http://dx.doi.org/10.5007/2175-7968.2016v36nesp1p62

REVISITING CORPUS CREATION AND ANALYSIS TOOLS FOR TRANSLATION TASKS

Claudio Fantinuoli*
Johannes Gutenberg University Mainz

Abstract: Many translation scholars have proposed the use of corpora to allow professional translators to produce high quality texts which read like originals. Yet, the diffusion of this methodology has been modest, one reason being the fact that software for corpora analyses have been developed with the linguist in mind, which means that they are generally complex and cumbersome, offering many advanced features, but lacking the level of usability and the specific features that meet translators' needs. To overcome this shortcoming, we have developed TranslatorBank, a free corpus creation and analysis tool designed for translation tasks. TranslatorBank supports the creation of specialized monolingual corpora from the web; it includes a concordancer with a query system similar to a search engine; it uses basic statistical measures to indicate the reliability of results; it accesses the original documents directly for more contextual information; it includes a statistical and linguistic terminology extraction utility to extract the relevant terminology of the domain and the typical collocations of a given term. Designed to be easy and intuitive to use, the tool may help translation students as well as professionals to increase their translation quality by adhering to the specific linguistic variety of the target text corpus.
Keywords: Corpus tools. Translation. Professionalization. Monolingual corpus.

* Claudio Fantinuoli: PhD in Applied Linguistics from the Johannes Gutenberg Universität Mainz (2012). Study of Translation and Interpreting at the Scuola Superiore per Interpreti e Traduttori of the University of Bologna/Forlì. Lecturer and researcher in the Faculty of Translation and Interpreting of the Johannes Gutenberg Universität Mainz in Germersheim, Germany. E-mail: fantinuoli@uni-mainz.de

Esta obra utiliza uma licença Creative Commons CC BY:
https://creativecommons.org/lice

A VUELTAS CON LA COMPILACIÓN Y HERRAMIENTAS DE ANÁLISIS DE CORPUS PARA LA PRÁCTICA DE LA TRADUCCIÓN

Resumen: Muchos investigadores han propuesto el uso de corpus como herramienta para que los traductores profesionales produzcan textos de alta calidad que puedan leerse como si fueran originales. Sin embargo, la difusión de esta metodología ha sido reducida. Una de las razones tiene que ver con el hecho de que los programas de análisis de corpus se han desarrollado teniendo en mente la figura del lingüista, lo que, en líneas generales, los ha llevado a ser complejos y engorrosos: si bien ofrecen muchas características avanzadas, carecen del nivel de usabilidad y características específicas que satisfagan las necesidades de los traductores. Ante este panorama, hemos desarrollado TranslatorBank, una herramienta gratuita de creación y análisis de corpus diseñada para la práctica de la traducción. TranslatorBank permite crear corpus monolingües especializados a partir de la web; extraer concordancias con un sistema de consulta similar al de un motor de búsqueda; utiliza medidas estadísticas básicas para indicar la fiabilidad de los resultados; accede directamente a los documentos originales para obtener más información contextual; incluye un extractor terminológico basado en datos estadísticos y lingüísticos para extraer la terminología relevante del ámbito, así como las colocaciones típicas de un término dado. Diseñada para ser intuitiva y fácil de usar, esta herramienta puede ayudar a los estudiantes de traducción, así como a los profesionales a aumentar su calidad de traducción ateniéndose a la variedad lingüística específica del corpus en lengua meta.
Palabras clave: Herramientas de corpus. Traducción. Profesionalización. Corpus monolingüe.

Introduction

During the last twenty years or so, corpora and corpus analysis software have been proposed in literature as an effective tool and methodology for providing a data-rich learning environment in translation training (cf. Aston, 2009; Bowker, 1998; Fantinuoli, 2013; Kübler 2011), and for improving translation quality in the profession (cf. Zanettin, 2012). Yet, with the exception of translation memories – a very specialized kind of parallel corpus

(Zanettin 2002b: 247) – and searchable online corpora, such as Linguee[1], it is difficult to deny that the use of corpora has not become widely established among professional translators (cf. Aston 2009; Bowker 2004). This has been confirmed by several surveys conducted during the past years (cf. Picton et al., 2015; Gallego-Hernández, 2015; Jaaskelainen and Mauranen, 2005; MeLLANGE, 2006; Scott, 2012). Despite several drawbacks, most translators still seem to prefer easy-to-use, out-of-the-box solutions, such as dictionaries, online databases and search engines, over corpora. There are several reasons for this. Firstly, only few translators were trained in using corpus analysis tools as translation aids. Secondly, even if such tools were mentioned during training, for example in specialized translation classes, they are not used in professional settings since "the design, compilation and exploitation of corpora can be very time-consuming while not providing a tangible immediate increase in productivity" (Bernardini, 2006, 19), especially when working under tight time-constraints. Aston describes the problem with the following words (2009, X):

> Regardless of its potential to improve translation quality and to provide a fruitful learning environment, corpus consultation remains time-consuming, and corpus construction enormously more so. One part of the problem is whether and how we can improve the efficiency of corpus use for the translator, facilitating both consultation and construction, and do so without compromising its quality as a translating and learning tool.

