Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

BOOTH-FRIENDLY TERM EXTRACTION METHODOLOGY BASED ON PARALLEL CORPORA FOR TRAINING MEDICAL INTERPRETERS

Lorena Arce Romeral and Míriam Seghiri

University of Málaga

Abstract

This article examines the partial results of a comprehensive, three-month study carried out in three. Since Mona Baker laid the foundations of Corpus-based Translation Studies, and as a result of the integration of new technologies in the current educational and professional environment, many proposals have advocated the use of *ad hoc* corpora in translation and interpreting disciplines due to their numerous advantages. These advantages have also been pointed out by researchers such as Laviosa (1998), Bowker (2002), Zanettin et al. (2003), Corpas (2008) or Seghiri (2015, 2017a and 2017b), as corpora are a very valuable source of grammatical, textual or terminological information. This article presents a process to implement a bilingual and bidirectional (English-

Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

Spanish/Spanish-English) glossary based on the compilation and exploitation of an *ad hoc* corpus to address an interpretation of a conference on dysphasia in the framework of a class lecture on interpreting. We illustrate how to semi-automatically extract the terms of the glossary using Terminology Extraction Suite (TES). In order to compile a quality corpus, it is necessary to apply a protocolised and systematic methodology. Therefore, in order to ensure the qualitative representativeness of the corpus, we have established clear design criteria and adapted the Seghiri compilation protocol (2006 and 2012) consisting of four phases—searching, downloading, text formatting and saving data—by adding an alignment phase (Castillo Rodríguez, 2009). We have also determined the quantitative representativeness of the corpus using the ReCor computer application (Seghiri, 2006 and 2015), which is designed specifically for this purpose.

Keywords: corpus linguistics, representativeness, specialised corpora, terminology.

1. INTRODUCTION

The term dysphasia, also called specific language impairment (SLI), was proposed to describe cases in which difficulties in the comprehension and/or expression of language cannot be explained. According to Aguado (2009), however, SLI is due to cognitive delay, morphological or motor alterations of the speech organs, perceptual deficiencies or social disorders. SLI is a developmental disorder that affects 5–7% of the

general population (Tomblin et al., 1997 and Leonard, 1998) and begins in the early stages of development. According to the Specific Language Impairment Association of Madrid (ATELMA), there is a sequential relationship between SLI and other conditions such as autism spectrum disorder (ASD), written language learning disorders and psychological impairment. In this regard, Conti-Ramsden (2002) found that 9% of 242 children with SLI studied from 1997 to 2001 (Nuffield Project) developed ASD over the study period. Moreover, in a 14-year follow-up study, Beitchman et al. (2001) observed that about 35% of young people diagnosed with SLI had psychiatric disorders such as anxiety, social phobia or certain types of antisocial behaviour. Consequently, people directly affected by SLI, as well as their families, undoubtedly need services and tools to enable the early detection and accurate diagnosis of the disorder. Thus, this study focuses on improving communication among the scientific community specialised in these disorders. Moreover, there is a growing need for professional medical translation. According to a study conducted by the Association of Specialised Centres in Translation (ACT, 2005), this type of translation already accounted for 14.6% of translation market demand in Spain in 2005; a figure which continues

to rise in response to the increasing amount of research being conducted in both the public and institutional sectors (Pan American Health Organization, World Health Organization, European Commission Directorate-General for Translation, etc.) and the private sector (e.g. the pharmaceutical industry, hospitals or research centres, etc.).

In the translation and interpreting field, the importance of documentation is evidenced by the presence of documentation content in higher education programmes of study. In Spain, documentation has been a core and compulsory course of the Bachelor's Degree in Translation and Interpreting since the 1980s. The importance of documentation has also been highlighted in the *White Paper on the Bachelor's Degree in Translation and Interpreting* (*Libro Blanco del Título de Grado en Traducción e Interpretación*), which sets the guidelines for curricular design in Spanish universities within the framework of the European Higher Education Area. Training in documentation is essential, since only effective documentation work will ensure correct translations and interpretations in any field of specialisation. The documentary sources available to professional

translators and interpreters are multiple and varied, ranging from terminological sources (glossaries, specialised dictionaries or terminology databases, etc.), consulting with experts, encyclopaedias, institutional sources, lists and discussion forums, manuals, parallel texts and thesauri, to name but a few. However, according to a study by Corpas et al. (2001) carried out among translation and interpreting students at the University of Malaga, Spain, despite the enormous variety of available resources, bilingual dictionaries continue to be the most widely used resource by students, followed far behind by monolingual dictionaries. The same results were obtained by Atkins and Knowles (1990) at the University of Tamepre, Finland, and Mayer (1988) and Roberts (1990, 1992) at the University of Ottawa, Canada. Excessive reliance on dictionaries, glossaries and terminology databases is problematic because these resources present words as isolated units without context. They also lack information on how words are combined. This is compounded by the fact that specialised dictionaries for specific discourse domains are often not available and, if they do exist, they are very deficient, which further justifies the need to learn a flexible, low-cost and user-friendly tool given the speed at which translation and interpreting are

performed. Therefore, the only resource that can offer us such advantages is the corpus, and according to Laviosa (1998), Bowker and Pearson, (2002) and Zanettin et al. (2003), the ideal type of corpus would be—and still is—the so-called *ad hoc* corpus.[1]

2. CORPUS LINGUISTICS IN TRANSLATION AND INTERPRETING DISCIPLINES

The concept of corpus has been addressed by numerous authors. For instance, Sinclair (1991: 171) defined a corpus as '[...] a collection of naturally-occurring language text, chosen to characterize a state or variety of a language'. Translators must make sure that the set of texts they are dealing with constitute a corpus, since it is precisely their representativeness that differentiates them from other types of texts. As Francis (1982: 17) stated, '[...] a corpus is a collection of texts assumed to be representative of a given language, dialect, or other subset of a language to be used for linguistic analysis'. The most accepted definition might be the one provided by EAGLES (1996a: 4), which identifies the three characteristics that

[1] An *ad hoc* corpus is also referred to as a corpus for specific, virtual, electronic, disposable or web purposes, among others.

differentiate a corpus from a set of texts: '[…] a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language'.

