FREQUENCY OF TECHNICAL TERMS AS OPPOSED TO COMMON WORDS IN THE SPECIALIZED TEXT CORPUS OF LIBRARY AND INFORMATION SCIENCE IN SLOVENE LANGUAGE

Ivan Kanič
University of Ljubljana
(Ljubljana, Slovenia)

Abstract: Frequency of technical terms as opposed to common words has been studied in a synchronous specialized text corpus in the field of library and information science, comprising 3.6 million words. The corpus had been designed to support the research in the field of library and information science terminology and dictionary construction in Slovene language.

Keywords: corpus linguistics, text corpus, word frequency, library science, terminology, Slovene language.

1. Introduction

The synchronous specialized text corpus represents the technical language in the specific field of library and information science, shared among the community of practitioners, researchers, translators, teachers and students in the present and very recent past in Slovene language. It helps to discover and determine 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. Insight into the occurrence of words, their combinations and frequency in technical and scientific texts support immensely the work of terminologists, thus the corpus 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.

[1] *Kanič I., Leder Z., Ujčič M., Vilar P., Vodeb G*. Bibliotekarski terminološki slovar. Ljubljana: Zveza bibliotekarskih društev Slovenije: Narodna in univerzitetna knjižnica, 2009. ISBN 978-961-6162-55-5.
URL: http://www.termania.net/slovarji/85/bibliotekarski-terminoloski-slovar

[2] *Kanič I., Vilar P., Dimec Z*. Angleško-slovenski slovar bibliotekarske terminologije = English-Slovenian dictionary of library terminology. Ljubljana: Narodna in univerzitetna knjižnica, 2002. ISBN 961-6162-53-5. URL: http://www2.arnes.si/~ljnuk4/dictionary/slovenian

2. The specialized corpus of library and information technical texts in Slovene language

The purpose of the project is to build and update 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 (mere 10.300 excerpts comprising half a million words) from texts published before 1999, but based on traditional hand-excerption then, of course, but processed digitally. The corpus is intended to the broader community of linguists and lexicographers, including librarians and students of Library and Information Science.

2.1. The scope and extent of the corpus

The web-based trial version of the corpus was launched in 2011 and upgraded in 2012 with the financial support of the Ministry of Culture of Slovenia which enabled additional processing and massive text input, the corpus exceeding thus an inventory of 3.6 million words now, 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, rare translations are an exception. A comprehensive list of 625 included texts with hyperlinks to original full texts is published in the project documentation on the web[1].

Type of publication	**Texts**	**Words**
Doctoral dissertations	4	215.000
Master theses	21	596.000
Graduate theses	17	319.000
Monographic publications	10	207.000
Scientific journal articles	484	2.058.000
Technical journal articles	30	212.000
Total	**625**	**3.661.000**

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

2.1.1. 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 can be consulted on mobile devices as well. It is installed as a separate page of the linguistic blog *Bibliotekarska terminologija*[2] with basic description and user documentation including help and some findings of the analyses. The user interface is simple and transparent, allowing some basic user settings and a selection of modes and criteria:

- **User settings** allow limiting the single word and concordance search to upper/lower case, truncation, and limiting the search to a specific document type range, i.e. standard search

[1] URL: http://www.cek.ef.uni-lj.si/ terminologija/Korpus/datoteke/seznam_besedil_si.html

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

performed across all the texts or restricted to one or several types of documents simultaneously (e.g. doctoral theses only, scientific articles only).

- **Search procedures –** The user interface allows for six different types of searches, special indexes have been prepared respectively. The simple 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. Each match is accompanied by an abbreviated bibliographic description of the document and a hyperlink to the original full-text on the server of its original publication.
- **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.
- **Word pairs -** Search for one or both words in the word pair is allowed implementing the right and/or left 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.
- **N-grams -** Search for N-grams (N=2, 3, 4 or 5) is allowed with any of the word(s) truncated. 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 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 Slovene language (Jakopin, 1999). The 3,660,900 words extracted from the texts need certain study and explanation to be interpreted correctly:

- In this context, a word is any string of characters delimited on both sides by a space, or a space and a punctuation mark, including numbers, paragraph headings, etc., so after an appropriate "cleaning" some 3,573,457 actual words remained.
- 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.