Since the problem seems to be of a cost-benefit nature, as corpus creation and analysis requires time and some computational skills many translators do not have or are not willing to acquire, it is our hypothesis that one of the reasons why corpora fail to establish among

[1] http://www.linguee.com

translators has to do with the tools that have been made available in the past. Although several corpus querying tools now exist, none has been specifically designed to meet translator's needs. Some programs, for example MicroConcord[2] and MonoConc Pro[3], were designed with a pedagogical application in mind while others, including WordSmith Tools[4], AntConc[5] and TextSTAT[6], to name but a few, were aimed at linguistic researchers, computational linguists, lexicographers, and so forth. Having to satisfy the needs of these target groups, some of the tools include a forbidding range of complex options which can easily confuse the user (Gavioli and Aston, 2001); they do not lend themselves to easy use in contexts which are different to those for which they have been developed (Anthony, 2013); and, even when they are easy to use, they do not implement the functions which could be regarded as relevant in a translation setting, for example the possibility to create on-the-fly corpora or to easily mine information to help solving translation problems.

Whereas the use of concordancers in translation practice and training has received closed attention during the last years, software applications for corpus analysis have never been thoroughly investigated in terms of their suitability for translation tasks and not much is known of how a corpus tool for translators should look. Yet, there is a general consensus about the fact that corpus construction and use has to be made substantially easier and faster. Bernardini (2006, 21), for example, suggests that, for corpora to be successful with translation professionals, "corpus construction and corpus searching tools should be made more user-friendly". Similarly, Zanettin (2002a, 10) suggests that for corpora and concordancing software to find a larger place in the translator workstation "corpus builders and software producers should take into account the specific needs of this group of users". In this paper

[2] http://lexically.net/software/index.htm
[3] http://www.athel.com/mono.html
[4] http://www.lexically.net/wordsmith/
[5] http://www.antlab.sci.waseda.ac.jp/software.html
[6] http://neon.niederlandistik.fu-berlin.de/textstat/

I would like to argue that the demands of a corpus program for translation tasks are quite different from the needs in the academic community and, therefore, need to be specifically addressed in the software design. A translators' corpus tool should be easy-to-use, facilitating both the creation and consultation of corpus data, and resemble the tools to which translators are accustomed for their documentation activities. As corpus software has to be adapted to this new context in order to be successfully integrated in the translation workflow, I shall finally propose a free tool specifically designed for this target user, called TranslatorBank[7].

1. Corpora and translation

In the field of translation, the use of corpora has had a growing impact during the last decades as it allows translators and researchers to move from the observation of small text samples to the investigation of larger collections of texts. Corpora have been used for descriptive and practical purposes: on the one hand, scholars have analyzed corpora of translations and interpretations, comparing them to corpora of original texts in order to establish the characteristics peculiar to translations (cf. Baker, 1996; Gellerstam, 1996; Hansen, 2003; Hansen-Schirra et al., 2012; Mauranen and Kujamäki, 2004) and interpretations (cf. Bendazzoli and Sandrelli, 2009; Pöchhacker, 2009; Shlesinger, 1998). On the other hand, they have been used in translator education and training (cf. Bowker, 1998; Zanettin et al., 2003) and have been proposed as aids in a professional environment, both in translation (cf. Bernardini and Castagnoli, 2008; Zanettin, 2002a) and in interpreting (cf. Fantinuoli, 2012; Gorjanc, 2006) settings.

In translation practice and education, both general and specialized corpora have been suggested to be effective tools in enhancing the quality of translations (cf. Gavioli and Zanettin, 1997; Varantola,

[7] http://www.staff.uni-mainz.de/fantinuo/translatorbank.html

2003). To put it simply, the advantage of using corpora is that corpora consist of a more comprehensive and diverse variety of source language items and possible translation solutions than a dictionary (cf. Hansen-Schirra and Teich, 2002; Zanettin et al., 2003); they allow autonomous learning (cf. Fantinuoli, 2013); they foster awareness-raising about language and translation and they obviously have a crucial documentation role (cf. Bernardini and Castagnoli, 2008). By browsing in a target language corpus, translators are able, for example, to reduce the amount of unwanted "shining through" (Teich, 2003), investigate terminology, explore phraseology and acquire working knowledge. Providing the translator with a set of "units of meaning" (Tognini-Bonelli, 2001), corpus analysis can support translators in emulating 'good' writing in the relevant text type and variety. This is particularly important in light of the fact that translators are often laypersons in the subjects they are translating and, even if they are specialized in one or more subjects, they do not generally share the same amount of domain-specific knowledge as the experts of that field. This is the reason why they need to acquire both factual information about a specific topic as well as its linguistic realizations in order to reduce the gap between them and the authorship/readership, and to 'clone' the customers' language, i.e. reproduce their wordings, phrases, styles, etc. (cf. Fantinuoli, 2013). Corpus analysis has been proposed as a fruitful methodology to cope with all these issues in a translation setting.