Since Mona Baker laid the foundations of the so-called Corpus-based Translation Studies, many proposals have supported the use and study of corpora in the field of translation and interpreting, and corpus linguistics in translation and interpreting studies has now become a consolidated line of research. As a result, there is a vast body of scientific literature that has examined the specific characteristics of different genres (Corpas, 2008; Sánchez Ramos and Vigier Moreno, 2016), as well as their pedagogical applications (Monzó Nebot, 2008 and Zanettin, 2003) or the use of corpus as a documentary resource in professional environments (Gallego-Hernández, 2015). Moreover, several studies have shown interesting results on the habits and uses of electronic documentary tools not only among translation and interpreting students (Cid-Leal and Perpinya-Morera, 2015), but also among professionals (Désinales et al., 2009). These works conclude that both students and professionals mainly use electronic resources and therefore training in this type of tools, as well as the efficient

search for them, must be incorporated into the training of translators and interpreters. In this context, the integration of information and communication technologies (ICT) has changed the approach of lecturers, professional translators and interpreters and students in these disciplines. Indeed, a large number of authors, such as Laviosa (1998), Bowker and Pearson (2002), Zanettin et al. (2003), Bernardini and Castagnoli (2008) or Corpas (2001 and 2008) have highlighted the virtues of using *ad hoc* corpora for the teaching and learning of translation and interpreting. According to these authors, corpora—as a specialised grammatical and discursive, lexicographic, terminological and cognitive resource—constitute a macro source of documentation. Corpora also provide models and patterns that guide translators or interpreters in their decision-making processes at the macro- and micro-structural level

However, despite the numerous advantages of using *ad hoc* corpora in translation and interpreting, the main problem, as Seghiri (2010) has stated, is that specialised corpora which have already been compiled are not available on the Internet, or if they do exist, they would hardly satisfy all documentation needs. Given this situation,

translators and interpreters have no alternative but to compile their own *ad hoc* corpora. In this study, we present a methodology to extract bilingual terminology from a parallel (Spanish-English) *ad hoc* corpus. The process will be exemplified in the context of a practical class in interpreting in which students must interpret a conference on dysphasia (SLI).

3. CREATING A GLOSSARY FOR INTERPRETERS BASED ON THE COMPILACION OF AN *AD HOC* CORPUS

The following section describes a method for creating a bilingual and bidirectional glossary based on the compilation of a parallel *ad hoc* corpus in the Spanish-English language pair that can be used for specialised interpreters in the field of medicine, specifically on dysphasia (SLI).

3.1 Compilation of an *ad hoc* corpus: determining qualitative representativeness

According to Seghiri (2006), in order for a collection of texts to be considered a corpus, it must be compiled according to specific parameters so that it can represent a state or a section of a language.

The author also points out that in order for a corpus to be representative, it must be correctly designed and the documents that compose it must be selected according to a specific design criteria and an appropriate compilation protocol. Thus, this method is divided into two well-differentiated phases, which will ensure the qualitative representativeness of the corpus.

3.1.1 Design criteria

Before starting the compilation process, it is essential to establish clear *design criteria.* With regard to the topic of the interpretation, we have used an interpretation dealing with dysphasia as an example. The corpus is comprised of abstracts drawn from research articles on this disorder, so it is completely *homogeneous* in terms of content. The corpus is fed exclusively by electronic resources, so it is *virtual*. It is also *parallel*, *bilingual* and *monodirectional*, as it includes original articles in Spanish and their translations into English. In addition, it is *partial* (we only include the abstracts of the articles) and *unbalanced* in terms of the number of documents, since the compilation is not determined by the availability of the texts on the web.

3.1.2. Compilation protocol

Once the initial design parameters have been established, the translator must follow a *protocolised methodology* for compiling the corpus. In our case, we have used an adapted version of the compilation protocol of Seghiri (2006 and 2012) comprising four phases: searching, downloading, text formatting and saving data. These typical phases to compile a comparable corpus are followed by a fifth step, alignment (Castillo Rodríguez, 2009), which is necessary for the subsequent management and exploitation of the bitexts in the Terminology Extraction Suite program.

The first phase consists of *searching* for the documents on the Internet. In this sense, the ability to identify the desired information on the web depends largely on the accuracy and effectiveness with which search engines are used by the translator and interpreter. Thus, an accurate and effective search does not consist of using the search engine *per se*, nor in reading multiple documents until we find the ones we are looking for, but in learning to locate the information with the necessary precision. To prevent the problem of retrieving an excessive

number of documents on the Internet that were not valid for our corpus, we performed two types of searches: an institutional search and a search using the descriptors and equations provided in the Google Advanced Search option. As regards the institutional search, we retrieved scientific and academic journals from the Scielo and Neurology Journals databases (see Figure 1).

Figure 1. Institutional search

Secondly, the search based on descriptors and search equations in Google's Advanced Search option was highly efficient, fast and simple. Among other functions, this tool allows restricting the search by language (Spanish and English) and geographical area. As far as the search format is concerned, we

opted not to specify it in order to obtain as many documents as possible. It should also be noted that it is essential to use clear and appropriate keywords or search descriptors to obtain the largest number of representative samples for the corpus and avoid the so-called 'documentary noise'. Table 1 shows the main descriptors and search equations used to restrict and specify the search results and obtain the maximum number of documents in accordance with the corpus design criteria.

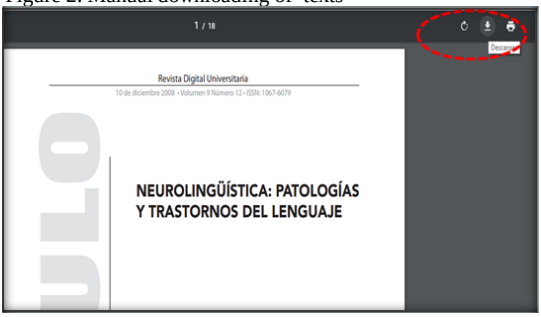
Table 1: Keywords and search descriptors used to access the information

	Text type	Keywords	Search descriptors
Spanish	Resumen artículo científico	Resumen, artículo científico, disfasia, trastrono específico del lenguaje, TEL	"resumen" AND "artículo científico" AND "disfasia" AND "trastorno específico del lenguaje" AND "TEL"
English	Abstract of scientific article on dysphasia	Abstract, scientific article, dysphasia, specific language impairment, SLI	"Abstract" AND "scientific article" AND "dysphasia" AND "specific language impairment" AND "SLI"

After locating the documents to compile the corpus, they are *downloaded*. Although documents are

usually downloaded manually (see Figure 2), this task can be automated for groups of pages using programs that allow them to be downloaded in batches, such as GNUWget[2] or GetBot[3] (see Figure 3 and Figure 4). With regard to this last phase, it is necessary to mention the multitude of formats in which the samples that comprise the corpus can be found (.pdf, .doc, .html, etc.), which is why the next step is necessary.

Figure 2. Manual downloading of texts



[2] Available at: <https://www.gnu.org/software/wget/>.
[3] Available at: <http://www.getbot.com/>.

Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

Figure 3. GNU Wget interface (Batch Download)

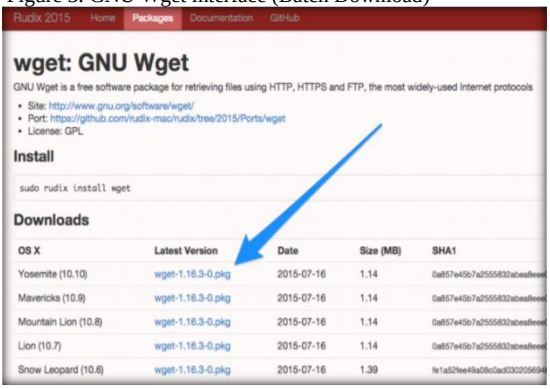
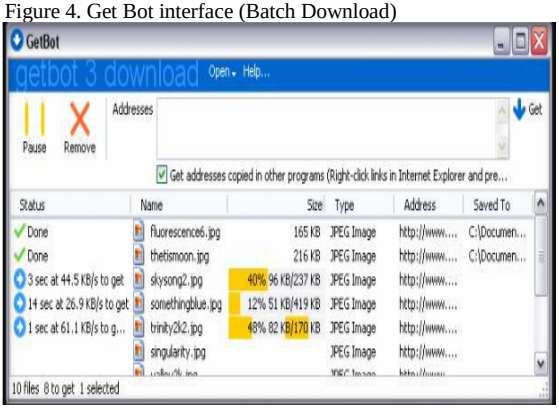


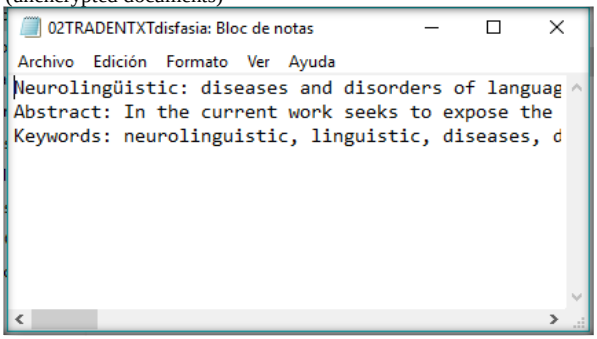
Figure 4. Get Bot interface (Batch Download)



The documents that we located and downloaded in the preceding steps can be found on the Internet, usually in .html, .doc, docx or .pdf formats. However, corpus management programs generally

only work in ASCII or plain text format (.txt). For this reason, it is necessary to perform a *format* conversion process. The procedure proposed by Seghiri (2012: 376) can used for this purpose: 'the conversion from any format to plain is as easy as to copy and paste it into a plain text document (.txt)' (see Figure 5), as long as the texts in .pdf format are not encrypted. However, if they are encrypted, it is necessary to use online programs such as freepdfconvert[4], documento.online-conver[5] or the powerful Abbyy Fine Reader[6], to mention some of the most common ones (see Figure 6).

Figure 5. Conversion to .txt format by copying to notepad (unencrypted documents)



[5] Available at: <https://documento.online-convert.com/es>.
[6] Available at: <https://www.abbyy.com/es-es/finereader/>.
[7] Available at: <http://rename.lupasfreeware.org/>.

Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

Figure 6. PDF Converter interface (conversion to. txt format for encrypted documents)

The last phase, *saving data*, consists of storing the documents in folders and subfolders. In this process, it is necessary to establish a clear code that permits the texts to be properly stored and easily located, as well as the possible extension of the corpus. To do so, we have first created a folder labelled 'Dysphasia (SLI)'. Two folders were then created within the first folder: one of which was labelled 'OT' and contained the original documents in Spanish (ES) and a second folder labelled 'TT' where the target texts were saved (i.e. the translations into English (EN). Two more folders were then created in these two subfolders: one which was labelled 'OF', which contained texts in their original format (.pdf, .html, etc.), and another labelled 'TXT', which contained

the abstracts in plain text. Finally, the topic is indicated: dysphasia (SLI). Table 2 shows the coding used for the data saving phase. Although corpus coding can be done manually, we have used the automatic coding program Lupas Rename[7].

Table 2. Coding process of the compiled corpus

DYSPHASIA (DSL)	**OT**	ES	OF	0OTESOFDSL
				02OTESOFDSL
				03OTESOFDSL
				…
			TXT	0OTESTXTDSL
				02OTESTXTDSL
				03OTESTXTDSL
				…
	TT	EN	OF	01TTENOFDSL
				02TTENOFDSL
				03TTENOFDSL
				…
			TXT	01TTENTXTDSL
				02TTENTXTDSL
				03TTENTXTDSL
				…

After completing the four steps to compile the corpus (searching, downloading, text formatting and saving data) , we obtain a parallel and bilingual corpus consisting of 20 original abstracts from scientific articles in Spanish (3246[8] words or

[7] Available at: <http://rename.lupasfreeware.org/>.
[8] To count the number of words we have used Word Count Tool available at:

tokens) and their corresponding translations in English (2488 words or tokens).

Finally, to exploit the samples of texts using the parallel corpus management program (TES[9]), it is necessary to align the corpus. Although many corpus alignment programs are currently available, we have used LF Aligner.[10] For the alignment process, the first step is to specify the format and coding of the texts (in this case, plain text with UTF-8 coding). Then, the language pair of the texts to be aligned is specified and, finally, the documents to be aligned are selected (see Figure 7 and Figure 8).

<http://wordcountool.com/>.
[9] Available at: <https://sourceforge.net/projects/terminology-extraction-suite/>.
[10] Available at: <https://sourceforge.net/p/aligner/wiki/Home/>.

Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

Figure 7. LF Aligner program interface (I)

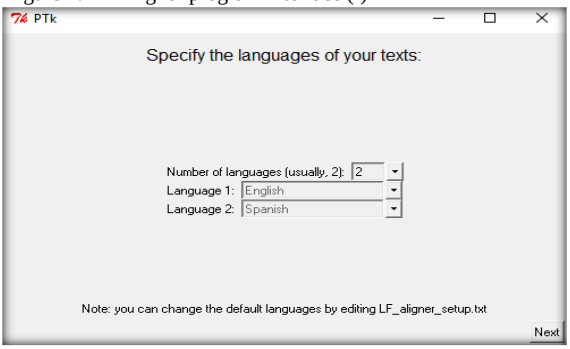
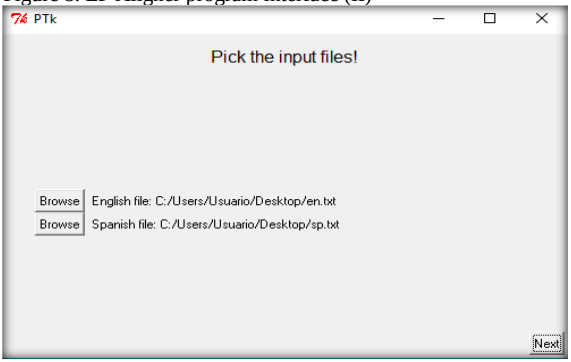


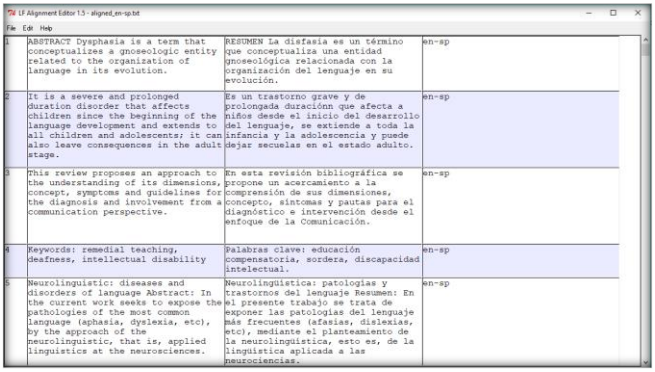
Figure 8. LF Aligner program interface (II)



Once the alignment is completed, LF Aligner displays how many lines have been created for each language (see Figure 9).

Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

Figure 9. Corpus alignment (Spanish-English) with LF Aligner

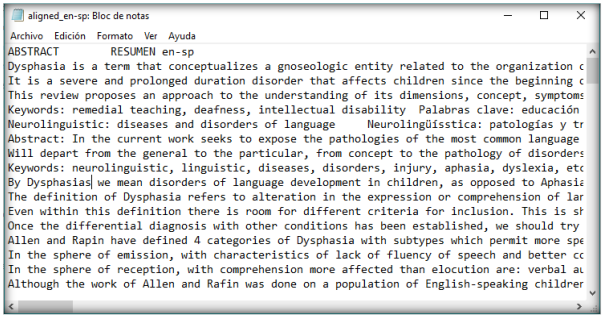


1	ABSTRACT Dysphasia is a term that conceptualizes a gnoseologic entity related to the organization of language in its evolution.	RESUMEN La disfasia es un término que conceptualiza una entidad gnoseológica relacionada con la organización del lenguaje en su evolución.	en-sp
2	It is a severe and prolonged duration disorder that affects children since the beginning of the language development and extends to all children and adolescents; it can also leave consequences in the adult stage.	Es un trastorno grave y de prolongada duraciónn que afecta a niños desde el inicio del desarrollo del lenguaje, se extiende a toda la infancia y la adolescencia y puede dejar secuelas en el estado adulto.	en-sp
3	This review proposes an approach to the understanding of its dimensions, concept, symptoms and guidelines for the diagnosis and involvement from a communication perspective.	En esta revisión bibliográfica se propone un acercamiento a la comprensión de sus dimensiones, concepto, síntomas y pautas para el diagnóstico e intervención desde el enfoque de la Comunicación.	en-sp
4	Keywords: remedial teaching, deafness, intellectual disability	Palabras clave: educación compensatoria, sordera, discapacidad intelectual.	en-sp
5	Neurolinguistic: diseases and disorders of language Abstract: In the current work seeks to expose the pathologies of the most common language (aphasia, dyslexia, etc), by the approach of the neurolinguistic, that is, applied linguistics at the neurosciences.	Neurolinguistica: patologías y trastornos del lenguaje Resumen: En el presente trabajo se trata de exponer las patologías del lenguaje más frecuentes (afasias, dislexias, etc), mediante el planteamiento de la neurolingüística, esto es, de la lingüística aplicada a las neurociencias.	en-sp

LF Aligner has an option that permits reviewing the alignment manually (which can be done with either the program's own graphic editor or with an Excel file). It is possible to combine or separate the paragraphs (Merge and Split), as well as move them up (Shift up) or down (Shift down), delete cells or modify the aligned segments. After completing the alignment review, LF Aligner can create a file with .tmx extension (translation memory) that can be used in computer-aided translation (CAT) programs or extract the aligned documents in plain text format, which is what we used for the management of the corpus (see Figure 10).

Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

Figure 10. Corpus aligned with LF Aligner (.txt format)



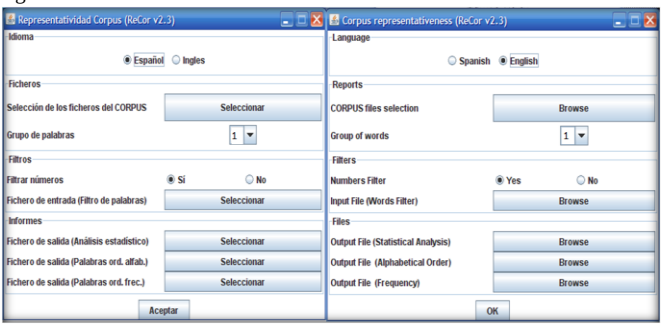
By following these five steps (searching, downloading, text formatting, saving data and alignment) and taking into account the previous design criteria, the quality of the corpus documents is ensured, in other words, it is a representative corpus from a qualitative point of view.

3.2 Determining quantitative representativeness

Although the set of texts we have obtained is representative from a qualitative point of view, it is necessary to verify whether the corpus is representative from a quantitative point of view, that is, whether the compiled documents cover the basic terminology of the field of specialty: dysphasia

(SLI). To do so, the ReCor[11] program (version 2.3) was used. Figure 11 shows how the program can be used to determine if the corpus is representative in quantitative terms. As shown in the figure, the first step is to load the two subcorpora.

Figure 11. ReCor interface

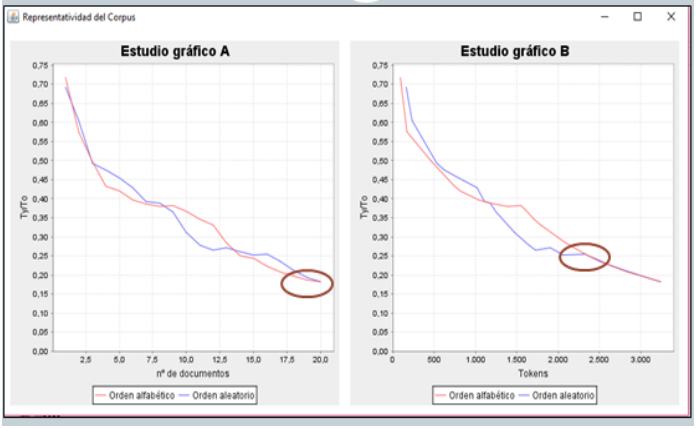


The result of both analyses is presented through graphical representations as output files in .txt format. In particular, it is assumed that the coefficient between real words of a text and total words (types/tokens)—i. e. the density or lexical richness of a text—does not increase proportionally from a certain number of texts (see figures 12A and 13A). The same applies when representativeness is

[11] Recor is an effective solution to determine a posterior, for the first time, the minimum size of a corpus or textual collection, regardless of the language or textual genre of that collection, establishing, therefore, the minimum threshold of representativeness through an algorithm (N-Cor) of analysis of the lexical density as a function of the incremental increase of the corpus.

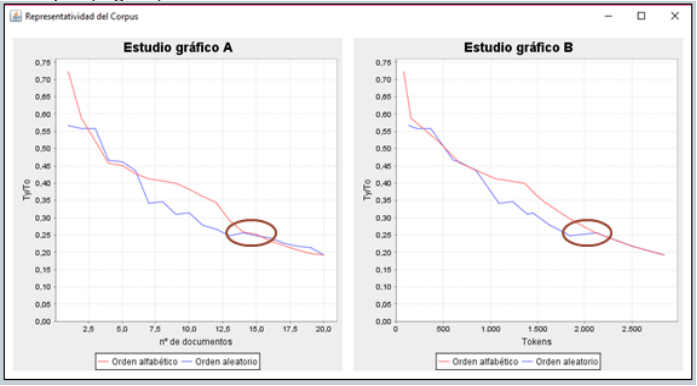
calculated on the basis of lexical density from word sequences or n-grams (see figures 12B and 13B). Each of the figures shows two lines that represent the documents ordered alphabetically (red line) and randomly (blue line). The lines merge together and stabilise as they approach the value of zero, which indicates the minimum size for the collection to be considered representative. Thus, figures A and B graphically illustrate the point at which the qualitative criteria begins to be representative in quantitative terms. We then proceed to determine the representativeness of each of the two subcorpora.

