COLLOCATIONS AND EXAMPLES OF USE: A LEXICAL-SEMANTIC APPROACH TO TERMINOLOGY

Nataša LOGAR
University of Ljubljana, Faculty of Social Sciences

Polona GANTAR
Fran Ramovš Institute of the Slovenian Language SRC SASA

Iztok KOSEM
Trojina, Institute for Applied Slovene Studies

Logar, N., Gantar, P., Kosem, I. (2014): Collocations and examples of use: a lexical-semantic approach to terminology. Slovenščina 2.0, 2 (1): 41–61.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2014/1/Slo2.0_2014_1_03.pdf.

The paper describes the compilation of an online terminological database that also includes a lexical-semantic framework of terms in the form of collocations and examples of use. Both types of information were extracted from a specialised corpus automatically, using Word Sketch and GDEX functions in the Sketch Engine corpus tool. Each entry contains links to two corpora: the LSP corpus of the public relations field KoRP and the Gigafida corpus, a reference corpus of Slovene. Preliminary results of the survey conducted among the target users of the terminological database indicate that the information on the term's typical collocations is very useful for fully understanding the term, its meaning and role in the context.

Key words: specialised corpus, terminological database, Sketch Engine, GDEX, user survey

1 INTRODUCTION

Lexis as an inventory of words in a language and a complex syntactically-semantic phenomenon has been at the forefront of lexicological

and lexicographical analyses, as well as general linguistic analyses, for at least three decades. The increased interest in this field is mainly the response to treating lexemes as fillers of independent syntactic models, governed by grammar (Chomsky 1965). Contemporary linguistic theories, especially the ones based on electronic corpora, i.e. large carefully designed collections of authentic texts in electronic form, are treating lexis as a central repository of language knowledge (Ooi 1998: 2).

The full potential of studying lexis, especially word relations, has been exploited by statistical methods in corpus linguistics that allow the extraction of relevant information about the regularity of a pattern or collocation, and about discourse, genre, textual and other characteristics of the lexis. One of the main principles of corpus linguistics is that natural languages contain the same amount of analogy and anomaly (Teubert 1999: 298) and that the search for universal structures of grammar and lexicon using generative grammar and cognitive semantics approaches cannot meet these two criteria effectively. Corpus linguistics therefore focuses on patterns and structures of semantic cohesion that exist in the area between word and sentence level, where a sentence is formed with grammar rules (Teubert 1999: 298–300). The latest corpus approaches, especially in lexicography, have moved the limits of lexical behaviour even further into the text, making it possible to identify its sociolinguistic and pragmatic characteristics, and to predict its future development by using larger and larger amounts of data and state-of-the-art tools for analysing them.

Even the earliest corpus analyses have paid particular attention to the notion of lexical unit as a semantic and syntactic phenomenon. These principles were especially thoroughly implemented in Collins COBUILD corpus-based dictionary projects. The arrival of these dictionaries has completely revolutionized lexicography: any general, bilingual or specialized dictionary that wanted to be state-of-the-art and reasonably useful, had to use corpus methodology. Further steps in connecting semantic and syntactic

characteristics of words, based on language databases and used in both lexicography and natural language processing, derive from a common Firthian and structuralist source, but disperse in various directions. Especially noteworthy are developments in word sense disambiguation, development of tools for corpus analysis, e.g. *the Sketch Engine* with advanced functions such as word sketches (Kilgarriff, Rundell 2002; Kilgarriff et al. 2004), and establishing "behaviour profiles" using quantitative corpus semantics, including sociolectometrical trend analysis and usage based linguistics (Geeraerts 2010).

Contemporary lexicographic approaches are slowly finding their way into terminography where we are witnessing an increase in the importance of lexical semantics, i.e. extraction of definitions from texts and of paradigmatic and/or syntagmatic relations between terms and between terms and other words. As Faber and L'Homme argue (2013, see also Pearson 1998):

> /T/he advent of corpus linguistics and corpus pattern analysis has brought many questions to the forefront in Terminology, such as term variation and polysemy, which were previously not envisaged in specialized language. Other issues include the identification of specialized meaning in running text, as well as the relations between terms and other lexical units. As a result, terminologists now have to deal with term meaning and how it is represented in texts.

