Knowledge Organization in the New Era
Using DIY Corpora as Writing Assistants

Abstract:
In academic contexts, students and researchers are expected to follow a set of strict conventions that are strange or unknown to them, especially when their dominant language is not English. This paper aims to show how tools such as the Corpógrafo (Maia and Sarmento 2005), a web-based environment for corpora research (DIY - Do it Yourself corpora), can be used to provide writing assistance, in this particular case on the domain of Knowledge Organization (KO). To conduct this study, we used a corpus of articles published in Knowledge Organization, the official bi-monthly journal of ISKO founded in 1973. The corpus is composed of articles published between 2016 and 2018 in the sections ‘Articles’ and ‘Reviews of Concepts’ in Knowledge Organization. The articles were processed in Corpógrafo, in order to enable results to be extracted. Given the results obtained, the relevance of using Corpógrafo as a writing assistance tool for the KO scientific field is discussed, in particular in the areas of: (a) terminology extraction; (b) retrieval of collocations and identification of specialized language phraseology.

1.0 Introduction
In academic contexts, students and researchers are expected to follow a set of (more or less) strict writing conventions that are often alien to them, especially when English is not their dominant language. This demands a command of both the general language (LGP), i.e. is the language used in routine daily interactions, and of the Language for Special Purposes (LSP), used to exchange and disseminate specialized knowledge, and hence facilitate specialized communication. Drafting texts in LSPs thus demands a command of the general language (as any LSP builds upon the general language grammar and lexis), but primarily of the specific conventions of that particular LSP. Appropriate information searching is thus crucial to assist users, as it allows them to produce texts using correct conventions, as well as the right linguistic form and contents. The main advantage of writing assistants, to which a significant volume of research has been devoted (e.g. Bourelle, Bourelle, and Rankins-Robertson 2015; Tarp, Fisker, and Sepstrup 2017), is that they offer novice and technical writers the support to produce texts on a particular topic. Therefore, LSP texts show some particular features: (a) they are terminology-rich (i.e. they contain a high density of specialized terminology, including formulae, acronyms and abbreviations, among others); (b) they contain particular word combinations (collocations); and (c) they use grammar differently from LGP (e.g. simple and short sentences, impersonal language, few adjectives, and word repetition to the detriment of co-referencing in an attempt to preserve accuracy, concision and precision, while avoiding ambiguities). Thus, tools that are commonly used in general language writing assistants, such as word processor spell and grammar checkers, can be useful to get the general language spelling and grammar right, but will hardly help them produce a text in a particular LSP.

There has been a growing interest in DIY (‘do it yourself’) corpora for LSP. These are composed of small to medium scale collections of texts, whose choice relies entirely on the user's specific needs within a given discipline (Maia 1997). The features offered by corpus linguistics, such as word lists and concordances, are particularly useful, as together they can show the most frequent words and word collocations, provide
information about the meaning of a word, detail the use of a word when combined with (an)other word(s), and provide a context for a certain word. Arguably, thus, DIY corpora can be used as a writing assistant, hence serving as a tool to improve LSP writing skills. This paper addresses writing assistance as a methodology, more than as a tool.

2.0 Objectives and methodology

Education for Knowledge Organization (KO) is a core, especially in LIS courses (Alajmi and Rehman 2016), including in countries where students are not native speakers of English. With the primary aim of providing students with a tool that allows them to explore specific vocabulary/terminology of KO, tests were run in the 2019 version of Corpógrafo (V5), a web-based suite of tools for collection and analysis of DIY corpora. This suite, which is freely available, allows text collection, corpus search (search for regular expression concordances, extract collocations, and produce frequency-based statistics - such as n-gram counts), information extraction (collect terminology, establish semantic relations, produce conceptual maps); knowledge-resource building (collect specific-domain glossaries, thesauri, terminological databases and ontologies), and categorized word-lists (Maia and Sarmento 2005). The tool has already been used for conceptual studies in the field of information science (Oliveira and Rodrigues 2011).