Figure 12. Determination of the quantitative representativeness of the Spanish subcorpus (1-gram)



Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

Figure 13. Determination of the quantitative representativeness of the English subcorpus (1-gram)



As can be seen in the figures, the Spanish subcorpus begins to be representative at 20 documents and 2200 tokens (see Figure 12), while the English subcorpus begins to be representative at 13 documents and 2000 tokens (see Figure 13). Therefore, the ReCor program has shown that the compiled corpus is not only qualitatively but also quantitatively representative, so it is ready to be exploited and managed for the subsequent semi-automatic extraction of terminology units that will be included the interpretation glossary.

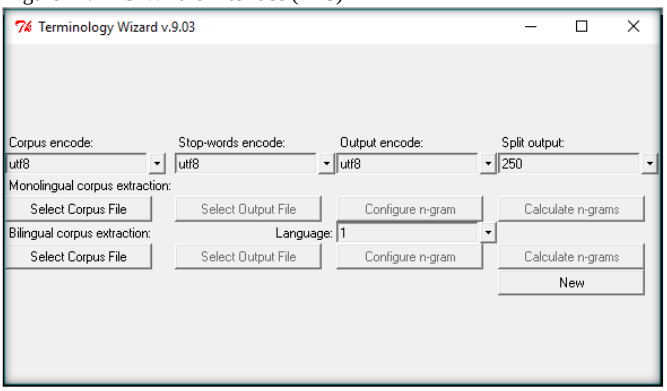
Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

3.3 Creating a bilingual glossary from an ad hoc corpus

Once we have compiled a representative corpus in both qualitative and quantitative terms, the next step is to exploit and manage the corpus to create a bilingual and bidirectional glossary on dysphasia (SLI) for medical interpreters. For the semi-automatic extraction of the glossary terms we have used Terminology Extraction Suite (TES), which is comprised of two smaller software applications: TES-Wizard and TES-Editor. Thus, the entire terminology extraction process must be done with two programs: TES-Wizard, which extracts candidate terms from a monolingual or bilingual corpus, followed by TES-Editor, which edits candidate terms and, if a parallel corpus is available, automatically searches for translation equivalents. When TES-Wizard is run, a screen like the one in Figure 14 will be displayed.

Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

Figure 14. TES-Wizard interface (TES)



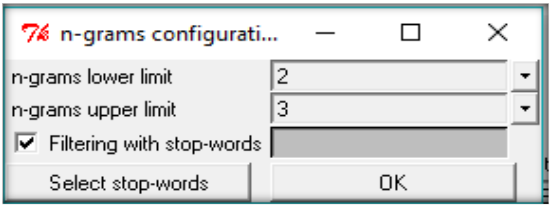
The 'Corpus encode', 'Stop-words encode' and 'Output encode' list boxes are used to select the character set that matches the corpus, the stop-words[7] list[12] and the output file, respectively. 'Split output' is used to choose the size of the partition allocated to the output file. Since a considerable number of candidates may be obtained from a corpus and editing files that are too large with TES-

[12] Stop-words lists are especially useful for creating glossaries and comprised of words that are empty of meaning (i.e. defined, indefinite, numerals or adverbs, etc.) and words with very general content. Although stop-words lists are available in different languages on the Internet, users can make their own stop-words list manually. To do so, an exclusion list must be created in a plain text file (.txt) with the words that to the user does not want to appear on the list. The words of the exclusion list must be separated from each other by commas (,) or paragraph breaks (¶). In our case, we have used the stop-words lists (in both English and Spanish) included in Terminology Extraction Suite.

Editor is cumbersome, it is possible to select the file partition in several candidate values (or decide not to partition the file). Either way, even if a partition size is selected, a complete file will also be generated. Once the previous configurations are determined, given the characteristics of our corpus, the terminology of the bilingual corpus is extracted as follows. First, click on 'Select Corpus File'. A dialog box will immediately open to indicate the corpus file. Bilingual corpora must be aligned in parallel in text format (.txt) and separated by tabs. Once the corpus has been uploaded, the following buttons can be activated or deactivated during the process. Secondly, to extract terms from the parallel corpus, select 'Language 1' or 'Language 2' to choose the second language. Thirdly, by clicking on 'Select Ouput File', a dialog box will open that allows the user to select the output file. If a 'Split output' value has been selected, a set of files with the same name but ending in part0, part1, etc., will be generated. Finally, click on 'Configure n-grams' to open a configuration screen for calculating n-grams. The lower and upper n and the stop-words list must be selected in this screen (see Figure 15).

Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

Figure 15. Configuration of n-grams and selection of the stop-words list (TES-Wizard)

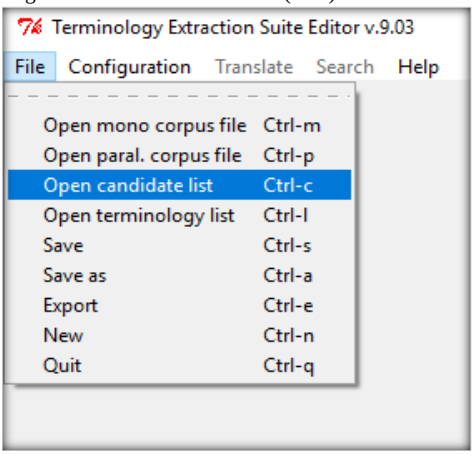


To end the process, click on 'Calculate n-grams'. This will start the n-grams calculation process and a progress bar will be displayed to indicate the status of the process. Once the process is complete, the file of candidates will be ready. To start a new extraction process, click on 'New'.

The aim of TES-Editor is to edit the term candidates extracted with TES-Wizard. When the program is run, a screen like the one in Figure 16 will be displayed.

Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

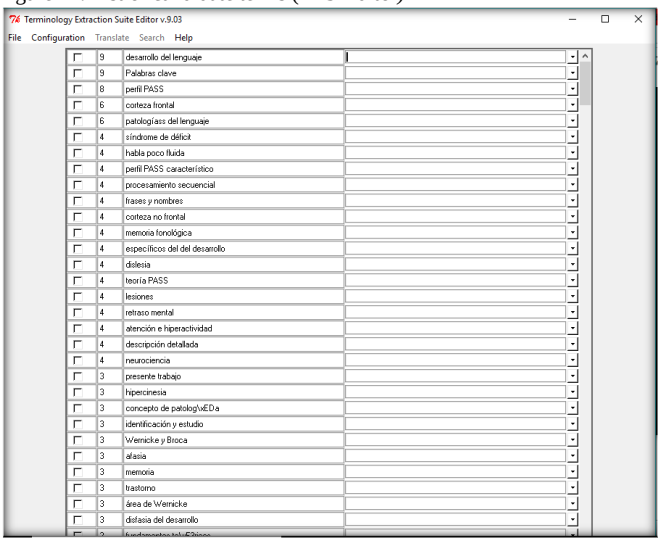
Figure 16. TES-Editor interface (TES)



The types of files that can be opened from the toolbar are monolingual corpus ('Open mono corpus file'), bilingual corpus ('Open paral. corpus file'), the list of candidate terms extracted using TES-Wizard ('Open candidate list') and a terminology list in text format with one term per line ('Open terminology list'). By selecting 'Open candidate list' (see Figure 17), a list of candidates terms for the files (part0, part1, etc.) will appear on the screen. The TES-Wizard division option also allows working with smaller word lists. When

'Open candidate list' is selected, a screen like the one in Figure 17 will be displayed.