In this paper we describe how some of the aforementioned approaches were used in the compilation of a freely available online terminological database of Slovene public relations terms.

2 THE PROJECT

An applied research project titled *Terminology data banks as the bodies of knowledge: The model for the systematization of terminologies* (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 of public relations terms called TERMIS (http://www.termania.net/,

Logar Berginc 2012). The database includes 2,000 entries with definitions, translation equivalents of headwords in English and contextual information (collocations and examples). Each entry also contains links to a specialised corpus of public relations texts called KoRP (Logar 2013) and to the reference corpus of Slovene language Gigafida (http://www.gigafida.net; Logar Berginc et al. 2012).

From the linguistic and technological aspect, three project features can be stressed: a corpus of public relations texts, which represents the entire discipline of public relations in Slovenia; the automatic extraction of terminological candidates from the corpus; and the automatic extraction of grammatical and collocation information for each term, together with good dictionary examples (Logar 2013; Logar Berginc, Vintar, Arhar Holdt 2012; Logar Berginc, Vintar, Arhar Holdt 2013; Logar Berginc, Verčič 2013; Logar Berginc, Kosem 2013). In this paper we focus on the latter: the lexical-semantic framework of terms presented in the database.

3 TERMINOLOGY DATABASE AS A BODY OF KNOWLEDGE

As established earlier, the inclusion of collocation 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 with terms and other lexemes. This information was included in the form of two elements: as collocations under the relevant grammatical relation, and as corpus examples.

3.1 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 linguistics and lexicography (e.g. Firth 1957; Halliday 1966; Church, Hanks 1990; Sinclair 1991; Krishnamurthy 2004).

This definition of collocation indicates that collocations are semantically transparent, their meaning is usually a combination of meanings of their components; they are normally syntactically acceptable, i.e. they follow grammatical rules, however they exhibit certain restrictions in their grammatical and lexical selection. Collocations can be divided into two groups: nominal collocations, consisting of two content words, and grammatical collocations, especially prepositional collocations (see Sicherl 1999; Benson et al. 1986). When determining the scope of collocation or so-called collocational paradigm (Čermák 1985: 173), which is defined by the set of a word's collocations, different perspectives of word relations on the syntagmatic level are combined with semantic relations between words at the paradigmatic level. In other words, this phenomenon, which exhibits a strong relation in the corpus, also has semantic properties.

In Slovene, typical collocates of nouns are adjectives, nouns and verbs; typical collocates of adjectives are adverbs and nouns; verb collocates fill valency positions or modify the verb (Gorjanc et al. 2005: 11). Grammatically relevant collocates for Slovene are prepositions. Considering those facts and our priority to make the extraction of collocations from the corpus as automatic as possible, we used the Sketch Engine tool and its Word sketch function (http://www.sketchengine.co.uk/; Kilgarriff et al. 2004; Krek, Kilgarriff 2006; Kilgarriff, Kosem 2012; Krek 2012; Figure 1), which requires a lemmatized and morphosyntactically tagged corpus, to extract typical grammatical relations, defined in the sketch grammar, collocations in those relations and their examples.

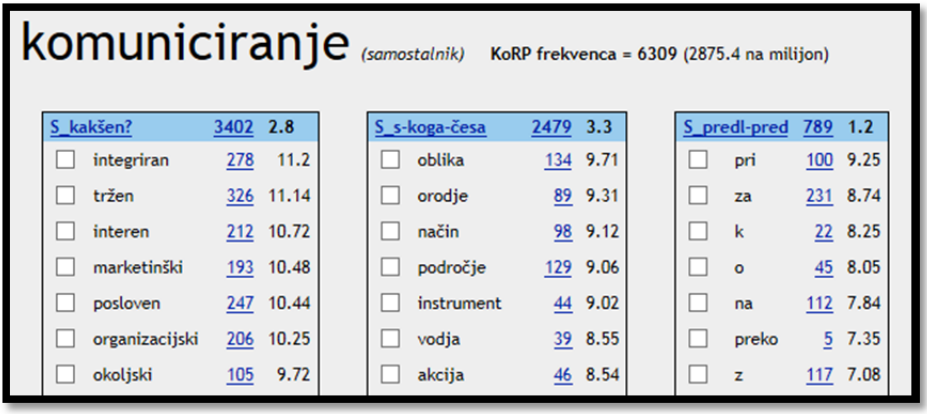

Figure 1: Partial word sketch for *komuniciranje ('communication')* in the KoRP corpus (the Sketch Engine).

