USING CORPUS LINGUISTICS TOOLS TO HELP TRANSLATION STUDENTS CREATE TECHNICAL GLOSSARIES

Alexandre Trigo Veiga
São Paulo Catholic University – Associação Cultura Inglesa
Brazil

Abstract

The creation of glossaries might become an arduous task if translation students rely simply on document analysis in order to choose the keywords to be included in their lists. This research project was carried out with the goal of developing techniques for identifying terms from a specialized field, by using both Portuguese and English comparable corpora, as well as computer tools designed for linguistic analysis. This method was based on corpus linguistics approaches, and the specific area in this study is Symbolic Freemasonry. The compiled corpora for this study were manuals and rituals used by freemasons during their meetings.

Introduction

When planning lessons for translation students, educators must include ways of improving work productivity due to the highly competitive profile of the translation market. Among the several ways available for improving productivity in translation services, we might emphasize the use of computer-assisted translation tools (CAT Tools) and the creation of integrated electronic glossaries.

CAT Tools, such as Trados or Wordfast, operate with translation memory, i.e., they make suggestions and/or substitutions based on pieces of language that have been previously translated. For instance, after successfully translating a lease agreement from English to Portuguese, a translator saves the original and the translated texts as a translation memory file. Therefore, the next time s/he needs to translate another lease agreement or similar document, the software will identify and make suggestions and/or substitutions for the extracts that are equal or similar to the ones in the translation memory file. Consequently, if a professional translator translates a great deal of lease agreements, s/he might reach a level of automated language translation that will increase considerably her/his work productivity by using CAT tools. If a professional translator wants to use CAT Tools, s/he must have in mind that "achieving full and complete memory is paramount: translation memory is at the heart of automated language translation" (Maylath, 2013, p. 42). Nevertheless, it is important to mention that if s/he translates something in an inaccurate way, unless s/he fixes the translation memory file before starting a new translation, the software will make suggestions and/or substitutions that will repeat the inaccuracies in the translation memory file.

CAT tools have been present in professional translation business for more than 30 years and they offer the feature of including personal glossaries and activating specific ones before starting working on a technical translation. Students must realize that creating and tailoring glossaries for the areas they intend to work with translations is essential for their professional development and might help them build more accurate translation memory files.

A glossary is a collection of terms from a specific science or area of knowledge with their meanings and/or translations. A *term*, in comparison to a general word "is a label - usually lexical - in the special language of a specific domain, designating a particular concept in the knowledge of that domain, and arguably less context-dependent with regard to its sense than a general-language word" (Ahmad, Davies, Fulford, & Rogers, 1992, p. 269). Selecting terms for including in glossaries might be an arduous task depending on how students perform such tasks.

One of the most basic ways in which students go about collecting terms in the process of creating a particular glossary is through *document analysis*, i.e., by reading texts in the specific science or area of knowledge in order to perform such a collection. However, document analysis is time consuming and relies on the reader's skill for identification and selection of terms. Souza (2012), for example, carried out the identification and validation of special nursing language terms in physical motor rehabilitation of adult patients through the usage of the document analysis approach. The author states that 1.425 electronically saved medical records were analysed one by one in the search for terms from the target area (Souza, 2012).

In the research project conducted by Souza, she made use of Excel spreadsheets for collecting terms. In the first column she put the medical record to be analysed and in the second column she listed the words and expressions from the medical record in the first column that were candidates to be terms. It is needless to say that analysing 1.425 medical records is a Herculean task that ought to be recognized as a genuine effort to identify and validate terms from a specific area. Notwithstanding, as it was previously mentioned, document analysis relies on the reader's skills and, therefore, factors like fatigue and human error might influence the final results.

The method proposed in this paper aims at restricting human interaction to the corpus assembling and cleaning and the term validation processes, since the potential term candidates were suggested by the software used in the research through an iteration procedure.

The Theoretical Constructions

The theoretical foundations for this research were based on Corpus Linguistics approaches and terminology studies.