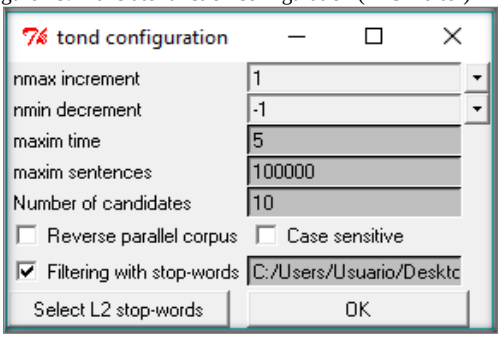
Figure 17. List of candidate terms (TES-Editor)



In the first column there are a series of boxes to choose the term to be exported. The second column shows the term frequency. The third column shows the term or candidate term. Finally, the fourth column is reserved for the results of the automatic search for translation equivalents. To edit the list in TES-Editor, select the relevant candidate terms to export a list of terms. By opening the parallel

corpus, the word list can be translated to automatically search for translation equivalents. To use the 'Translate' function, it must be set by clicking on Configuration>Translate. The following window will appear to where the desired parameters can be chosen (see Figure 18).

Figure 18. Translate function configuration (TES-Editor)

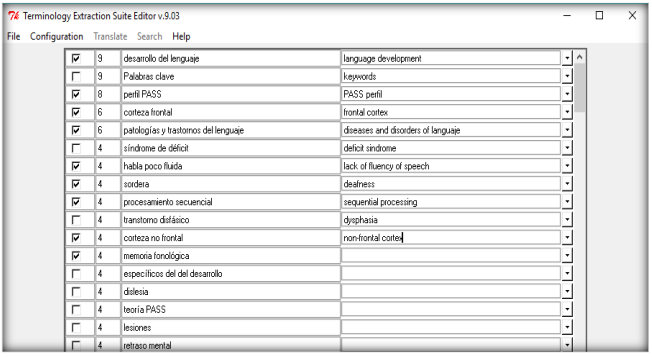


The 'Nmax increment' option indicates the maximum increase in n compared to the original term's n. 'Nmin decrement' permits selecting the decrease in n compared to the original term's n. 'Maxim time' is the maximum time spent by the algorithm to find the equivalent translation. 'Maxim sentences' is the maximum number of parallel corpus sentences that will be read before returning a possible equivalent. 'Number of candidates' indicates how many candidates will be displayed in the dropdown box. 'Reverse parallel corpus' is used to reverse the order of the languages in the parallel

corpus. 'Case sensitive' permits distinguishing between uppercase and lowercase. The 'Select L2 stop-words' option opens the stop-words file corresponding to the target language (English in our case). Finally, tick the 'Filtering with stop-words' option and click on 'OK' to accept all the options.

After making these selections, the equivalent translation can be searched for automatically by clicking on the translate menu. The program will display the most probable translation equivalent, but a list of candidate terms can also be displayed and a search for other possible translations can be performed. It is also possible to write directly in the box (see Figure 19).

Figure 19. Management of list of candidate terms (TES-Editor)

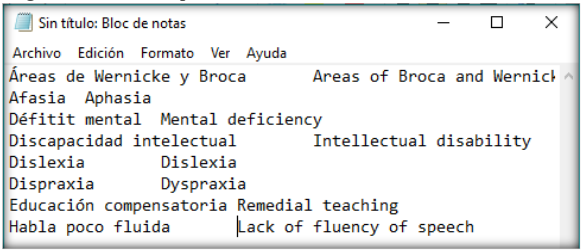


When searching for the translation equivalent, the term is automatically marked by its export. If

desired, the box can be unchecked. Once the process is finished, there are three options: a) 'File>Save' and 'File>Save As' to save the candidate terms file and continue to work with it later on; b) 'File>Export' to export the search results or c) 'File>New' to prepare the program to edit a new file.

In our case, we chose the second option (File>Export). A browser will open to select the location of the generated glossary, name it and save it in plain text format. After carrying out these steps, the result is a document in which the Spanish terms appear on the left and their English translation equivalents on the right, separated by tabulation (see Figure 20).

Figure 20. Terms exported from TES-Editor



The document in .txt format exported with TES-Editor is then used to copy the terms (in Spanish) and their equivalents (in English) to paste them into a MicroSoft Excel sheet. Once copied, other useful

fields can be included if desired. In our case, we have added a third column containing pronunciation.[13] in the target language (English). If several glossaries have been generated from the folder division (part0, part1, part2, etc.) created by TES-Wizard, all individual glossaries must be pasted into the Excel sheet and then sorted alphabetically (see Figure 21).

Figure 21. Spanish-English glossary

	A	B	C
1	Áreas de Wernicke y Broca	**Areas of Broca and Wernicke**	| ˈeəriəz əv <broca> ənd ˈwɜːnɪkə |
2	Agnosia auditiva	**Verbal auditory**	| ˈvɜːbl ˈɔːdɪtəri |
3	Afasia	**Aphasia**	| əˈfeɪzɪə |
4	Déficit mental	**Mental deficiency**	| ˈmentl dɪˈfɪʃnsi |
5	Discapacidad intelectual	**Intellectual disability**	| ˌɪntəˈlektʃuəl ˌdɪsəˈbɪlɪti |
6	Dislexia	**Dislexia**	| <dislexia> |
7	Dispraxia	**Dyspraxia**	| <dyspraxia> |
8	Educación compensatoria	**Remedial teaching**	| rɪˈmiːdɪəl ˈtiːtʃɪŋ |
9	Habla poco fluida	**Lack of fluency of speech**	| læk əv ˈfluːənsi əv spiːtʃ |
10	Hipercinesia	**Hyperkinesia**	| <hyperkinesia> |
11	Lesiones	**Injury**	| ˈɪndʒəri |
12	Lesiones cerebrales focales	**Focal cerebral lesions**	| ˈfəʊkl ˈserəbrəl ˈliːʒnz |
13	Memoria fonológica activa	**Phonological working memory**	| ˌfəʊnəˈlɒdʒɪkl ˈwɜːkɪŋ ˈmeməri |
14	Neurociencia	**Neuroscience**	| ˈnjʊərəʊsaɪəns |
15	Neurolingüística	**Neorolingüistic**	| <neorolingüistic> |
16	Patologías y trastornos del lenguaje	**Diseases and disorders of language**	| dɪˈziːzɪz ənd dɪsˈɔːdəz əv <languaje> |
17	Recuerdo de series	**List recall**	| lɪst rɪˈkɔːl |
18	Retraso mental	**Mental retardation**	| ˈmentl ˌriːtɑːˈdeɪʃn |
19	Síndrome de atención deficitaria	**Cognitive función developemental disorder**	| ˈkɒgnətɪv <función> <developemental> dɪsˈɔːdə |
20	Síndrome de déficit fonológico sintáctico	**Syntactic phonological deficit sindrome**	| sɪnˈtæktɪk ˌfəʊnəˈlɒdʒɪkl ˈdefɪsɪt <sindrome> |
21	Síndrome de déficit sintáctico lexical	**Lexical syntactic deficit sindrome**	| ˈleksɪkl sɪnˈtæktɪk ˈdefɪsɪt <sÍndrome> |
22	Sordera	**Deafness**	| ˈdefnəs |
23	Transtorno específico del lenguaje (TEL)	**Specific Language Disorder (SLD)**	| spəˈsɪfɪk <language> dɪsˈɔːdə |
24	Transtorno disfásico	**Dysphasia**	| <dysphasia> |
25	Trastorno de hiperactividad	**Hyperactivity disorder**	| ˌhaɪpərˈæktɪv dɪsˈɔːdə |