The results of the automatic procedure, described in more detail in Logar Berginc, Kosem (2013), were XML files (Figure 2) containing 462 lexical-grammatical sketches for noun terms, 58 sketches for verb terms, and 718 sketches for multi-word terms (adjective + noun, noun + noun). For 479 noun terms, 141 verb terms and 122 multi-word terms we extracted all corpus data as there was not enough data available to create word sketches.

Figure 2: Partial XML export of the word sketch for *komuniciranje ('communication')* in the KoRP corpus.

Once the collocation information was imported into the terminological database, it was manually edited. This procedure was limited to few activities: putting the collocations into the correct case (they were extracted as lemmas), splitting and merging semantically related collocations, and consequently re-ordering the corpus examples (more on examples below). We also had to delete some "false" collocations, as they were exemplified by only one example or by two or more identical examples, which was caused by the repeated occurrence of textual elements in the corpus (book titles, institutions etc.).

Collocations were listed in the database under the relevant grammatical structure, as shown in Figure 3.

Figure 3: Part of the entry for the term *komuniciranje ('communication')* in the terminological database of public relations (www.termania.net): collocations.

Figure 3 shows the top quarter of the entry *komuniciranje*, which in total contains 25 different grammatical structures. The formulae *pbz0 SBZ0*, *sbz0 SBZ2* etc. denote the structure of collocations; upper case is used for the headword (i.e. the term) and lower case for the collocation. Thus, *pbz0 SBZ0* means that the term *komuniciranje* (SBZ), which can appear in any case (thus

0), is preceded by an adjective or adjectival phrase (pbz), which can also appear in any case; the formula *sbz0 SBZ2* means that the term *komuniciranje* appears in the genitive, i.e. the second case (thus *2*), and is preceded by a noun or noun phrase in any case; the formula *SBZ1 gbz* means that the term *komuniciranje* appears in the nominative, its collocate being a verb or verb phrase (gbz).

Each collocation in the database is exemplified by two automatically extracted corpus examples.

3.2 Examples of use

Examples of use show the headword in its syntactic environment. They are authentic examples as opposed to invented ones. 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). If it has been said that terminological dictionaries rarely contain collocations, this is even more true of authentic examples, as the implementation of such a concept requires a corpus-driven approach.

Part of the method for extracting lexical information with the help of the Sketch Engine tool is the GDEX tool. 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 et al. 2011; Kosem et al. 2012).

Each collocation is exemplified with two examples. We used two settings for minimum collocation frequency: the frequency of 3 was used for verbs with frequency higher than 200 (there were 20 such verbs), for nouns with frequency higher than 700 (there were 38 such nouns), and for multi-word noun terms with frequency higher than 130 (there were 155 such nouns). Higher frequency of a term also meant more examples to choose from for the

GDEX function. In fact, in only about 10% of cases we decided to replace the extracted example by a manually selected one; but even in those cases we quite often found that there were no significantly different or better examples available in KoRP, as the authors quoted the same source, and also in the same or very similar manner.

Examples were not shortened as the online database format did not pose any restrictions, normally faced by lexicographers and terminologists working on printed dictionaries. The only modifications made to examples were deleting any non-final punctuation at the end, any numbers denoting footnotes, any redundant spaces before commas and full stops and around brackets and quotation marks (it can be assumed that these redundant spaces were caused by corpus annotation), and any extra spaces between words and adding missing spaces between words. All other "errors" in examples (e.g. typos, spelling errors, and inconsistencies) have been left uncorrected for now.

The users of the terminological database of public relations can access examples from the KoRP corpus by clicking on the *Več…* ('*More…*') link, which follows the group of collocations (see Figure 3). In Figure 4, showing the entry for *komuniciranje,* examples are opened for the second collocation group.