2. Translators' needs and corpus tools

Documentation activities are of vital importance to translators. Especially within the context of specialized translation, the access to the right information at the right time is considered to be a key asset for the delivery of high-quality translations. Traditionally, translators rely on three types of documentation resources:

dictionaries, including lexicographical and terminological databases, reference texts, and web queries using commercial search engines. These resources are generally used opportunistically in order to acquire or enhance – while translating – linguistic, cultural and domain competence: they are used to look up terms and expressions, confirming or rejecting translation hypothesis, or tentatively find solutions to lexical problems, for example by cross-reading domain-related reference texts collected on the web. In particular, the fact that reference texts, mainly in the target language, are generally retrieved from the web, has transformed the Internet and search engines into the most widespread and informal documentation environment for translators.

The way search engines and reference texts are used for documentation and problem-solving activities closely resembles the uses of corpora discussed in literature, for example for explorative learning and translation enhancement. Surveys conducted among professional translators point out that monolingual corpora, as is the case with search engines, are mainly used to find equivalents at a terminological and phraseological level and to confirm or reject translation hypothesis (Picton et al., 2015). In particular, Gallego-Hernández (2015, 7) found out that "term extraction stands out as the main purpose for which translators make use of corpora (86%), followed by collocations and phraseology (64%)". Other possible uses of corpora, such as to explore and understand the source text or the subject, are indeed less widespread.

As introduced in Section 1, traditional corpus tools offer several advanced functions, such as statistical measures, clustering, plotting and so forth. Yet, these functions are researcher-centric and not among the tools translators generally need when dealing with corpora. Other functions, such as keyword extraction, for example, which is considered a useful task by translators, generally require language-dependent resources, the acquisition or creation of which is not straightforward. Other functions again, such as extraction of complex terms or corpus building, are not available at all. Corpus creation functionalities, in particular, are of extreme

importance for our target group. Since corpora are especially used in the context of specialized translation (Beeby et al., 2009) and since there is no ready-to-use corpus for any specialized subject and language, translators need to create their own corpus prior to a translation assignment or project. The compilation of these types of corpora, which are referred to as ad-hoc (Aston, 1999), disposable (Varantola, 2003) or DIY (Zanettin, 2002a) corpora, is not straightforward. The process of collecting and creating a corpus is generally time-consuming if done manually, for example by downloading texts from the web, or requires the use of extra software for automatic compilation. Indeed, the time needed to compile a corpus is considered as one of the main arguments against the use of such corpora, especially in the light of the fact that DIY corpora are often used only in the context of a single translation (Castagnoli, 2006).

Various tools have been developed in the past years to speed up the process of corpus creation, for example the well-known BootCat (Baroni and Bernardini, 2004), a suite of Perl script for bootstrapping specialized language corpora from the web[8]. Only two concordancers integrate some sort of corpus creation functionality in their software: WordSmith offers WebGetter, which is basically a simplified version of BootCat, while TextStat offers Web2corpus, a simple web spider. Yet, at the time of writing, the only tool still maintained and working is BootCat. The main drawback of this tool is that it collects texts only from Websites (HTML) and does not support PDF, which can be considered the most important format in which reliable specialized texts are disseminated[9]. Furthermore, it is available as a standalone program, which means it requires users to install and learn a new program as well as integrate it in the translation workflow.

Since traditional corpus tools have been loaded with too many options, making the graphical user interface rather cumbersome

[8] http://bootcat.sslmit.unibo.it/

[9] See for example Odlyzko, 2000.

and complex too use, the learning curve of available concordancers may be regarded by many translators as too steep to embark in the use of corpora. If we consider that the steepness of this learning curve is not only related to software use, but has also to do with grasping the basics of corpus exploration (Frankenberg-Garcia, 2012) and the high level of analytical skill and attention to detail needed (Braun, 2007), a first contribution towards facilitating corpus consultation, and consequently the acceptance of this methodology, could be to provide translators with a tool which is easy to use and makes the info mining tasks more straightforward.

It is our assumption that a corpus program specifically designed for translators should comprise at least the following features:

- Quick creation of disposable, domain-related corpora
- Easy-to-use, versatile query language which resembles the way search engines work
- Basic statistics to interpret the reliability of the results
- Seamless integration of part-of-speech (POS) analysis
- Hypertextuality (direct access to original documents) for documentation
- Terminology extraction to obtain the most important terms of the subject
- Computing of co-occurrences (collocates) of selected terms

The program should be easy-to-use, the graphical user interface clear and well-structured and all features should work without requiring any computational skill, both for its installation and usage[10]. The software should work with as many languages as possible and should easily accommodate other languages as soon as the needed resources are made available.

The next sections explain in detail the implemented architecture of the tool.

[10] See for example Zanettin (2002a) who reports how several studies among professional translators and trainees have underlined that the user-friendliness of the concordancing software was very low.