As can be seen in the figure, we have obtained a bilingual and monodirectional glossary (Spanish-English) from a bilingual parallel corpus that was aligned using LF Aligner and managed with

Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

Terminology Extraction Suite. In order to obtain the English-Spanish glossary, the columns must be ordered in Excel. To do so, select the column containing the English translation equivalents and cut it. Then right-click on column A (where the terms are in Spanish) and choose the option 'Insert cut cells'. The terms in English will appear in the left-hand column and the terms in Spanish in the right-hand column. Finally, as with the Spanish-English glossary, the entries are then ordered alphabetically and their corresponding phonetic transcriptions are included.[14]

Figure 22. English-Spanish glossary

Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

	A	B	C
1	Areas of Broca and Wernicke	**Áreas de Wernicke y Broca**	[ˈɾe az ðe we r ˈnik ke i ˈβro ka]
2	Aphasia	**Afasia**	[a ˈfa sja]
3	Cognitive función developemental disorder	**Síndrome de atención deficitaria**	[ˈsiN dro me ðe a teN ˈθjoN de fi θi ˈta rja]
4	Deafness	**Sordera**	[sor ˈðe ra]
5	Diseases and disorders of languaje	**Patologías y trastornos del lenguaje**	[pa to lo ˈxi as i tras ˈtor nos ˈðel leŋ ˈgwa xe]
6	Dislexia	**Dislexia**	[diz ˈlek sja]
7	Dysphasia	**Transtorno disfásico**	[trans ˈtor no ðis ˈfa si ko]
8	Dyspraxia	**Dispraxia**	[dis ˈprak sja]
9	Focal cerebral lesions	**Lesiones cerebrales focales**	[le ˈsjo nes θe re ˈβra les fo ˈka les]
10	Hyperactivity disorder	**Trastorno de hiperactividad**	[tras ˈtor no ðe j pe rak ti βi ˈðad]
11	Hyperkinesia	**Hipercinesia**	[i per θi ˈne sja]
12	Injury	**Lesiones**	[le ˈsjo nes]
13	Intellectual disability	**Discapacidad intelectual**	[dis ka pa θi ˈðað iN te lek ˈtwal]
14	Lack of fluency of speech	**Habla poco fluida**	[ˈa βla ˈpo ko ˈfluj ða]
15	Lexical syntactic deficit síndrome	**Síndrome de déficit sintáctico lexical**	[ˈsiN dro me ðe ˈðe fi θi t siN ˈtak ti ko lek si ˈkal]
16	List recall	**Recuerdo de series**	[ře ˈkwer ðo ðe ˈse rjes]
17	Mental deficiency	**Défitit mental**	[ˈde fi θi t meN ˈtal]
18	Mental retardation	**Retraso mental**	[ře ˈtra so meN ˈtal]
19	Neorolingüistic	**Neurolingüística**	[new ro liŋ ˈgwis ti ka]
20	Neuroscience	**Neurociencia**	[new ro ˈθjeN θja]
21	Phonological working memory	**Memoria fonológica activa**	[me ˈmo rja fo no ˈlo xi ka ak ˈti βa]
22	Remedial teaching	**Educación compensatoria**	[e ðu ka ˈθjoŋ kom pen sa ˈto rja]
23	Specific Languaje Disorder (SLD)	**Transtorno específico del lenguaje (TEL)**	[trans ˈtor no es pe ˈθi fi ko ˈðel leŋ ˈgwa xe ˈtel]
24	Syntactic phonological deficit síndrome	**Retraso mental**	[ře ˈtra so meN ˈtal]
25	Verbal auditory	**Agnosia auditiva**	[ag ˈno sja aw ðiˈti βa]

As shown in Figure 22, we have also obtained an English-Spanish bilingual glossary from the bilingual parallel corpus. Thus, the final result is a two-directional, bilingual glossary (Spanish-English/English-Spanish) on dysphasia (SLI) comprising a total of 68 terms in each language.

4. CONCLUSIONS

Parallel corpora are particularly useful for meeting interpreters' documentation needs. A representative and properly managed corpus is a very effective tool

for identifying, extracting and translating lexical units in the form of a bilingual glossary to help interpreters in the research and documentation process prior to and during interpretation. The advantages of using corpora in interpreting are undeniable due to their objectivity and reusability for multiple purposes. Corpora are also easy to use and allow accessing and managing large amounts of information in a matter of seconds.

In this paper, we have described a protocolised method for terminology extraction based on a bilingual parallel corpus in order to generate a glossary on dysphasia that can be of use in medical interpreting. In order to determine the qualitative representativeness of the texts, we have first taken into account the design criteria of the corpus and, secondly, adapted the Seghiri compilation protocol (2006 and 2015), which consists of four phases—searching, downloading, text formatting and saving data. A fifth alignment phase was added to the compilation protocol following Castillo Rodríguez (2009).

With a view to the subsequent management of the corpus, it was aligned using the LF Aligner program, resulting in a corpus of bitexts formed by ten

abstracts drawn from scientific articles in Spanish and their corresponding translations in English. In addition, the quantitative representativeness of the corpus was determined using the ReCor program. Terminology Extraction Suite (TES) was used to extract the terminology in both languages and export the candidate terms to implement a glossary. The process has resulted in a bilingual and bidirectional glossary (Spanish-English/English-Spanish) consisting of 68 terms in each language. The pronunciation of the terms in English and Spanish was also included in the glossary using the automatic phonetic transcribers PhoTransEdit[13] and Aucel[14], respectively.

REFERENCES

Cronin, M. (2003). *Translation and Globalization*. London: Routledge.

Aguado, G. (2009). El trastorno específico del lenguaje (TEL): un trastorno dinámico. *Audición y Lenguaje. Revista de la Federación Española de Profesores de Audición y Lenguaje, 88*, 13-22.