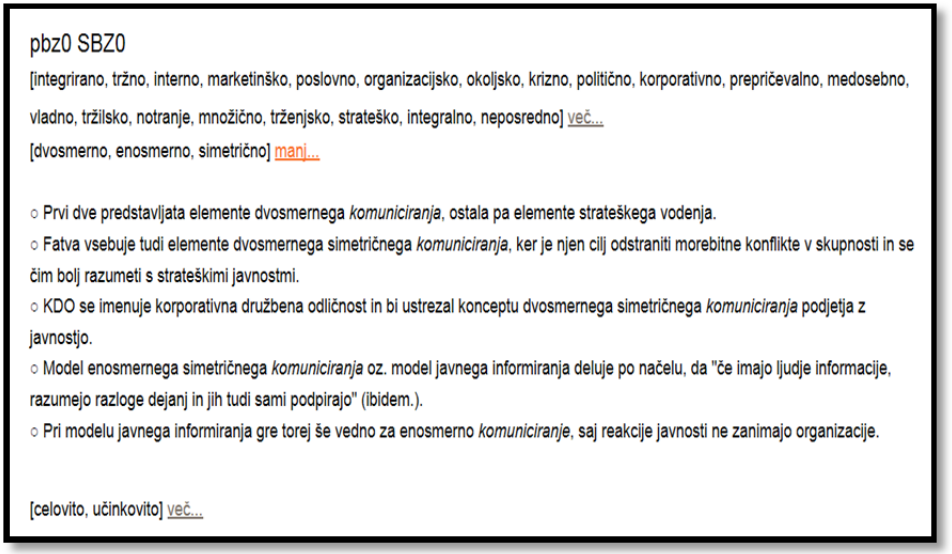

Figure 4: Part of the entry for the term *komuniciranje ('communication')* in the terminological database of public relations (www.termania.net): examples of use.

Collocations and good dictionary examples require an additional comment, namely that the corpus size of 1.8 million words resulted in certain limitations for creating word sketches for low frequency single- and multi-word terms. Therefore, approximately half of the headwords in the final database do not contain collocation information; however they are still exemplified by two corpus examples (the exception being the terms with only a single occurrence in the KoRP corpus, which contain only one example). It is of course also possible that a part of the terminological lexemes on the headword list does not form any relevant collocations, or features in grammatical structures with very diverse lexical elements.

3.3 Linking to other parts of the database, and to the Gigafida and KoRP corpora

The final part of each entry in the TERMIS database contains links to related entries (Figure 5), and as shown in Logar Berginc (2014), users of the

database can access two corpora: the reference corpus of Slovene Gigafida (http://www.gigafida.net; Logar et al. 2013) and the KoRP corpus in the NoSketch Engine and CUWI concordancer (Erjavec 2013). In the latter, 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 former corpus the users can see how a term is used in general language (a majority of public relations terms are found in general language as well).

