et al. 2004; Kilgarriff & Kosem 2012), so we had to prepare and upload the KoRP corpus in our local installation of the Sketch Engine. Some changes had to be made to the extraction algorithm and its constituent parts. For example, Sketch Grammar had to be slightly adapted (Krek 2012) and minor tweaks to API script (Application Programming Interface) had to be made (Kosem, Gantar & Krek 2012; Kilgarriff et al. 2008; Kosem, Husak & McCarthy 2011). In addition, a new DTD for the Termania dictionary portal (http://www.termania.net; Romih & Krek 2012) had to be prepared to enable importing of information in the database, as well as its visualisation.

Fig. 2 shows a partial word sketch for the term communicator with the grammatical structure

G. Budin and V. Lušicky (eds.), Languages for Special Purposes in a Multilingual, Transcultural World. Proceedings of the 19th European Symposium on Languages for Special Puporses, 8-10 July 2013, University of Vienna, Vienna. Vienna: University of Vienna.

388

"adjective + communicator", Fig. 3 shows its incorporation into the terminological database and its final visualization at the Termania web portal.

Figure 2: Partial word sketch for *komunikator* (communicator) in the KoRP corpus (the Sketch Engine)[1] English translations of adjectives: *vladen* = government, *poslovni* = business, *akreditiran* = accredited, *profesionalen* = professional, *poklicen* = business, *britanski* = British, *glaven* = head, *organizacijski* = organizing, *dober* = good, *slovenski* = Slovene." on page 391

Figure 3: Partial entry of the term *komunikator* (communicator) at the Termania web portal

4.1.2. Examples

Examples are included in dictionaries to confirm the existence of the word, to assist with understanding of the definition, and to exemplify syntactic, collocational, textual and other characteristics of the word (Atkins, Rundell 2008: 452–455).

As shown in Fig. 3, examples have been included in public relations terminological database, in two parts of the entry:

G. Budin and V. Lušicky (eds.), Languages for Special Purposes in a Multilingual, Transcultural World. Proceedings of the 19th European Symposium on Languages for Special Puporses, 8-10 July 2013, University of Vienna, Vienna. Vienna: University of Vienna.

389

○ after the English translation of the headword (two examples) and

○ under each collocation (two examples for each collocate).

As Kilgarriff and Kosem (2012: 46) say: "Good dictionary examples are hard to find" – if one needs to search for them in a very large corpus, the procedure becomes even more difficult and time-consuming; for this reason, the GDEX (Good Dictionary Examples) tool has been developed, when preparing the online version of Macmillan English Dictionary.

GDEX ranks corpus examples according to their dictionary potential by using criteria such as sentence length, whole-sentence form, sentence complexity, presence/absence of rare words, presence of URLs etc., and is therefore a very useful function for lexicographers (Kilgarriff et al. 2008; Kosem, Husak, McCarthy 2011; Kosem, Gantar & Krek 2012). Using GDEX, we extracted good examples for almost all 2000 terms in the database.

In addition, users of the database can access two corpora (Fig. 4): the KoRP corpus and Gigafida, the reference corpus of Slovene (http://www.gigafida.net; Logar et al. 2013). In the former, the users can see all the concordance lines of a term, and a wider context (each paragraph has the information on the text source), and in the latter corpus the users can see how a term is used in general language (majority of public relations terms are found in general language; e.g. *javnost (public), odnos (relation), organizacija (organisation); sporočiti (to send a massage), komunicirati (to communicate), izvajati (to perform); blagovna znamka (brand), lokalna skupnost (local community), neprofitna organizacija (nonprofit organisation)*).