3. TranslatorBank architecture

Basically, the architecture of TranslatorBank is comprised of three main parts:

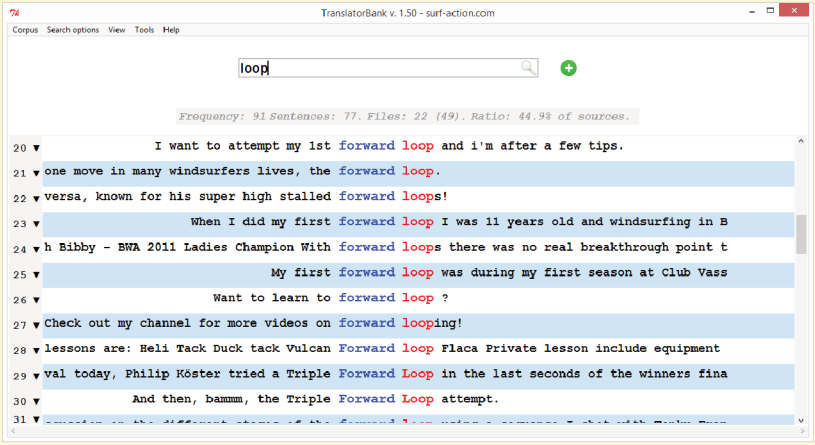
- A concordance with a "search engine-like" query system
- A tool to automatically create specialized monolingual corpora from the web
- A tool to automatically extract specialized terminology and collocates from the corpus

The three parts are designed to be seamlessly linked to each other and to be used off-the-shelf.

3.1 Concordancer

In corpus linguistics, the concordancer is considered to be the central tool to access and analyze corpus data. Listing all occurrences of the query item together with some surrounding context in the form of words to the left and right, it allows to show how words or phrases are used in the immediate contexts in which they appear. As introduced above, the main idea in developing a translator-oriented concordancer is to design a user-friendly and intuitive interface, giving priority to clarity and simplicity over a large number of options. As shown in Fig. 1, the entire area of the concordancer is occupied by the input field and the query results. Advanced options are displayed at the top of the window only if the user wishes it. At the top of the results area, basic statistical information about the query are shown.

Fig. 1: Simple GUI: query results, ordered first word left[11]



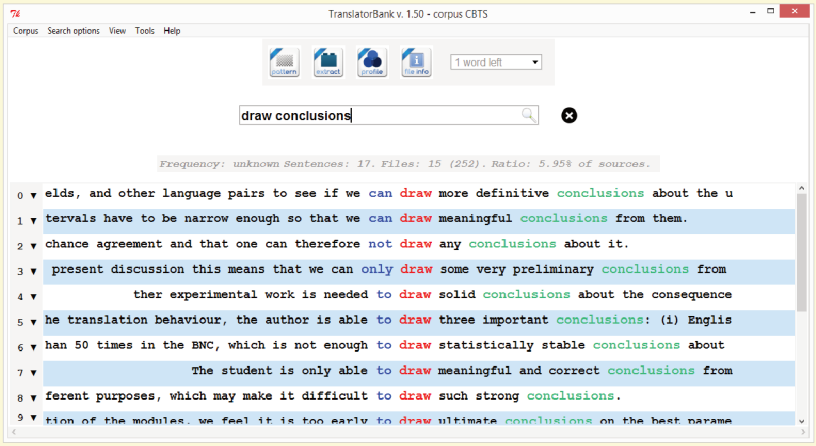
To enhance clarity, the query string is highlighted in red. The number of characters shown on both sides, the so-called "span", is limited in size to fit in the window and can be adjusted by the user. When more context for a specific result is needed, the span can be increased to a given set of sentences preceding and following the query item. Furthermore, the user can directly access the original text in its entire length and layout (both PDF and webpages), for example to see pictures, tables etc., making the concordancers a proper, hyperlinked documentation tool.

In order to increase the usability of the software, the query system is designed to replicate the behavior of a search engine as far as possible, as this is considered a very familiar environment for translators. By default, queries are performed in a case-insensitive way. If the input string is a single word, all sentences containing that word are shown among the results. If the input string is made up of two or more words; subsequently, the so called *proximity search* is performed: all sentences containing the words within a certain window

[11] Source: Author's personal archives (permission to use granted by author)

span are displayed and the user is prompted to choose whether or not the word order is relevant. Contrary to traditional corpus analysis tools, this functionality is already activated by the program by default and does not need to be defined by means of advanced masks. The *proximity search* is particularly useful in offering a flexible and easy way to explore the corpus with a serendipity approach (Johns, 1988): when translating an academic paper into English, for example, to confirm or reject the translator's hypothesis that the verb "draw" collocates with the noun "conclusions", the user can search for both words. The proximity search produces the following results, which easily confirm that the initial hypothesis is correct.

Fig. 2: Results with proximity search, ordered first word left[12]



To change the standard search behavior of the software, the operators used by the major search engines are implemented:

- Quotes ("): when a word or a phrase is included in quotes, the search will include sentences with the same words and in the same order as in the quotes. This is useful if

[12] Source: Author's personal archives (permission to use granted by author)

the user is looking for the exact word or phrase. Example: "corpus linguistics".

- Dash (-): when a dash is used in front of a word, the search excludes sentences containing that word. This feature can be used to filter out unwanted sentences. Example: "corpus –based studies".
- Asterisk (*): an asterisk can be used in a query performed within quotes. It works as a placeholder for an unknown word. Example: "a * car".