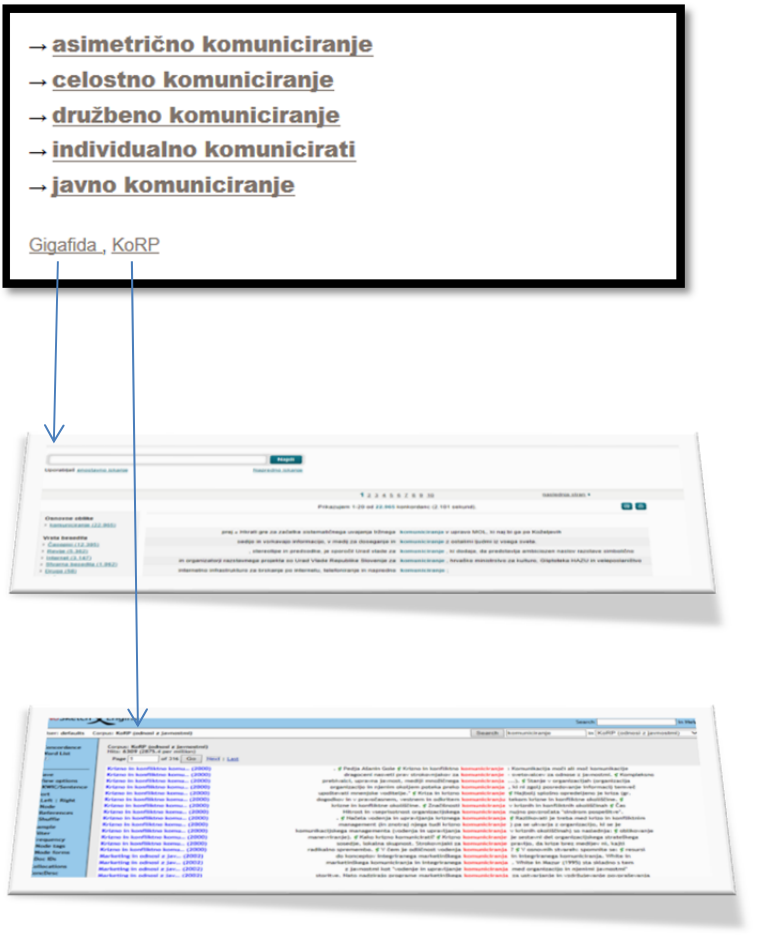

Figure 5: Part of the entry for the term *komuniciranje* ('*communication*') in the terminological database of public relations (www.termania.net): related terms, Gigafida, KoRP.

4 USER FEEDBACK: PRELIMINARY RESULTS

Users of terminological dictionaries are not used to seeing collocation information and full-sentence or multi-sentence examples, despite being able to read them online (as already mentioned, the terminological database of public relations also contains definitions and English equivalents of the terms; these elements are offered at the beginning of each entry, as these types of information are most frequently consulted).

Understandability, clarity and relevance of terms' collocation information in the TERMIS database is something that can be comprehensively measured after a certain period of usage, however during the compilation of the database we have already conducted a small survey about this part of the database entry among 24 Slovenian experts in public relations. The respondents were shown two types of display for the database entry, as planned at that time, and asked two multi-choice questions:

1. *After clicking on the More... link after the two examples, the users will be offered information on the term's typical context. Is this information shown in a clear and straightforward manner?*

A. Yes, one can quickly understand what the information means.

B. Yes, however one needs to get used to this way of presenting information.

C. Yes and no; certain information is clear and understandable, other is not.

D. Mostly no; it took me a long time to understand what this information means.

E. No, I don't understand at all what this information means.

2. *Do you consider the information on the term's typical context to be relevant for the terminological dictionary of public relations?*

A. Yes; all this information helps me fully understand the term, its meaning and role in context.

B. Yes and no.

C. No; it is enough to read only the first part of the entry (definition, translation, two examples).

Distributions of answers to the questions are shown in Figure 6 and Figure 7, respectively.

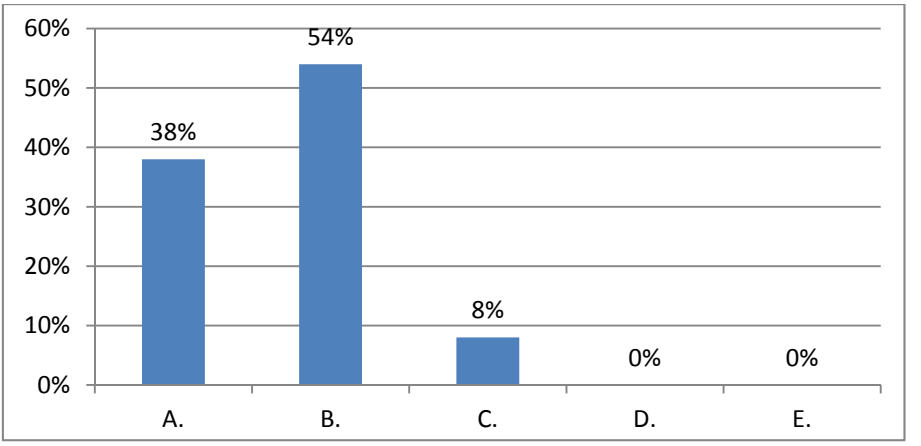

Figure 6: Distribution of answers to the survey question 1: *After clicking on the More... link after the two examples, the users will be offered information on the term's typical context. Is this information shown in a clear and straightforward manner?*