Corpus Linguistics

Corpus Linguistics studies the language in use, in its natural and authentic form and it is present in a *corpus*, which is a collection of texts that was electronically compiled specifically for linguistic analysis through the use of computer tools (Berber Sardinha, 2004). When analysing language in two or more corpora, we can use both parallel and comparable corpora. Parallel corpora consist of two (or more) corpora that contain the original texts and their translations in different languages whereas comparable corpora consist of two (or more) corpora that contain original texts from a similar variety in different languages (Hunston, 2002). Kenning (2010, p. 487) reinforces this idea, stating that "the prototypical parallel corpus consists of a set of texts in language A and their translations in language B" and what puts them together is meaning whereas the texts compiled in a comparable corpus were selected according to one certain criteria such as texts from a specific area of knowledge, which is the case in this study.

The corpora used for this research were compiled from the Internet in free access, i.e., without any special access, and all the texts are available to the general public. They consist of comparable corpora, containing original texts in English and Portuguese. The chosen varieties were manuals and rituals of Symbolic Freemasonry used by Freemasons during their work. Thirteen manuals and rituals were compiled in the Portuguese corpus, and fifteen manuals and rituals were compiled in the English corpus.

Corpus linguists make use of corpora and computer tools for language analysis. The computer tools used in this research project were:

- WordSmith Tools 6.0: Collection of tools for linguistic data analysis. The tool used for this research was the Keywords Tool, which helps identifying words that stand out in a text or set of texts.
- ZExtractor: Software used for the automatic extraction of term candidates in a text or set of texts.
- SketchEngine: A linguistic data analysis system that has several features. The feature used in this research was the Keywords, which helps identifying words that stand out in a text or set of texts.
- Notepad for Windows: Text editor that allows the work with simple text files or ASCII, which is the file type accepted by the computer tools mentioned above and with which better results are obtained.
- Microsoft Excel: Software for creating spreadsheets that allows to filter, calculate, compare and analyse data.

In order to identify keywords and, consequently, potential candidates for terms in a given area of study, we compared the wordlist for the target corpora with a wordlist for the reference corpus. During this comparison, "a word will be key if its frequency is either unusually high or unusually low in comparison to a reference corpus" (Berber Sardinha, 2001, p. 89). The English reference corpora used in this study were the British National Corpus (WordSmith Tools and zExtractor), enTenTen12 (Sketch Engine), and the Portuguese reference corpus was the Corpus Brasileiro. The advantages of using corpus linguistics approaches to language analysis is that we are able to process large amounts of texts and obtain faster results than any document analysis procedure carried out exclusively by humans.

Terminology

Terminology might be defined as a methodology for collecting, describing and presenting terms (Sager, 1990). It is also seen as "a *theory,* i.e. the set of premises, arguments and conclusions required for explaining the relationships between concepts and terms" (Sager, 1990, p.3). In addition, it might also be a vocabulary from a specific area of study (Sager, 1990).

Traditional terminologists tend to work with terms in isolation mainly because they are only concerned with giving names and creating concepts and vocabulary. Modern terminologists prefer to take into consideration the language usage, and the focus is in the use of authentic texts as the main data sources (Ahmad, Davies, Fulford, & Rogers, 1992). The terminological approach in this research project is closer to modern terminologists' points of view since it makes use of authentic texts for term collection and identification. The selected area of knowledge is Symbolic Freemasonry, which is a

division of Freemasonry and consists of three symbolic degrees: Entered Apprentice, Fellow craft and Master Mason. As it was previously mentioned, the texts collected for compiling the corpora were rituals and manuals used by Freemasons during their work that are freely available in the Internet.

Methods

The method to analyse the manual and rituals in the collection of terms consisted of four phases, which are:

1. Phase 1 -- the corpora compilation and cleaning
2. Phase 2 -- an iteration procedure that included the use of the WordSmith Tools 6.0, ZExtractor SketchEngine
3. Phase 3 -- a normalization procedure
4. Phase 4 -- a matching procedure

Phase 1: Corpus Compilation and Cleaning

At this phase, the corpora were compiled and prepared as described below to be analysed by the computer tools selected for this research project. This was a unique process and did not need to be repeated for each tool that was used.

Step 1 – Corpus compilation. Thirteen manuals and rituals in the Portuguese corpus and fifteen manuals and rituals in English were copied from the Internet and individually saved in ASCII format (Notepad format), which is the format accepted by all the computer tools used in this research project. In order to be representative, a corpus must be as large as possible (Berber Sardinha, 2004), and in terminological research, one of the challenges to be overcome is the lack of adequate corpora for the research (Pearson, 1998). The corpora collected for this study have around 200 thousand words each (English and Portuguese) and if we compare them to general language corpora such as the British National Corpus (100 million words), they are quite small.