Since translators are supposed to use the concordancer opportunistically, i.e. with very specific and practical questions in mind, and since they will do this under strict time constraints and without the possibility of performing extensive analyses, the tool offers a mechanism that suggests the level of reliability of the results. When using a search engine to confirm translation hypothesis, it is common, for example, to do it on the basis of sheer frequency values. If a term or a phrase has a certain minimum frequency, the hypothesis is confirmed. Similar behavior can be observed in translators using a corpus analysis tool. Yet, absolute frequencies do not take into consideration distributional tendencies inside the corpus, for example, the idiosyncratic use of a particular term by one author or company. If a lexical item is found only in one document of the corpus, but with high frequency, the user, seeing a high number of results, could be tempted to consider it as a typical item of the domain. To avoid this risk, the query results are accompanied by some basic statistics, comprising:

- Absolute frequency of the query item
- Number of files in which the query item recurs
- Ratio between number of files in which the item recurs/total number of files in the corpus
- Distribution of the item among the sources

The ratio, for example, could be a useful indicator to discriminate between two possible lexical alternatives (quasi-synonyms): the higher the ratio, the more common the unit should be in the domain of interest.

With all these functions implemented, the concordancer becomes a sort of indexer and archive for reference texts, looking and behaving in the way of common search engines. Yet, contrary to search engines or archiving tools, it keeps the peculiar query and visualization features which are so important for corpus analysis. Furthermore, it integrates a series of advanced features potentially useful for translators, as the next sections will show.

3.2 Monolingual corpus creation

The corpus creation utility is designed to build on-the-fly specialized corpora using the web as a text repository (Baroni et al., 2006). To reduce the amount of time needed to collect and create the corpus, the software design focuses on ease of use and process automation.

The following features are available:

- Automatic search for URL of domain related documents (PDF and webpages)
- Download, conversion and cleaning of PDF and HTML in plain text and XML files
- Annotation of each text with basic information, such as URL, document name, etc.
- Creation of a SQLite database for quick search even with large corpora
- Part-of-speech tagging for morphological insensitive queries and language depended features such as terminology extraction

The implemented corpus creation procedure, which is almost unsupervised, shows some similarities with that proposed by the

work of Baroni and Bernardini (2004). It uses an API of the search engine Bing[13] to collect URLs of pages dealing with the subject of interest, downloads and prepares them for corpus query. The workflow is straightforward: the corpus building procedure starts with a small set of terms that are expected to be representative of the domain. These terms are used as a Bing query string and the top pages (PDF or HTML) returned for each query are downloaded and formatted as text. The result of this procedure is a monolingual collection of XML-annotated texts. During the last years, several experiments have successfully used this procedure to create corpora from the web for translation or interpreting tasks, see for example Fantinuoli (2006) and Bernardini & Castagnoli (2008). To prevent the software from collecting unrelated texts, the initial terms must be unambiguous, highly specialized and possibly used only within the domain of interest. The number of terms which are needed to create a corpus depends on the quantity of queries the user wants to start. To build a medium size corpus of approx. 100,000 tokens, between 4 and 6 specialized terms are needed.

The user can influence the corpus building procedure by setting a small range of parameters such as: *Count*, i.e. the number of URLs to be collected for each query (influencing the size of the corpus); *Language*, the language of the documents to be retrieved; *Format*, the format of the documents (PDF/ HTML); and *Domain,* a value which allows to restrict the search to a specific domain or Internet address (for example to create a company related corpus).

Once the URLs have been collected, users can assess the quality and the relatedness of the collected URLs and discard what is considered un suitable for the corpus. The remaining list of URLs is downloaded and each PDF file or HTML is converted into plain text, cleaned and saved in unique files. Whenever available, meta-information, like original URL, source, date and so forth, are saved in the XML structure.

[13] http://www.bing.com

In order to reduce the data noise, texts retrieved from both PDF and HTML documents need to undergo various cleaning steps. This operation is generally considered trivial, but it is very important for the usability of the corpus (see Maher et al., 2008). In order to clean up the retrieved corpus, we apply a series of heuristics to the texts extracted from both HTML and PDF files, for example, we keep in the final corpus only HTML texts between 5 KB and 150 KB in size (Fletcher, 2004), we strip off the HTML tags and remove codes (for example Java-scripts) and remove boilerplates, as they will invalidate statistics collected from the corpus, impair attempts to analyze the text by looking at KWIC concordances (Ferraresi et al., 2010) and produced biased terminology lists.

The texts are saved in plain text format for use in other tools, and in XML with a simple annotation schema containing URL, title, time stamp and other information automatically retrieved from the converted text, for the database creation. Additional annotations can also be set manually by the user when starting the retrieving procedure. The collected texts are imported in a SQLite database, which is automatically loaded in the concordancer to be looked up.