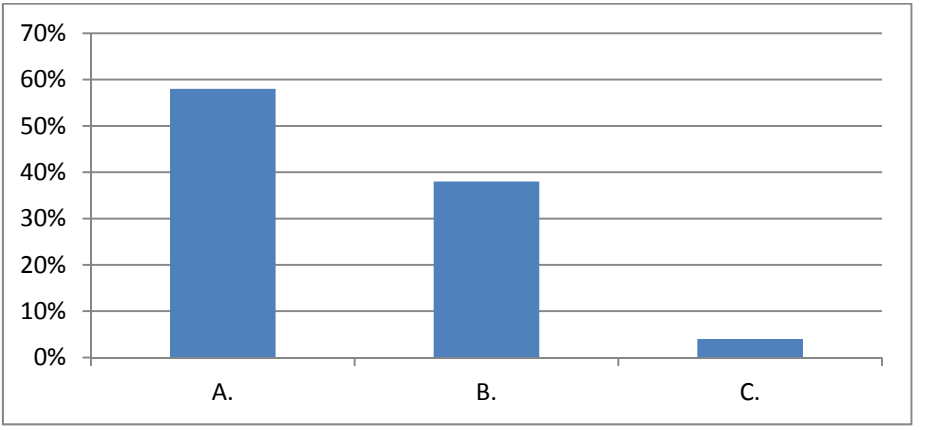

Figure 7: Distribution of answers to the survey question 2: *Do you consider the information on the term's typical context to be relevant for the terminological dictionary of public relations?*

The answers were encouraging as 38% and 54% of the respondents answered the first question with "Yes, one can quickly understand what the information means" and "Yes, however one needs to get used to this way of presenting information" in a terminological dictionary, respectively. The most frequently selected answer (58%) to the second question was "Yes; all this information helps me fully understand the term, its meaning and role in the context", followed by "Yes and no" (38%). The respondents' opinion that the collocation information and examples of use contribute to a better understanding of the terms confirms our assumption that the terminological database of public relations is a step away from traditional terminological dictionaries towards a dictionary that functiones as a body of knowledge.

5 CONCLUSIONS

Electronic (especially online) media offer different and better possibilities of including a variety of information in language resources. Terminological dictionaries are no exception. Wüster's General Theory of Terminology that sees a concept as a central phenomenon that can be described in detail and has a clear relation to other concepts, with denominations of those concepts – terms – carefully created in systematic manner, has been substantially developed and expanded over the years. One of the developments was that terms are not context-independent (Pearson 1998: 1–2). As soon as we accept the claim that

> In spite of extensive research in the field of terminology and in the field of sublanguages, there is no usable definition of term and no adequate communication model which allows us to identify when words are being used as terms. While we accept that there are indeed differences between words and terms, we find that, without human intervention, it is not possible to use any of the proposed definition of term as a means of distinguishing between terms and

> words. (Pearson 1998: 8)

we can apply to terminology several approaches of corpus lexicography, which is concerned with compiling general language dictionaries. One of such approaches, as shown in this paper, is the inclusion of information on the term's collocations. 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 (in our case, a term). This has been confirmed by the experts in the public relations field who participated in the survey on the understandability, clarity and relevance of the collocation information in the terminological database.

At the moment it appears that the TERMIS project has chosen the correct approach, and we will continue to carefully monitor user feedback to confirm this. There are already new trends on the horizon, for example:

> To cope with the challenge posed by the documentary and communicative explosion behind Big Data, the descriptive dictionary of the future should optimize the use of computational corpus techniques, and should consider the inclusion of longitudinal lexical analyses at aggregate level, complementing the traditional analyses at the level of the word. (Geeraerts 2014)

These are definitely approaches that might or should be transferred to and adapted for terminology.

REFERENCES

Atkins, S., and Rundell, M. (2008): *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Benson, M., et al. (1986): *The BBC Combinatory Dictionary of English: A Guide to Word Combinations*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Bergenholtz, H., and Tarp, S., eds. (1995): *Manual of Specialised Lexicography*. Amsterdam, Philadelphia: John Benjamins.