3.1. Parts of speech

Automatic part of speech tagging 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 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 and part of speech tagging there were 13,074 words recognized as possibly belonging to two or more part of speech categories.

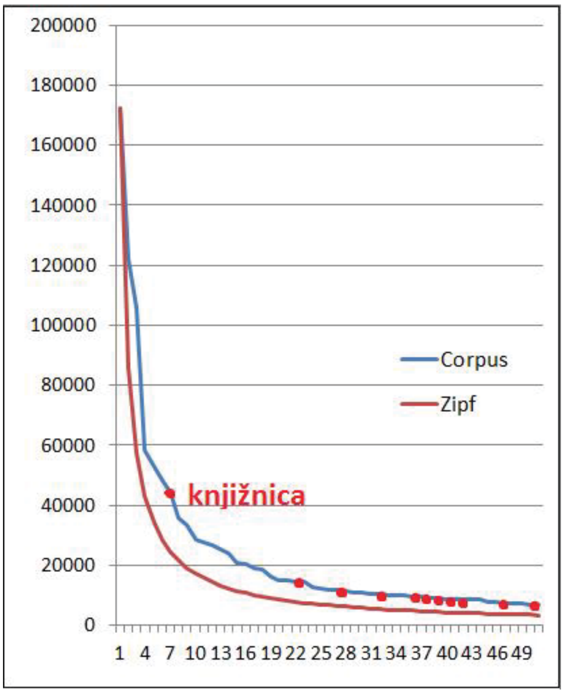
3.2. Word frequency

Regarding their frequency of occurrence the words extracted in the corpus 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. An exception are a few basic terms of the library terminology, ranking very high according to their frequency (q.v. Table 2 and Graph 1).

Rank	Term	*English translation*	Frequency
6	knjižnica	*library*	48214
23	gradivo	*materials*	14348
27	knjiga	*book*	11876
33	podatek	*data*	10046
36	uporabnik	*user*	9568
37	sistem	*system*	9510
38	informacijski	*information* (adj.)	9060
39	leto	*year*	8878
40	knjižničen	*library* (adj.)	8820
41	informacija	*information*	8807
47	delo	*work*	7343
50	vir	*source*	6828
51	zbirka	*collection*	6717

Table 2: Top ranked technical terms belonging to the library terminology.



Graph 1: Frequency of words compared to the theoretical Zipf curve. Top ranked technical terms are marked.

- *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 of library and information science.

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 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* 6th most common word with 48,214 occurrences, *knjiga* (book) on the 27th place with 11,876 occurrences, and *podatek* (data) the 33rd 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 thousand, at the rank 500 the frequency reaches under one thousand (21,215 words) and the total count of *hapax legomena* is **7,310**.

4. Conclusion

The specialized technical 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 Department of Library and Information Science and Book Studies at the University in Ljubljana (graduate, master and doctoral theses).

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 and information science remains a further goal. Word occurrence analysis will be upgraded to visualisation and word cloud presentation. Thus the corpus will gain the representative role in the inventory and study of the Slovene library terminology and further lexicographic work.

References

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/

Каніч І. Словенський одномовний “Словник бібліотечної термінології” тлумачного типу. Українська термінологія і сучасність. Збірник наукових праць. Інститут української мови НАНУ, 2013. pp. 119-147. URL: http://www.term-in.org/images/img/pdf/utis_2013.pdf

Kanič I. Slovene specialized text corpus of Library and Information Science – An advanced lexicographic tool for library terminology research., 2013. In International Conference "Corpus Linguistics – 2013", St. Petersburg , 25 – 27 June 2013. [Conference paper, ISBN 978-5-8465-1335-8], pp. 52-59.