3.3 Terminology extraction

The purpose of this function is to extract a list of monolingual specialized terms and phrases from the collected corpus that can be used by translators to 'clone' a particular language, to see how particular companies, perhaps competitors, write, etc. Monolingual terminology extraction systems are traditionally based on two basic approaches: on one hand, the linguistically-based or rule-based approach (Ananiadou, 1994; Dagan and Church, 1994), and on the other the statistical corpus-based approach (Khurshid et al., 2000). The implemented method, which is hybrid as it combines linguistics knowledge and statistical measures, is similar to the monolingual term extraction approach described in an interpreter-oriented experiment by Fantinuoli (2006). To improve the usability

of the software, the focus is on obtaining a high level of precision rather than recall[14].

The terminology extraction process consists of two separate steps: extraction of single-word terms and of multi-word terms. The tool is designed to work for a vast number of languages, provided the following resources are available for the respective language:

- morphosyntactic rules for multi-word terms
- parameter file for the TreeTagger[15]
- word frequency list from a general reference corpus

The first step of the terminology extraction procedure is to compare the relative frequencies of terminology units in a specialized and a general reference corpus, which provides a text norm or standard, to find single-word terms typical of the former. The basic idea is that a general corpus covers many domains and can therefore be used as a reference, whereas a specialized corpus emphasizes a given domain. Assuming that specialized terms are more frequent in a specialized than in a balanced corpus, the terminology extraction tool considers the items that show a high relative frequency in the specialized corpus as term candidates. This is a fairly common approach in terminology extraction and corpus comparison areas (Ahmad and Rogers, 1992; Baroni and Bernardini, 2004). TranslatorBank integrates the free available corpora of the Europarl project (Koehn 2005), a collection of bilingual corpora extracted from the proceedings of the European Parliament. The varied nature of the EU proceedings, where no specific topic or domain has dominance and the number of languages available makes these texts suitable for corpus comparison.

The multi-word terms extraction combines a statistical and linguistic approach. In this approach, a morphological category

[14] For more details on the values of precision and recall obtained with this method with German, English and Italian refer to Fantinuoli (2006).
[15] http://www.cis.uni-muenchen.de/ ~ schmid/tools/TreeTagger/

is associated to each word. The extraction is based on the assumption that single-word and multi-word term candidates have a certain fixed set of linguistic properties, for example "Noun + Preposition + Noun" are likely to be candidate terms in Italian ("catena del freddo", "barca da riporto", etc.). After assigning a part-of-speech tag to each word, it is possible, with a set of regular expressions, to extract all candidate terms that adhere to this and other patterns.

We proceed in two steps: we first apply a linguistic filter based on a part-of-speech analysis, selecting candidate multi-word terms from the corpus; we then apply statistical measures to rank the candidate terms and select the most appropriate. The tool applies a predefined, language dependent set of POS-patterns to generate candidate multi-word terms. Users can change or extend the set of rules proposed. The algorithm extracts all word combinations matching the defined set of POS-patterns. The candidate multi-word terms are ranked according to their frequency. At this stage, we choose to retain only the frequency as a means, to rank the multi-word terms extracted with the linguistic approach. Daille (1996, 33) points out that:

> Frequency is the most significant score to detect terms of a technical domain. This results [sic.] contradicts numerous results of lexical resources, which claim that association criteria are more significant than frequency.

A series of informal tests conducted with the extraction algorithm confirms this observation. However, in the same paper, Daille points out that "The remaining problem with the sort proposed by frequency is that it very quickly integrates bad candidates". The results of our algorithm contradict this, at least for corpora with a high specialization level. When extracting n-grams from very repetitive corpora of reasonable dimensions (of at least 50,000 tokens), the key terms occur very frequently. Ranking the POS

filtered multi-word terms by their frequency seems to produce reliable results in terms of selection of common specialized terms, leaving out most of the poorly formed candidates as they will statistically occur less frequently.

4. Collocates extraction

TranslatorBank adheres to the "directional" view on collocations (Evert, 2005), which starts from a given keyword (the node) and aims at identifying its collocates. The goal of such an approach, which is widely used in computational lexicography (Sinclair, 1991), is to identify those collocates which are the most characteristic for the given node, i.e. collocates which are very frequent in the specific domain, leaving out rather atypical collocational patterns. In fact, for our target users, we assume that they predominately need the most typical and therefore more frequent linguistic information (terms or collocations) in order to 'clone' a certain language for special purposes. This is in contrast with the needs of terminographers and lexicographers, who generally want to (linguistically) cover the entire domain under investigation.

The node is initially defined by the user and corresponds to the query string. The collocates are identified statistically by counting the number of occurrences of all tokens conforming to the POS pattern of interest, which occur in a defined window span. The most frequent collocates are presented in the GUI as a list of collocates and their frequency or as a word cloud. For example, from a specialized corpus about "electrical discharge machining" the tool computes for the node "surface" the following collocates: finest, eroded, spark-machined and rougher.

The implemented solution is certainly a very simple form of collocates extractions as it lacks any notion of syntactical knowledge and does not use any advanced statistical measures, such as Mutual Information or T-test. Still, as it replicates what corpus users generally do by ordering the KWIC to the left or the right in search

of recurrent patterns, the results should satisfy the demands of most translators and interpreters.