Chomsky, N. (1965): *Aspects of the Theory of Syntax*. Cambridge: The MIT Press.

Church, K. W., and Hanks, P. (1990): Word Associations Norms, Mutual Information and Lexicography. *Proceedings of the 27th Annual Conference of the Association for Computational Linguistics*: 76–82. Vancouver.

Čermák, F. (1985): Frazeologie a idiomatika. In F. Čermák and Josef Filipec (eds.): *Česká lexikologie*: 166–248. Praha: Academia.

Čermák, F. (2006): Collocations, Collocability and Dictionary. *Proceedings of the 12th EURALEX International Congress*: 929–937. Torino.

Erjavec, T. (2013): Korpusi in konkordančniki na strežniku nl.ijs.si. *Slovenščina 2.0*, 1 (1): 24–49.

Faber, P., and L'Homme, M.-C. (24 August 2013): *Call for Papers for Special Issue of Terminology on Lexical-semantic Approaches to Terminology*. [Corpora-List.]

Firth, J. R. (1957): A Synopsis of Linguistic Theory 1930–55. *Philological Society: Studies in Linguistic Analysis* (spec. issue): 1–32.

Geeraerts, D. (2010): *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Geeraerts, D. (2014): Corpus Linguistics in the Netherlands, 17th January, Leiden. Invited talk.

Gigafida, a reference corpus of Slovene. Available at: http://www.gigafida.net/ (16 January 2014).

Gorjanc, V., Krek, S., and Gantar, P. (2005): Slovenska leksikalna podatkovna zbirka. *Jezik in slovstvo*, L (2): 3–19.

Halliday, M. A. K. (1966): Lexis as a Linguistic Level. In C. E. Bazell et al. (eds.): *In Memory of J. R. Firth*: 148–162. London: Longman.

Hanks, P., and Pustejovsky, J. (2004): Common Sense About Word Meaning: Sense in Context. *TSD 2004*: 15–17. Brno.

Hanks, P., and Pustejovsky, J. (2005): A Pattern Dictionary for Natural Language Processing. *Revue Francaise de linguistique appliquée* 10: 2.

Hanks, P. (1994): Linguistic Norms and Pragmatic Exploitations, or Why Lexicographers Need Prototype Theory and Vice Versa. In F. Kiefer, G. Kiss and J. Pajzs (eds.): *Papers in Computational Lexicography: Complex '94*: 89–113. Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences.

Heid, U., and Gouws, R. H. (2006): A Model for a Multifunctional Dictionary of Collocations. *Proceedings of the 12th EURALEX International Congress*: 979–988. Torino.

Kilgarriff, A., and Kosem, I. (2012): Corpus Tools for Lexicographers. In S. Granger and M. Paquot (eds.): *Electronic lexicography*: 31–55. Oxford: Oxford University Press.

Kilgarriff, A., and Rundell, M. (2002): Lexical Profiling Software and its Lexicographic Applications: Case Study. *Proceedings of the 10th EURALEX International Congress*: 807–819. Copenhagen.

Kilgarriff, A., et al. (2004): The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*: 105–116. Lorient.

Kilgarriff, A., et al. (2008): GDEX: Automatically Finding Good Dictionary Examples in a Corpus. *Proceedings of the 13th EURALEX International* Congress: 425–432. Barcelona.

KoRP corpus. Available at: http://nl.ijs.si/noske/sl-spec.cgi/first_form?corpname=korp_sl (29 January 2014).

Kosem, I., Gantar, P., and Krek, S. (2012): Avtomatsko luščenje leksikalnih podatkov iz korpusa. In T. Erjavec and J. Žganec Gros (eds.): *Zbornik Osme konference Jezikovne tehnologije*: 117–122. Ljubljana: Institut "Jožef Stefan".

Kosem, I., Husak, M., and McCarthy, D. (2011): GDEX for Slovene. *Proceedings of eLex* 2011: 150–159. Ljubljana.