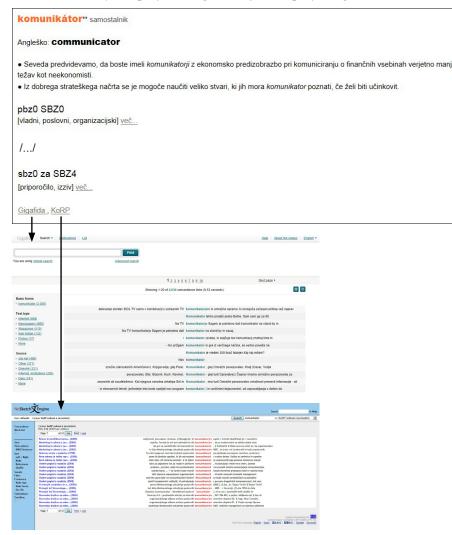Figure 4: Partial entry of the term *komunikator* (communicator) at the Termania web portal; and links to the KoRP corpus and the Gigafida corpus

G. Budin and V. Lušicky (eds.), Languages for Special Purposes in a Multilingual, Transcultural World. Proceedings of the 19th European Symposium on Languages for Special Puporses, 8-10 July 2013, University of Vienna, Vienna. Vienna: University of Vienna.

390

5. Conclusions

"No matter how many features are used to summarize the data, the lexicographer still needs to critically review the summary" (Kilgarriff & Kosem 2012: 48). The editing of the extracted data mainly included redistribution and grouping of semantically related collocates, identification of compounds, and moving and reordering of corpus examples. In rare cases, we had to re-examine the word sketch of the term and manually select another example.

Language technologies can certainly speed up the building of terminological databases. By using Word sketches and GDEX, we almost totally avoided manual corpus analysis. After initial preparations it only took us three hours to obtain all the lexical information for almost all 2000 terms. The use of lexicographic tools, described in this paper, has not only facilitated a quicker building of terminological database for the discipline of public relations, but also made the analysis more objective.

Research shows that collocations strengthen (terminological) definition and/or facilitate its understandability (Bergenholtz, Tarp 1995: 117-126, 141-142) – together with examples they enable quicker understanding of the concept of the lexeme (term). This can undoubtedly increase the informative value and usability of any language resource; terminological resources should not be an exception. By adding considerably more information on the public relations terms, as opposed to merely providing a short definition and English translation, we have developed a body of knowledge for the field. In order to be able to evaluate how successful we were in the preparation of this terminographical product, which is quite different to what the users have been used to so far, we intend to carefully monitor its use and user feedback.

6. Acknowledgements

7. Notes

[1] English translations of adjectives: *vladen* = government, *poslovni* = business, *akreditiran* = accredited, *profesionalen* = professional, *poklicen* = business, *britanski* = British, *glaven* = head, *organizacijski* = organizing, *dober* = good, *slovenski* = Slovene.

8. References

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

Atkins, Sue B. T., & Rundell, Michael (2008). *The Oxford guide to practical lexicography*. Oxford: Oxford University Press.

Bergenholtz, Henning, & Tarp, Sven, eds. (1995). *Manual of specialised lexicography*. Amsterdam, Philadelphia: John Benjamins.

Caruso, Valeria (2011). Online specialised dictionaries: a critical survey. *Proceedings of eLex 2011*. Ljubljana: Trojina, zavod za uporabno slovenistiko, 66-75.

Church, Kenneth Ward, & Hanks, Patrick (1990). Word associations norms, mutual information and lexicography. *Proceedings of the 27th annual conference of the Association for Computational Linguistics*. Vancouver: Association for Computational Linguistics, 76-82.

Čermak, František (2006). Collocatians, collocability and dictionary. *Proceedings of the 12th EURALEX international congress*. Torino: Edizioni dell'Orso, 929-937.

Firth, John Rupert. (1957). A synopsis of linguistic theory 1930–55. *Philological Society: studies in linguistic analysis* (spec. issue), 1-32.

G. Budin and V. Lušicky (eds.), Languages for Special Purposes in a Multilingual, Transcultural World. Proceedings of the 19th European Symposium on Languages for Special Puporses, 8-10 July 2013, University of Vienna, Vienna. Vienna: University of Vienna.

391

Gantar, Polona (2009). Leksikalna baza: vse, kar ste vedno želeli vedeti o jeziku. *Jezik in slovstvo*, *54*(3/4), 69-94.

Gantar, Polona, & Krek, Simon (2011). Slovene lexical database. In Daniela Majchraková & Radovan Garabík (eds.), *Natural language processing, multilinguality: 6th international conference*. Modra: Tribun EU, 72-80.

Grčar, Miha, Krek, Simon, & Dobrovoljc, Kaja (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, 89-94.