Nevertheless, it is also important to consider that:

- Symbolic Freemasonry is composed of three degrees (Entered Apprentice, Fellow craft and Master Mason);
- there is one ritual for each of these degrees;
- the terms used remain almost the same in the different rites of Freemasonry (Ancient and Accepted Scottish Rite, Emulation, York and others);
- around 89% of the Lodges in Brazil work the Ancient and Accepted Scottish Rite; and
- in the United States of America and in the United Kingdom there is a balance between the Ancient and Accepted Scottish Rite, York and Emulation.

The representativeness of the corpora of this study is based on the fact that the manuals and rituals in Portuguese and English include the three degrees and the mentioned rites. Adding more similar texts to the corpora would not have a significant impact on the results since there is almost no change in the language used in them.

Step 2 – Corpus cleaning. Cleaning a corpus is the process of getting rid of all the elements that will not interfere in the language analysis. This interference, also known as *noise*, might be, for instance, page numbers, pictures, graphics and everything that is

not linguistically relevant. For this research project, as the texts were copied to the Notepad file, the linguistically irrelevant elements were eliminated.

Phase 2: Iteration Procedure to Identify Terms
At this phase, all the steps described below (Compilation/update of stop words lists or blacklists, extraction of the word list from the corpora, extraction of keywords, qualitative analysis – judgement of the term candidate status, saving the list of terms, iteration procedure with the zExtractor and Iteration procedure with the Sketch Engine) were used in the iteration procedure, which is the repetition of a sequence of steps that provides results closer to what is expected. In this case, it resulted in a list of terms of Symbolic Freemasonry. The five first steps in this phase were used for the first computer tool, WordSmith Tools 6.0 and then repeated for the zExtractor and the Sketch Engine with each corpus (English and Portuguese).

Step 1 – Compilation/update of stop words lists or blacklists. Stop words lists or blacklists are lists of words saved in ASCII format that are unlikely to be considered terms. The computer tools used in this research project have the feature of including lists of words that will not be considered for linguistic analysis. These lists include words such as articles, prepositions, auxiliary verbs and common verbs/nouns and they were loaded in the WordSmith Tools 6.0 and updated for the use with the zExtractor and the Sketch Engine.

Step 2 – Extraction of the word list from the corpora. After loading the reference corpus, the corpus of study and the stop words lists, the first computer tool (WS Tools 6.0) provided a list of words from the corpora.

Step 3 – Extraction of keywords. After the word list operation, the WordSmith Tools 6.0 provided a list of keywords, which are words that are candidates for being terms.

Step 4 – Qualitative analysis – judgement of the term candidate status. After acquiring the keywords list, it was saved in an Excel spreadsheet and each term candidate was judged individually if it was a term of Symbolic Freemasonry or not. The WordSmith Tools 6.0 produced a list of 517 keywords in English and 771 keywords in Portuguese. After the quantitative analysis, 393 keywords in English and 325 keywords in Portuguese were considered terms.

Step 5 – Saving the list of terms. After the judgement of the term candidate status, the keywords from the first computer tool (WS Tools 6.0) that were considered terms were saved into an Excel spreadsheet, and the keywords that were not considered terms were included in the stop words lists or blacklists to be used in the zExtractor and the Sketch Engine.

Step 6 – Iteration procedure with the zExtractor. As it was mentioned before, the five first steps in this phase were repeated for the zExtractor. Out of 428 keywords in English, 176 terms were identified and out of the 441 keywords in Portuguese, 271 terms were identified. After adding the terms to the list and excluding the repeated ones, the use of the second computer tool helped identify 41 new terms in English and 140 new terms in Portuguese. Again, the keywords that were not considered terms were included in the stop words lists or blacklists to be used in the Sketch Engine. At the end

of this stage the lists of terms in English had a total of 434 words, and the list of terms in Portuguese had a total of 465 words.

Step 7 – Iteration procedure with the Sketch Engine. Similar to step 6, the five first steps in this phase were repeated for this third computer tool. Out of 343 keywords in English, 258 terms were identified and out of the 179 keywords in Portuguese, 169 terms were identified. After adding the terms to the list and excluding the repeated ones, the use of the third computer tool helped identify 32 new terms in English and 6 new terms in Portuguese. At the end of this stage the lists of terms in English had a total of 466 words and the list of terms in Portuguese had a total of 471 words.