5. Conclusion

In this paper, we argued that a reason why corpus analysis failed to become widely established among professional translators is because of the lack of software specifically designed to meet translators' needs. In order to narrow the gap between the evident advantages of using corpora in a translation setting, as indicated in literature, and the scarce diffusion of corpus analysis among translators, we propose a user-friendly tool which, resembling the way commercial search engines work, facilitates both corpus consultation and construction, and do so without compromising the typical quality of corpus analysis methods.

Adjusting the design of corpus software to the practical requirements of translators is a substantial step towards the successful integration of corpus analysis in the everyday life of professional translators. This paper has focused on the role of corpus tools as a translation aid, and in particular on the features that are mostly needed by translators in their professional workflow, an aspect which has been somewhat underestimated in the past. Proposing a tailor-made corpus technology, TranslatorBank may contribute to establish the use of corpora among professional translator and foster the scientific debate about the needs of this particular target group.

References

Ahmad, K. and Rogers, M. "Terminology management: a corpus-based approach." *Proceedings of Translating and the Computer 14: Quality Standards and the Implementation of Technology in Translation*. London, 1992. 33–44.

Ananiadou, S. "A methodology for automatic term recognition." *Proceedings of the 15th conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, 1994. 1034–1038.

Anthony, L. "A critical look at software tools in corpus linguistics." *Linguistic Research* 30.2 (2013): 141–161.

Aston, G. "Corpus use and learning to translate." *Textus* 12 (1999): 289–314. Available at: http://www.sslmit.unibo.it/guy/textus.htm

Aston, G. "Foreword." *Corpus use and translating: Corpus use for learning to translate and learning corpus use to translate*. Ed. A. Beeby, P. Rodríguez-Inés, & P. Sánchez-Gijón. Amsterdam: John Benjamins, 2009. IX–X.

Baker, M. "Corpus-based Translation Studies: The Challenges that Lie Ahead." In: H. Somers (ed.). *LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Amsterdam: John Benjamins, 1996. 175–186.

Baroni, M. and Bernardini, S. "BootCaT: Bootstrapping corpora and terms from the web." *Proceedings of LREC*. 2004.

Baroni, M., Bernardini, S. & Evert, S. "A WaCky Introduction." In: M. Baroni & S. Bernardini (eds.). *Wacky! Working papers on the Web as Corpus*. Bologna: GEDIT, 2006. 9–40.

Beeby, A., Rodríguez-Inés, P. and Sánchez-Gijón, P. *Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate*. Amsterdam: John Benjamins, 2009.

Bendazzoli, C. and Sandrelli, A. "Corpus-based interpreting studies: Early work and future prospects." *Tradumatica* 7 (2009). Available at: http://www.fti.uab.cat/tradumatica/revista/num7/articles/08/08.pdf

Bernardini, S. "Corpora for translator education and translation practice: Achievements and challenges." *Third International Workshop on Language Resources for Translation Work, Research & Training*, 2006. 17–22.

Bernardini, S. and Castagnoli, S., "Corpora for translator education and translation practice." *Topics in language resources for translation and localization*. Ed. E. Y. Rodrigo. Amsterdam: John Benjamins, 2008. 39–55.

Bowker, L. "Corpus resources for translators: academic luxury or professional necessity." *Tradterm* 10 (2004): 213–247.

Bowker, L. "Using specialized monolingual native-language corpora as a translation resource: a pilot study." *Meta: Translators' Journal*, 43.4 (1998): 631–651.

Braun, S. "Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora." *ReCALL* 19.03 (2007): 307–328.

Castagnoli, S. "Using the Web as a Source of LSP Corpora in the Terminology Classroom." *Wacky! Working papers on the Web as Corpus*. Ed. M. Baroni and S. Bernardini. Bologna: GEDIT, 2006. 159–172.

Dagan, I. and Church, K. "Termight: Identifying and translating technical terminology." *Proceedings of the fourth conference on Applied natural language processing. Association for Computational Linguistics*, 1994. 34–40.

Daille, B. "Study and implementation of combined techniques for automatic extraction of terminology." *Workshop On The Balancing Act: Combining Symbolic And Statistical Approaches To Language,* 1996. 49–66.

Evert, S. *The statistics of word cooccurrences: word pairs and collocations*, Doctoral Dissertation, University of Stuttgart, 2005.

Fantinuoli, C. *InterpretBank: design and implementation of a terminology and knowledge management software for conference interpreters*, Germersheimer Dissertations, 2012.

Fantinuoli, C. "Projekte und Projektionen in der translatorischen Kompetenzentwicklung.". *Einbindung von Korpora im Übersetzungsunterricht als Schlüssel zur Professionalisierung*. Ed. S. Hansen-Schirra and D. Kiraly. Frankfurt: Peter Lang, 2013. 173–188.

Fantinuoli, C. "Specialized corpora from the Web and term extraction for simultaneous interpreters." *Wacky! Working papers on the Web as Corpus*. Ed. M. Baroni and S. Bernardini. Bologna: GEDIT, 2006. 173–190.

Ferraresi A., Bernardini S., Picci G., and Baroni M. "Web corpora for bilingual lexicography: a pilot study of English/French collocation extraction and translation." *Using Corpora in Contrastive and Translation Studies*. Ed. R. Xiao. Newcastle: Cambridge Scholars Publishing, 2010.