[13] Available at: <http://www.photransedit.com/>.
[14] Available at: <http://www.aucel.com/pln/transbase.html>.

Atkins, B.T.S. & Knowles, Frank E. (1990). Interim report on the Eurlex/AILA research project into dictionary use. In T. Magay and J. Zigány (Eds.), *Budalex '88 procedings: Papers from the Eurlex third international congress* (pp. 381-392). Budapest: Akadémiai Kiado.

Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and Applications. In M Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology: in Honour of John Sinclair* (pp. 233-250). Amsterdam and Philadelphia: John Benjamins.

Beitchman, J.H., Wilson, B. Johnson, C.J., Atkinson, L., Young, A., Adlaf, E., Escobar, M. & Douglas, L. (2001). Fourteen-year follow-up of speech/language impaired and control children: Psychiatric outcome. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40*, 75-82.

Bernardini, S. & Castagnoli, S. (2008). Corpora for Translation Education and Translation Practice. In E. Yuste Trigo (Eds.), *Topics in Language Resources for Translation and Localization* (pp. 39-55). Amsterdam & Philadelphia: John Benjamins.

Bowker, L. & Pearson, J. (2002). *Working with*

Specialized Language: A practical guide to using corpora. London: Routledge.

Castillo Rodríguez, C. (2009). La elaboración de un corpus *ad hoc* paralelo multilingüe. *Revista Tradumática*, *7,* 1-11.

Cid-Leal, P. & Perpinyá-Morera, R.(2015). Competencia informacional en Traducción: análisis de los hábitos de los estudiantes. *BiD: textos universitarios de biblioteconomía y documentación* (34).

Conti-Ramsden, G. (2002). Continuidad académica y educativa en niños con trastorno específico del lenguaje (TEL). *Revista Chilena de Fonoaudiología*, *3*, 25-38.

Corpas Pastor, G. (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfurt am Main: Peter Lang.

Corpas, G., Leiva, J & Varela, M.J. (2001). El papel del diccionario en Traducción e Interpretación: análisis de necesidades y encuestas de uso. In M. C. Ayala Castro (Eds.), *La utilidad de los diccionarios para la enseñanza de las lenguas* (pp. 237-272). Sevilla: Servicio de Publicaciones de la

Universidad de Sevilla.

Désilets, A., Melaconc, Ch., Patenaude, G. & Brunette, L. (2009). How translators use tools and resources to resolve translation problems: an etnography study. Actas del Congreso Beyond Translation Memories. NRC Publications Archive (NPArC).

EAGLES (1996a). Preliminary Recommendations on Corpus Typology, documento EAGLES (Expert Advisory Group on Language Engineering) EAG-TCWG-CTYP/P.

EGLES (1996b). Text Corpora Working Group Reading Guide, documento EAGLES (Expert Advisory Group on Language Engineering) EAG-TCWG-FR-22.

Francis, W. Nelson (1982). Problems of assembling and computerizing large corpora. In Stig Johansson (Eds.), *Computer Corpora in English Language Research* (pp. 7-24). Bergen: Norwegian Computing Centre for the Humanities.

Gallego- Hernández, D. (2015). The use of corpora as translation resources. A study based on a survey of Spanish professional translators. *Perspectives: Studies in Translatology*, *23* (3), 375-371.

Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

Laviosa, S. (1998). The Corpus-based Approach: A New Paradigm in Translation Studie. *Meta: Translator's Journal*, *43* (4), 474-479.

Leonard, L.B. (1998). *Children with specific language impairment*. Cambridge: MIT Press.

Meyer, I. (1988). The General Bilingual Dictionary as a Working Tool in Thème. *Meta, 33* (3), 368-376.

Monzó Nebot, E. (2008). Corpus-based Activities in Legal Translator Training. *The Interpreter and Translator Trainer, 2* (2), 221-252.

Orf, D. (2005). Estudio de situación del mercado español de servicios profesionales de traducción (ACT), Madrid, Agrupación de Centros Especializados en Traducción

Roberts, R.P. (1990). Translation and the Bilingual Dictionary. *Meta, 35* (1), 74–81.

Roberts, R. P. (1992). Traslation pedagogy: strategies for improving dictionary use. *Traduction, Terminologie et Rédaction, 5* (1), 49-76.

Sánchez Ramos, M. M. & Vigier Moreno, F. (2016). Using Monolingual Virtual Corpora in Public

Service Legal Translator Training. In E. Martín-Monje, I. Elorza & B. García Riaza (Eds.), *Technological Advances in Specialized Linguistic Domains: Practical Applications and mobility* (pp. 228-239). London & New York: Routledg.

Seghiri, M. (2004). *Compilación de un corpus especializado ad hoc para la traducción de contratos de seguros turísticos.* Memoria de Investigación. Málaga: Universidad de Málaga.

Seghiri, M. (2006). *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad*. Tesis doctoral. Málaga: SPICUM.

Seghiri, M. (2011). Metodología protocolizada de compilación de un corpus de seguros de viajes: aspectos de diseño y representatividad. *Revista de lingüística teórica y aplicada*, *49* (2), 13-30.

Seghiri, M. (2012). Creating Electronic Corpora: Design, Compilation Protocol and representativeness. In E. Parra-Membrives, M.A. García Peinado & A. Classen (Eds.), *Aspects of Literary Translation: Building Linguistic and Cultural Bridge* in Past and Present (pp. 373-382). Tübingen: Narr Verlang.

Seghiri, M. (2014). Too big or not too big: Establishing the minimum size for a legal ad hoc corpus. *Hermes-Journal of languaje and Communication in Business, 27* (53), 85-98.

Seghiri, M. (2015). Determinación de la representatividad cuantitativa de un corpus ad hoc bilingüe (inglés-español) de manuales de instrucciones generales de lectores electrónicos/Establishing the quantitative representativiness of an E-Reader User's Guide *ad hoc* corpus (English-Spanish). In M. T. Sánchez Nieto (Eds.), *Corpus-based Translation and Interpreting Studies: From description to application* (pp.125- 146). Berlín: Frank & Timme.

Seghiri, M. (2017a). Metodología de elaboración de un glosario bilingüe y bidireccional (inglés-español/español-inglés) basado en corpus para la traducción de manuales de instrucciones de televisores. *Babel, 63* (1), 43-64. Seghiri, M. (2017b). Corpus e interpretación biosanitaria: extracción terminológica basada en bitextos del campo de la Neurología para la fase documental del intérprete. *Panacea, 18* (46), 123–132.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Arce Romeral, L. and Seghiri, M. (2018). Booth-friendly term extraction methodology based on parallel corpora for training medical interpreters. *Current Trends in Translation Teaching and Learning E, 5, 1 – 46.*

Tomblin, J.B., Records, N.L., Buckwalter, P., Zhang, X., Smith E. & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research, 40*, 1245-1260.

Zanettin, F., Bernardini, S. & Dominic, S. (2003).*Corpora in Translator Education*. Manchester: St. Jerome.