Phase 3: Normalization
The normalization process consisted of grouping together terms that presented variations in spelling, abbreviations, gender, number and conjugation and placing them in lines for building the glossary. Words with variations in spelling, gender, abbreviations, number or conjugation were placed together. For instance, in the English terms list, the words *bro*, *brother* and *brethren* were put in one line, whereas in the Portuguese terms list, the words *venerável*, *ven* and *veneráveis* were put together. After the normalization process, 466 terms in the English list were grouped in 350 lines and the 471 terms in Portuguese were grouped in 368 lines.

Phase 4 – Matching. Having the two lists in different languages, the final phase consisted of matching the terms in one language with their respective translation in an Excel spreadsheet. The terms that had no match were searched in dictionaries for possible translations. The translated words were searched in the corpus of that specific language and then included in the glossary.

Conclusions

A glossary with 392 words was achieved using the method proposed in this study without having to read all the texts in the corpora. The glossary created can be easily inserted in CAT tools that offer the feature of glossary inclusion. Even considering that around one thousand keywords had to be judged if they were terms or not, the amount is much smaller than the document analysis carried out by Souza (2012) that initially extracted 827.047 terms after reading 1.425 medical records. Therefore, in comparison to document analysis, the method in this study proved to be less time consuming. In addition to that, the iteration process contributed for identifying more terms in comparison to the use of only one computer tool, a fact that adds relevance to the proposed method. The teaching and application of this method shall facilitate the lives of translation students and professionals as it may be used for building glossaries in other areas of knowledge or interest. However, in order to obtain satisfactory results through the use of this method, translation students and professionals must be aware of the following aspects:

- **Corpus relevance:** The corpus must be relevant and not necessarily very large. When compiling the corpus, students must have in mind that sometimes bigger does not mean better, and they should focus on selecting genres that contain a high incidence of terms, such as rituals and manuals in this study or medical records in Souza's case (2012).

- **Cleaning the corpus:** This procedure excluded the possibility of contamination because non-linguistically relevant elements such as page numbers or pictures were eliminated.
- **Stop words lists or blacklists:** The update of stop words lists or blacklists helped reduce the number of the keywords lists and facilitated the qualitative analysis.
- **Reference corpus selection:** The reference corpus should be at least twice the size of the corpus of study and should be representative of the general language in use.
- **Iteration procedure:** The use of three computer tools in a sequential mode, updating stop words lists or blacklists, helped identify more terms in a precise manner.
- **Qualitative analysis:** The judgement of the term candidate status is something that needs to be conducted by someone who knows the terms in both languages, once computer tools make suggestions by crossing frequency and they are not able to validate terms.

References

Ahmad, K., Davies, A., Fulford, H., & Rogers, M. (1992). What is a term? The semi-automatic extraction of terms from text. In M. Snell-Hornby, F. Pöchhacker, & K. Kaindl (Eds.), *Translation Studies: An Interdiscipline* (p. 267-277). Vienna, Austria: John Benjamins Publishing Company.

Berber Sardinha, T. (2001). Comparing Corpora with Wordsmith Keywords. *The ESPecialist. Pesquisa em Língua para Fins Específicos. Descrição, Ensino e Aprendizagem*, *22*(1), 87-99.

Berber Sardinha, T. (2004). *Lingüística de Corpus*. São Paulo, Brazil: Manole.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge, United Kingdom: Cambridge University Press.

Kenning, M. M. (2010). What are parallel and comparable corpora and how can we use them? In A. O'Keeffe & M. McCarthy, M. (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 487-500). New York, NY: Routledge.

Maylath, B. (2013). Current trends in translation. *Communication & Language at work 2*(2), 41–50.

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

Sager, J. C. (1990). A practical course in terminology processing. Amsterdam, Netherlands: John Benjamins Publishing Company.

Souza, D. R. P. de. (2012). *Identificação e Validação de Termos de Linguagem Especial de Enfermagem em Reabilitação Física Motora de Pacientes Adultos* (Master's thesis). Retrieved from the digital library from the Federal University of Minas Gerais, Belo Horizonte.

Author Details

Alexandre Trigo Veiga
aletrigo@globo.com