Fletcher, W. H. "Making the web more useful as a source for linguistic corpora." *Language and Computers* 52.1 (2004): 191–205.

Frankenberg-Garcia, A. "Raising teachers' awareness of corpora." *Language Teaching* 45.04 (2012): 475–489.

Gallego-Hernández, D. "The use of corpora as translation resources: A study based on a survey of Spanish professional translators." *Perspectives* 23.3 (2015): 375–391.

Gavioli, L. and Aston, G. "Enriching reality: language corpora in language pedagogy." *ELT Journal* 55.3 (2001): 238–246.

Gavioli, L. and Zanettin, F. "Comparable corpora and translation: a pedagogic perspective." Paper presented at *Corpus Use and Learning to Translate (CULT)*. Bertinoro, 1997.

Gellerstam, M. "Translations as a source for cross-linguistic studies." *Lund Studies in English* 88 (1996): 53–62.

Gorjanc, V. "Terminology resources and terminological data management for medical interpreters." *Spürst Du, wie der Bauch rauf-runter? Fachdolmetschen im Gesundheitsbereich*. Ed. D. Andres and S. Pöllabauer. Frankfurt: Peter Lang, 2006. 85–95.

Hansen, S. *Nature of translated text: an interdisciplinary methodology for the investigation of the specific properties of translations*, German Research Center for Artificial Intelligence, Saarland University, 2003.

Hansen-Schirra, S., Neumann, S., and Steiner, E. *Cross-linguistic Corpora for the Study of Translations: Insights from the Language Pair English-German*. Berlin: de Gruyter, 2012.

Hansen-Schirra, S. and Teich, E. "Corpora in human translation." *Corpus Linguistics. An International Handbook*, Vol. 1. Berlin: de Gruyter, 2002. 1159–1175.

Jaaskelainen, R. and Mauranen, A. "Translators at work: a case study of electronic tools used by translators in industry." *Meaningful texts: the extraction of semantic information from monolingual and multilingual corpora*. Ed. G. Barnbrook, P. Danielsson, M. Mahlber. London: Continuum, 2005. 48–53.

Johns, T. "Whence and whither classroom concordancing." *Computer applications in language learning* (1988): 9–27.

Khurshid, A., Gillman, L. and Tostevin, L. "Weirdness Indexing for Logical Document Extrapolation and Retrieval." *Proceedings of the Eighth Text Retrieval Conference (TREC-8)*, 2000.

Koehn, P. "Europarl: A Parallel Corpus for Statistical Machine Translation." *MT Summit*. 2005.

Kübler, N. "New Trends in Corpora and Language Learning." *New Trends in Corpora and Language Learning*. Ed. A. Frankenberg-Garcia, L. Flowerdew, and G. Aston. London: Continuum, 2011. 62–80.

Mauranen, A. and Kujamäki, P. *Translation universals: do they exist?* Amsterdam: John Benjamins, 2004.

MeLLANGE. "Corpora and e-learning questionnaire. Results summary." 2006. Available at: http://mellange.eila.jussieu.fr/Mellange-Results-1.pdf

Picton, A. et al. "Defining the Notion of "Corpora" in Translation: Addressing the Gap between Scholars' and Translators' Points of View." Paper presented at conference *Corpus use and learning to translate (CULT)* Alicante, 2015.

Pöchhacker, F. *Introducing interpreting studies*, London: Routledge, 2009.

Scott, J. "Towards professional uptake of DIY electronic corpora in legal genres." In: M. Sánchez (ed.). *Salford working papers in translation and interpreting,* 2012. Available at: http://www.salford.ac.uk/__data/assets/pdf_file/0010/229492/WorkingPapersT-and-I.Scott.pdf

Shlesinger, M. "Corpus-based interpreting studies as an offshoot of corpus-based translation studies." *Meta: Translators' Journal*, 43.4 (1998): 486–493.

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

Teich, E. *Cross-linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. New York: de Gruyter, 2003.

Tognini-Bonelli, E. *Corpus Linguistics at Work*. Amsterdam: John Benjamins, 2001.

Varantola, K. "Translators and Disposable Corpora. *Corpora in Translator Education*. Ed. F. Zanettin, S. Bernardini, and D. Stewart. Manchester: St. Jerome, 2003. 55-70.

Zanettin, F. Corpora in translation practice." *Proceedings of the First International Workshop on Language Resources (LR) for Translation Work and Research,* 2002a. 10–14.

Zanettin, F. "DIY corpora: the WWW and the translator." *Training the language services provider for the new millennium*. Ed. B. Maia, J. Haller, & M. Ulrych. Universidade do Porto, 2002b. 239–248.

Zanettin, F. *Translation-Driven Corpora*, Manchester: St. Jerome, 2012.

Zanettin, F., Bernardini, S. & Stewart, D. *Corpora in translator education,* St. Jerome, 2003.

Recebido em 04 janeiro de 2016
Aceito em 22 de fevereiro de 2016
Publicado em abril de 2016