The tests run allowed the study of the linguistic specificities in KO, with a view to further its linguistic enquiry and the analysis of the evolution of this scientific area. In this context, the following features of Corpógrafo were used: (a) terminology extraction; (b) retrieval of collocations and identification of specialized language phraseology. The corpus used in this study builds on articles published in Knowledge Organization, the official bi-monthly journal of ISKO founded in 1973. It consists of 16 articles published between 2016 and 2018 in the sections ‘Articles’ and ‘Reviews of Concepts’ in KO. The terminology extraction was performed using the Terminology Database (TDB) feature, whereas the collocations and phraseology were obtained using the concordances search functions (KWIC, sentence search and window search). In order to establish the frequency counts of the words, we undertook a study of n-grams and thus have an overview of the text contents. This analysis allows the visualization of some prominent features of the corpus, e.g. systematic use of certain words or word strings that enable the identification of terms, collocations or writing styles. Finally, a list of the most frequent terms was created, which was used as the basis to select terms for retrieval of collocations and identification of specialized phraseology.

3.0 Results and discussion

The results presented in this section are based on searches performed over the KNOWA - Knowledge Organization Writing Assistant corpus (our own DIY corpus of highly specialized KO publications) using the Corpógrafo. Overall, these results show that the methodology proposed has a significant potential as writing assistant. A list of n-grams was produced in order to establish which words/word combinations are most frequent. Linguistic elements that were not relevant for the KO LSP were discarded. We manually selected 5 terms from the word list produced by Corpógrafo that are core to KO and often used to build more complex terms, based on two criteria: (1) they are

1 Available at https://www.linguateca.pt/corpografo/
highly technical terms; (2) they are some of the most frequent terms in KNOWA. These are: ‘classif’ and ‘index’ (using the * wildcard); ‘concept’; ‘domain’; and ‘subject’.

3.1 Terminology extraction
Terminology, like collocations and specialized phraseology, can also be extracted from highly specialized language corpora. Corpógrafo includes a terminology extraction function that is particularly helpful, as it returns a list of candidate terms, which the user can then choose to include in the terminological database, or else simply discard (in case they are not terms). As the terminology search function in Corpógrafo is based on user-selectable search strings (from 1- to 5-grams), it is likely that the candidate terms include words that are not part of the term. While the strings containing only general language should be discarded, those strings containing terms should be included in the Terminology Database (TDB) and edited to remove unnecessary words. Nevertheless, as we consider that (a) not all corpora management software enables terminology extraction, (b) not all those that do operate in the same way, and (c) Corpógrafo is localised in Portuguese, in this research we use simple n-gram searches, followed by human selection of terms, to illustrate how corpora management tools can provide resourceful writing assistance. The following terms, sorted alphabetically, are illustrative examples: “Aboutness is a concept used in LIS”; “work domain analysis in CWA (like DDD) differs from domain analysis”; “Formal Concept Analysis (FCA)”; “Indexing”; “Library of Congress Subject Headings (LCSH)”; “MeSH (Medical Subject Headings)”; “The alternative principle is request-oriented indexing”; “mainstream automatic indexing is not purely document oriented”; “subject access points”.

This list reveals that there are one-word terms that are core to KO (e.g. ‘indexing’), as well as multi-word terms that show different possibilities within a broader term (e.g. ‘request-oriented indexing’ and ‘mainstream automatic indexing’). When writing about information science it is crucial to know that there are ‘subject access points’, as well as different subject headings, including ‘Library of Congress Subject Headings (LCSH)’ and ‘MeSH (Medical Subject Headings)’. Given the requests for precision and concision underlying terminology (Cabré 1999), it is essential to consider that terms are often expressed in shortened forms (including acronyms and abbreviations), and defined in the text. The list of terms above shows several examples of the former: ‘LIS’, ‘FCA’, ‘LCSH’ and ‘MeSH’. Use of these shortened versions, rather than the full versions, will show that the writer has a command of the terminology in the area.

3.2 Retrieval of collocations and identification of specialized phraseology
Retrieval of collocations from corpora has a significant potential for providing writing assistance to inexperienced writers. Information on the collocations of a certain word will steer the writer, by providing them with lexical combination possibilities. By resorting to this method, users can do searches similar to the ones provided by Benson et al.’s (1986) resource, with the significant advantage that they have access to an updated, more easily and faster search capability - and, most importantly, to a resource that applies to their LSP and not to LGP. Word collocations also help writers select the right option from a taxonomy of possible options. Additionally, those collocations allow the writer to learn from specialized writers how to use field-specific phraseology. This is crucial since, similarly to what happens with general language — where a non-native
speaker is identified as such by their use of uncommon collocations and wrong prepositions — the use of uncommon collocations in LSPs may indicate lack of proficiency in the field. The following table illustrates examples of collocations and phraseology for the 5 terms mentioned above.