Krek, S., and Kilgarriff, A. (2006): Slovene Word Sketches. In T. Erjavec and

J. Žganec Gros (eds.): *Jezikovne tehnologije*: 62–67. Ljubljana: Institut "Jožef Stefan".

Krek, S. (2012): New Slovene Sketch Grammar for Automatic Extraction of Lexical Data. *SKEW3*. Brno. Available at: http://www.sketchengine.co.uk/documentation/wiki/SKEW-3/Program# (24 January 2014).

Krishnamurthy, R., ed. (2004): *English Collocations Study: the OSTI Report*. London: Continuum.

Logar Berginc, N., et al. (2012): *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Ljubljana: Trojina, Institute for Applied Slovene Studies; Faculty of Social Sciences.

Logar Berginc, N., Vintar, Š., and Arhar Holdt, Š. (2012): Luščenje terminoloških kandidatov za slovar odnosov z javnostmi. In T. Erjavec and J. Žganec Gros (eds.): *Zbornik Osme konference Jezikovne tehnologije*: 135–140. Ljubljana: Institut "Jožef Stefan".

Logar Berginc, N., Vintar, Š., and Arhar Holdt, Š. (2013): Terminologija odnosov z javnostmi: korpus – luščenje – terminološka podatkovna zbirka. *Slovenščina 2.0*, 1 (2): 113–138.

Logar Berginc, N. (2012): Slovene Terminologies / TERMIS: Preserving Slovene Terminology in a Globalising World. *International Innovation: Disseminating Science, Research and* Technology: 79–81. Bristol: Research media.

Logar Berginc, N. (2014): A Corpus Based E-dictionary of Terminology as a Body of Knowledge. *Languages for Special Purposes in a Multilingual, Transcultural World: Proceedings of the 19th European Symposium on Languages for Special Purposes*: 386–392. *Vienna*.

Logar, N., and Kosem, I. (2013): TERMIS: A Corpus-driven Approach to Compiling an E-dictionary of Terminology. *Proceedings of the eLex 2013 Conference*: 164–178. Ljubljana; Tallinn.

Logar, N., and Verčič, D. (2013): Terminological Databanks as the Bodies of Knowledge: Slovenian Public Relations Terminology. *Public Relations*

Review, 39 (5): 569–571.

Logar, N. (2013): *Korpusna terminografija: primer odnosov z javnostmi*. Ljubljana: Trojina, Institute for Applied Slovene Studies; Faculty of Social Sciences.OOI, B. Y. V. (1998): *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

Pearson, J. (1998): *Terms in Context*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Sicherl, E. (1999): *Predložne kolokacije v slovenščini in angleščini: Doktorska disertacija*. Ljubljana: Filozofska fakulteta.

Sinclair, J. (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Termania, dictionary web portal. Available at: http://www.termania.net/ (22 December 2013).

TERMIS. Available at: http://www.termis.fdv.uni-lj.si/ (20 April 2014).

Teubert, W. (1999): Korpuslinguistik und Lexikographie. *Deutsche Sprache* 4: 292–313.

KOLOKACIJE IN ZGLEDI RABE: LEKSIKALNO-SEMANTIČNI PRISTOP K TERMINOLOGIJI

Prispevek prikazuje pripravo spletno dostopne terminološke podatkovne zbirke, v kateri so v obliki kolokacij in primerov rabe vključeni tudi podatki o leksikalno-semantičnem okolju terminov. Kolokacije in primere rabe smo pridobili iz korpusa strokovnih besedil s pomočjo aplikacije Besedne skice in funkcije GDEX v orodju Sketch Engine. Vsak geselski članek je povezan z dvema korpusoma: korpusom besedil odnosov z javnostmi KoRP in referenčnim korpusom slovenščine Gigafida. Prikazani so tudi nekateri preliminarni rezultati anketne raziskave, izvedene med ciljnimi uporabniki terminološke podatkovne zbirke, ki kažejo, da so podatki o značilnem

besedilnem okolju terminov zelo uporabni za celostno razumevanje njihovih pomenov in njihovo ustrezno rabo v besedilu.

Ključne besede: korpus strokovnih besedil, terminološka podatkovna zbirka, Sketch Engine, GDEX, anketa med uporabniki