I. Kanič

SLOVENE SPECIALIZED TEXT CORPUS OF LIBRARY AND INFORMATION SCIENCE – AN ADVANCED LEXICOGRAPHIC TOOL FOR LIBRARY TERMINOLOGY RESEARCH

Abstract. To support the research in the field of library and information science terminology and dictionary construction in Slovene language a specialized text corpus has been designed and constructed. The corpus has reached 3,6 million words extracted from 625 Slovene technical and scientific texts of the field. It supports a variety of specialized search methods, display of search results, and their statistic computation. The web based application is in open public access.

Keywords: corpus linguistics, text corpus, library science, terminology, Slovene language

1. Introduction

The purpose of the project is to build a modern linguistic tool to effectively support the codification of the library and information science terminology in Slovene language, based on the inventory and evaluation of the current usage in modern technical and scientific texts in the field. The synchronous specialized text corpus was primarily designed and constructed to assist the work of the Commission on Library Terminology in the frame of the Slovene Library Association, which had already built a sample corpus of text fragments in the nineties (10.300 excerpts comprising half a million words) from texts published before 1999, but based on traditional hand-excerption, of course. The corpus is intended to the broader community of linguists and lexicographers, including librarians and students of Library and Information Science.

The synchronous specialized text corpus represents the technical language in the specific field, shared among the community of practicioners, researchers, translators, teachers and students in the present and very recent past. It helps to discover the exact inventory

and verify the occurrence of words and phrases in technical and scientific texts, enabling researchers to obtain a variety of structured lists of words and phrases, be it in their original form or lemmatized and tagged with part of speech labeling. It has proven an indispensable and powerful tool for the preparation of modern dictionaries and updating the existing monolingual explanatory[1] and multilingual translating dictionaries[2] of library terminology.

2. The scope and extent of the corpus

The preparatory one-year design project resulted in the web-based trial version of the corpus in July 2011 which already included all the intended functionalities. In 2012 the Ministry of Culture of Slovenia supported the extension of the project and above all inclusion of more than 400 scientific journal articles, exceeding thus an inventory of 3.6 million words, excerpted from 625 texts by 353 authors. All the included works have been originally published in electronic form, mostly born digital or digitized by publishers. Data capture was focused predominantly on texts published in the last decade, depending on their availability, of course. Original texts in Slovene language were chosen as a rule, translations are more an exception. A selective list of potentially interesting texts comprises some 500 additional units, they will be dealt with in 2013.

Table 1. Type and number of texts, their contribution in words

Type of publication	**Texts**	**Words**
Doctoral dissertation	4	215.000
Master theses	21	596.000
Graduate theses	17	319.000

[1] *Kanič I., Leder Z., Ujčič M., Vilar P., Vodeb G*. Bibliotekarski terminološki slovar. Ljubljana: Zveza bibliotekarskih društev Slovenije: Narodna in univerzitetna knjižnica, 2009.

[2] *Kanič I., Vilar P., Dimec Z*. Angleško-slovenski slovar bibliotekarske terminologije = English-Slovenian dictionary of library terminology. Ljubljana: Narodna in univerzitetna knjižnica, 2002. URL: **http://www2.arnes.si/~ljnuk4/dictionary/slovenian**

Monographic publications	10	207.000
Scientific journal articles	484	2.058.000
Technical journal articles	30	212.000
Total	625	3.661.000

A comprehensive list of 625 included texts with hyperlinks to original full texts is published in the project documentation on the web[3].

2.1. Software and application support

For the technical and structural preparation of texts, word indexing and tagging, concordances as well as the public accessible web application of the corpus a customized PC and web application EVA and its Internet version NEVA[4] were used. Their specific functions had been successfully operational for several years with the general reference corpus of Slovene language *Nova beseda*[5], the online version of the normative *Dictionary of Slovene Literary language*[6] and some other lexicographic and linguistic tools of the Slovene Academy of Science.