Halliday, Michael Alexander Kirkwood (1966). Lexis as a linguistic level. C. E. Bazell, J. C. Catford, M. A. K. Halliday, R. H. Robins (eds): *In memory of J. R. Firth*. London: Longman, 148-162.

Heid, Ulrich, & Gouws, Rufus H. (2006). A model for a multifunctional dictionary of collocations. *Proceedings of the 12th EURALEX international congress*. Torino: Edizioni dell'Orso, 979-988.

Kilgarriff, Adam, & Kosem, Iztok (2012). Corpus tools for lexicographers. In Sylviane Granger & Magali Paquot (eds.), *Electronic lexicography*. Oxford: Oxford University Press, 31-55.

Kilgarriff, Adam, Husák, Miloš, McAdam, Katy, Rundell, Michael, & Rychlý, Pavel (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, 425-432.

Kilgarriff, Adam, Rychlý, Pavel, Smrz, Pavel, & Tugwell, David (2004). The Sketch Engine. *Proceedings of the 11th EURALEX international congress*. Lorient: Universite de Bretagne-Sud, 105-116.

Kosem, Iztok, Gantar, Polona, & Krek, Simon (2012). Avtomatsko luščenje leksikalnih podatkov iz korpusa. In Tomaž Erjavec & Jerneja Žganec Gros (eds.), *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, 117-122.

Kosem, Iztok, Husak, Miloš, & McCarthy, Diana (2011). GDEX for Slovene. *Proceedings of eLex 2011*. Ljubljana: Trojina, zavod za uporabno slovenistiko, 150-159.

Košmrlj Levačič, Borislava (2006). O terminološkem slovarju in njegovi izdelavi z vidika terminografske ravnine. *Jezikoslovni zapiski, 12*(1), 71-87.

Krek, Simon (2012). New Slovene sketch grammar for automatic extraction of lexical data. *SKEW3*. Brno. Accessed at: http://trac.sketchengine.co.uk/wiki/SKEW-3/Program# (24 August 2013).

Krishnamurthy, Ramesh (2008). Corpus-driven lexicography. *International journal of lexicography, 21*(3), 231-242.

Logar Berginc, Nataša, & Kosem, Iztok (2013). TERMIS: A corpus-driven approach to compiling an e-dictionary of terminology. *Proceedings of eLex 2013*. Tallinn: Institute of the Estonian Language; Trojina, zavod za uporabno slovenistiko, 166-180.

Logar Berginc, Nataša, Grčar, Miha, Brakus, Marko, Erjavec, Tomaž, Arhar Holdt, Špela, & Krek, Simon (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, zavod za uporabno slovenistiko; Fakulteta za družbene vede.

Logar Berginc, Nataša, Vintar, Špela, & Arhar Holdt, Špela (2012). Luščenje terminoloških kandidatov za slovar odnosov z javnostmi. In Tomaž Erjavec & Jerneja Žganec Gros (eds.), *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, 135-140.

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

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

Romih, Miro, & Krek, Simon (2012). Termania – prosto dostopni spletni slovarski portal. In Tomaž Erjavec & Jerneja Žganec Gros (eds.), *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, 163-166.

Sinclair, John (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Teubert, Wolfgang (2005/1999). Korpusno jezikoslovje in leksikografija. In Vojko Gorjanc & Simon Krek (eds.), *Študije o korpusnem jezikoslovju*. Ljubljana: Krtina, 103-136./Korpuslinguistik und Lexikographie. *Deutsche Sprache* 4.

Vintar, Špela (1999). Računalniško podprto iskanje terminologije v slovensko-angleškem vzporednem korpusu. *Uporabno jezikoslovje*, 7-8, 156-169.

Vintar, Špela (2003a). *Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov: doktorska disertacija*. Ljubljana: Filozofska fakulteta.

Vintar, Špela (2003b). Kaj izvira iz jezikovnih virov. *Jezik in slovstvo, 48*(3-4), 77-88.

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

G. Budin and V. Lušicky (eds.), Languages for Special Purposes in a Multilingual, Transcultural World. Proceedings of the 19th European Symposium on Languages for Special Puporses, 8-10 July 2013, University of Vienna, Vienna. Vienna: University of Vienna.

392