<table>
<thead>
<tr>
<th>Classif*</th>
<th>Indexes may be classified according to / Objects may be classified based on / systems of classification and categorization / the classification of knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification [systems]</td>
<td>Dewey Decimal Classification (DDC) / Universal Decimal Classification (UDC) / Library of Congress Classification (LCC) / Bibliographic classification / reader-interest classification / Aristotelian classification</td>
</tr>
<tr>
<td>Concept</td>
<td>instances of the concept / representation of a concept / practice of concept systems construction / knowledge-based approaches to concept formation</td>
</tr>
<tr>
<td>Domain</td>
<td>creation of the domain / domain knowledge organization / develop the domain</td>
</tr>
<tr>
<td>Index*</td>
<td>processes of cataloging, subject analysis, indexing and classification / alphabetical index / the development of citation indexes / indexing principles / documents are being indexed / the activity of assigning index terms</td>
</tr>
<tr>
<td>Subject</td>
<td>lists of subject headings / subject knowledge / subject fields / subject retrieval / subject classification schemes / LIS specialists assign subject labels to documents / What is the criterion that a given subject should be attributed to a given document?</td>
</tr>
</tbody>
</table>

These examples illustrate the potential of corpora management software as writing assistants. A writer can learn, after performing a corpus search of the stem ‘classif’ + the wildcard * (in order to obtain all derivatives of ‘classif’), that indexes are classified according to a set of criteria, whereas objects are classified based on other criteria. We also learn that systems of classification and categorization often work closely together, and that knowledge is subject to classification, not grouping or systematization. A corpus search for the same stem also unveils a taxonomy of classification systems: Dewey Decimal Classification (DDC); Universal Decimal Classification (UDC); Library of Congress Classification (LCC); Bibliographic classification; reader-interest classification; Aristotelian classification. Likewise, when searching for the word ‘concept’, one learns that: there are instances, not categories or examples, of a concept; a concept has a representation, not a formulation; concept systems are constructed, not built; concepts can be formed - not created - using knowledge-based approaches. General language speakers and dictionary users will know that a domain refers to an area or territory owned or controlled by a particular power, but will be unaware that domains, in KO, can be created or developed, but not produced.

One crucial term in KO is ‘index’ and its derivatives. The concept is so complex that less experienced writers on the topic can struggle to match the right terms and concepts. Again, a corpus search will be useful to reveal that there are several processes that are approximate: cataloging, subject analysis, indexing and classification. By being aware that these terms are used to name close concepts, writers can look for the fine detail that set each concept apart. It will also show them that indexes can be alphabetical; there are citation indexes, and these are developed; indexing follows a set of principles; documents are indexed; and index terms are assigned. Finally, when writing about ‘subject’,
users might like to know that subjects may have headings and lists can be created from them; there is subject knowledge, but not subject comprehension; subjects are structured into fields and can be retrieved, rather than accessed; subject classification follows schemes, but not plans; there are LIS specialists that assign, but do not allocate, subject labels to documents to make them retrievable, rather than searchable; and that subjects are attributed, not ascribed, to a given document.

4.0 Conclusion

In this paper we described the operation of Corpógrafo as a writing assistance tool for students and researchers, using its original set of tools for terminology extraction, retrieval of collocations and identification of specialized phraseology. The examples selected for this study illustrate a small part of all the available possibilities, but they show the potential of Corpógrafo to allow writers to grasp vocabulary changes and new terms in a specialized domain. The focus of this paper is more on writing assistance as a method, rather than as a tool. Therefore, one of our aims was to demonstrate how general, freely available tools can be used by any writer to provide writing assistance, especially in LSP scenarios where a command of the terminology, phraseology and collocations is essential to write precisely and concisely. The texts included in the corpora are a research basis for extracting very recent vocabulary; hence some fundamental concepts may be less visible. The results can be enriched using a corpus of texts from a longer time interval. It is however important to note that diachronic corpora, which include texts published over a long period of time, are extremely helpful to analyze evolution in time. A systematic study is planned as future development to build solid terminology and phraseology databases in the field of KO, using the methods described.

References