2.2. The use and functions of the corpus

The text corpus is offered to the public in open access as a web application and does not need any components to be downloaded to the user's computer. It is installed as a separate page of the blog Bibliotekarska terminologija[7] with basic description and user documentation including help and some findings of the analyses. The

---

[3] Seznami vključenih besedil: URL: **http://www.cek.ef.uni-lj.si/terminologija/Korpus/datoteke/seznam_besedil_si.html**

[4] *Jakopin P.* NEVA – Internet version of EVA. URL: **http://www.laze.org/neva/index.html**

[5] Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. Nova beseda. URL: **http://bos.zrc-sazu.si/a_beseda.html**

[6] Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. Slovar slovenskega knjižnega jezika. URL: **http://bos.zrc-sazu.si/sskj.html**

[7] Korpus bibliotekarstva. URL: **http://terminologija.blogspot.com/p/korpus.html**

user interface is simple and transparent, allowing some basic user settings and selection of mode and criteria.

User settings allow limiting the single word and concordance search to upper/lower case, truncation of a single or multiple word(s) in the query, and limiting the search to a specific document range, i.e. standard search performed across all the texts or restricted to one or several types of documents simultaneously (e.g. doctoral theses only).

Search procedures – The user interface allows for six different types of searches, special indexes have been prepared respectively. The default search procedure may be combined i.e. limited with other criteria, e.g. with the length of words and/or their frequency of occurrence.

Concordances – Search and display of the word(s) is performed in context with an indication of the full source text to which there is a direct hypertext link, so any text can be consulted in full immediately. The results of a query are shown in the form of a concordance list, the search string shown in the nearby context of the sentence not exceeding 100 characters.

- Standard search can consist of one or several words, any of them may be truncated. Only right truncation is allowed (symbol *).
- Each match is accompanied by an abbreviated bibliographic description of the document with a hyperlink to the original full-text on the server of its original publication.

Single word search

- Left and/or right truncation is allowed (symbol *).
- The results display an alphabetical list of the hits with an indication of the frequency of occurrence (no context).
- The next step allows the display of individual hits in the context and with an indication of the source which is hyperlinked.

Word pairs

- Search for one or both words in the word pair is allowed implementing the right truncation (symbol * may replace an entire

word as well). Useful in Slovene e.g. for the adjective + noun string as the adjective always precedes the noun.

- In the results lists word pairs occur with their respective frequencies, sorted by descending frequency. Hyperlinks to the original source text are enabled.

N-grams

- Search for N-grams (N=2, 3, 4 or 5) is allowed with any of the word(s) truncated (symbol * may replace an undefinite string of characters or a single word).
- In the results lists N-grams occur with their respective frequencies, sorted by descending frequency. Hyperlinks to the original source text are enabled.

3. Insight into the Corpus

The vocabulary itself, the diversity and frequency of individual words and their co-occurrence are reflecting the nature of the texts selected so far, therefore much is expected from the further growth and expansion of the corpus. The greatest richness and diversity of library and information science related terminology is expected in numerous scientific articles which are waiting to be digitized by the publisher Library Association of Slovenia shortly, and the academic works of the Library and Information Science Department at the University in Ljubljana (graduate, master and doctoral theses).

3.1. Word frequency

The analysis of the corpus comprising 625 Slovene technical and scientific texts in the field of library and information science, written by 353 authors of different age and scholarly level, altogether 3,6 million captured words, has completely confirmed the basic theoretical assumptions concerning text corpora and findings in the

Slovene language[8]. The words extracted may be categorized into three specific groups:

- *Very common words* which do not contribute to the presentation of the contents and meaning of the documents, among them also function words, that represent the very top frequency count; in this group there are relatively few different words but distinctly stand out with their high frequency (the absolute champion is the auxiliary verb *biti* (to be) with 172.031 occurrences, followed by the conjunction *in* (and) (120.870) and the preposition *v* (in, 93.847), etc. A very steep drop in frequencies is completely in accordance with the Zipf's Law; e.g. the frequency of the thirtieth most common word is already below ten thousand.
- *Very rare words*, including *hapax legomena* and personal names, which also do not represent the contents of the documents, prolonging into the long tail of frequency=1.
- A relatively narrow strip of words in the middle, representing the most important drivers of content and in our case rather potential candidates for the study and inclusion into the terminological dictionary.

Nevertheless, in conflict with the Zipf's law there are 13 for the library terminology relevant words among the first 100 most frequently represented words. This is due to the fact that the corpus represents a narrow and very specialized choice of technical language rather than the general everyday language.

The 3.660.900 words extracted from the texts need certain study and explanation to be correctly interpreted:

- In this context, a word is any string of characters delimited on both sides by a space, including numbers, paragraph headings, etc., so after an appropriate «cleaning» some 3.573.457 actual words remained.

---

[8] *Jakopin P.* Zgornja meja entropije pri leposlovnih besedilih v slovenskem jeziku. Doktorska disertacija. Ljubljana, 1999. URL: **http://www.ff.uni-lj.si/hp/pj/disertacija/**

- Taking into consideration their frequency and repetition less than 150,000 different forms resulted.
- Since Slovene is a highly inflected language, we implemented automatic lemmatization in order to group the different inflected forms of a word into a canonical form so they could be analysed as a single item. The word *knjižnica* (library) appears in 21 different forms (depending on the grammatical case and number, but also with the distinction of the upper and lower case). Thus the real number of different words was restricted to **28.808** only.
- It is necessary to take into account, however, that despite the «manual cleaning» a few foreign language words (e.g. from citations and notes) and names still remain as unwanted.

The frequency of use of individual words in the texts is very different, of course. In accordance with expectations the auxiliary verb *biti* (to be) leads with 172.031 occurrences, followed by other function words. Nevertheless, our technical termin *knjižnica* (library) *is the 6*th most common word with 48.214 occurrences, *knjiga* (book) on the 28th place with 11.876 occurrences, and *podatek* (data) the 34th most common word with 10.046 occurrences. Among the fifty most common words there are 13 full terms, the rest are function words. The occurrence of individual words declines abruptly from the most common (172.031), so only the first 35 most frequent words belong to the leading group with frequency above tens thousands, at the rank 500 the frequency reaches under one thousand (21.215 words) and the total count of *hapax legomena is* **7.310**.

3.2. Parts of speech

Automatic part of speech tagging[9] has identified 13.128 nouns, 7.064 adjectives, 6.460 verbs and 3.877 adverbs, and altogether a hundred prepositions, numerals, conjunctions and pronouns. It has to be stressed that these are estimates only since the automatic part of speech tagging still cannot discern particular forms of homographs

[9] Oblikoslovni označevalnik za slovenski jezik. URL: http://**www.slovenscina. eu/ tehnologije/oznacevalnik**

without human intervention (e.g. *dela* may be a form of the verb *delati*, a noun *delo* or *del*; *uporabnikov* may be a noun or an adjective, etc.). Resulting from the lemmatization[10] and part of speech tagging there were 13.074 words recognized as possibly belonging to two or more part of speech categories.

4. Conclusion

Even though the corpus already covers a wide range of different types of documents ranging from doctoral dissertations and scientific articles to conference papers and has reached a rather wealthy selection of words used by some 353 Slovene authors, a dynamic growth of the corpus by inclusion of recently published texts within the wider field of library science remains a further goal. Thus the corpus will gain the representative role in the inventory and study of the library terminology and further lexicographic work. The vast professional and scholarly community may profit from its open web access.

[10] Inštitut za slovenski jezik Frana Ramovša ZRC SAZU. Določevanje osnovnih besednih oblik (lem) in besednih vrst ali oblikoslovnih oznak. URL: **http://bos.zrc-sazu.si/dol_lem1.html**