Miriam Seghiri

Creating electronic corpora: design, compilation protocol and representativeness

1 Introduction

Many authors have underlined the advantages of using electronic corpora[1] in translation like Laviosa, Bowker and Pearson, or Zanettin, Bernardini, and Stewart, *inter alia*.[2] Actually, researchers and teachers are in agreement over the importance of using electronic corpora in translation training and practice because they offer information about the text structure, terminology, collocations and phraseology used in a determined field of specialization that cannot be found in dictionaries. Corpora are user-friendly and allow management of a huge quantity of information in almost no time. In addition, once a corpus has been created, it can be always reused for different purposes (translating new texts on the same topic, revising translations, studying a concrete genre in different languages, etc…).[3] And all this at minimal cost because electronic corpora are based exclusively on texts mined from the Internet. However, the main problem that translators come up against is that a corpus for the particular speciality is not available for consultation on the web, so translators have no alternative other than to

---

1 Electronic corpora may are also called *ad hoc* (see Gloria Corpas Pastor: *Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada.* Málaga 2001), *disposable* (cfr. Federico Zanettin, Silvia Bernardini and Dominique Stewart (eds): *Corpora in translator education*. Manchester 2003.), *do-it-yourself/DIY* (cfr. Federico Zanettin: *DIY Corpora: The WWW and the Translator.* Porto 2002), *virtual* (see Miriam Seghiri: *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad [Compilation of a trilingual corpus of travel insurance contracts (English-Italian-Spanish): evaluation, classification, design and representativeness].* Málaga 2006), or *special purpose* (cfr. Jennifer Pearson: *Terms in Context, Studies in Corpus Linguistics.* Amsterdam/Philadelphia 1998), among other denominations.

2 See Sara Laviosa (ed): *L'approche basée sur le corpus / The Corpus-based Approach.* Montreal 1998; Lynne Bowker and Jennifer Pearson: *Working with Specialised Language: A practical guide to using corpora.* London 2002; Federico Zanettin, Silvia Bernardini and Dominique Stewart (eds): *Corpora in translator education*. Manchester 2003.

3 The research reported in this paper has been carried out in the framework of the R&D project "Ecoturismo: espacio único de sistemas de información ontológica y tesauros sobre el medio ambiente" (ref. FFI2008-06080-C03-03,2008-2011).

compile their own virtual corpora.[4] In order to do so, a set of clear *design criteria* and a *systematic compilation protocol* have to be established,[5] if the translator wants the collection of text compiled to be considered a *corpus* in the strict sense of the term.

2 Methodology for Corpus Design and Compilation

According to the words expressed above, in this section we will present a proposal of design criteria for the creation of an electronic corpus of travel insurance policies followed by a compilation protocol divided into four steps.

So, firstly, it is vital to establish a set of *clear design criteria* when compiling a corpus. We will illustrate this methodology by creating a corpus of travel insurance policies.[6] This corpus will be *monolingual* (English), and diatopically restricted to the United Kingdom, due to the large number of countries in which this language is spoken. It will be a *full-text* corpus because it will include complete policies, all of them downloaded from the web, so the corpus will be *electronic*. Finally, as the corpus will only include policies, it will be *homogenous* (in genre).

Once the set of design criteria is clear, a *compilation protocol* divided into four steps —(i) finding data, (ii) downloading, (iii) formatting and (iv) storage— should be followed for the creation of the corpus:

The first step, *finding data*, will consist in searching relevant documents on the web. There are two main types of searches that may be carried out online: institutional searches and thematic searches.[7] On the one hand, the

---

4 Cfr. Miriam Seghiri: *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad [Compilation of a trilingual corpus of travel insurance contracts (English-Italian-Spanish): evaluation, classification, design and representativeness].* Málaga 2006. This publication is available online at http://riuma.uma.es/xmlui/bitstream/handle/10630/2715/16754888.pdf?sequence=1. Last accessed on December 05th, 2011.

5 Cfr. EAGLES: *Corpus Typology: A framework for classification. EAGLES Document 080294*. Oxford 1994; EAGLES: *Text corpora Working Group reading Guide. EAGLES Document EAG-TCWG-FR-2*. Oxford 1996a; *EAGLES: Preliminary Recommendations on Corpus Typology. EAGLES Document EAG-TCWG-CTYP/P*. Oxford 1996b.

6 European consumers have the right to demand translations of this type of documents under the auspices of European directives on insurance matters (92/49/EEC and 92/96/EEC). These directives recognise the right of the party taking out insurance to receive the contract written not only in the official language of the member state where the agreement is made, but also in a language which they may specify.

7 For further information about institutional searches and thematic searches see Miriam Seghiri: *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad [Compilation of a trilingual corpus of travel insurance contracts (English-Italian-Spanish): evaluation, classification, design and representativeness].* Málaga 2006. This publication is available online at

institutional search is the one carried out on the web sites of international companies, organisations and institutions. The information one can find on these sites is of a high standard of quality and reliability because the writers are specialists in the field. Travel insurance policies have been mainly downloaded from web sites of British insurance companies such as *AT Bell Insurance Brokers Ltd*,[8] *Lloyds of London*[9] or *Royal and Sun Alliance*,[10] to mention only a few of the most representative examples. On the other hand, thematic search is normally carried out by using key word searches on good search engines. There are many search engines on the Internet, like Google or Yahoo, Live Search, among others. However, according to a great number of analysts Google is the best search engine in terms of the quality of search results.[11] On this point, it is clearly essential to establish descriptors and using Boolean operators, truncation and phrase searches, as illustrated in Table 1, in order to avoid a large amount of irrelevant information to be returned. At the same time, search engines (like Google) allow to restrict the finding to a specific domain. In this case, it will be selected "pages from the UK" (.uk) in order to filter pages from other English spoken countries.

Language	**Domain**	**Genre**	**Descriptors**	**Search equation**
English	.uk	Policy	Policy Travel Insurance	Policy AND "Travel Insurance"

Table 1: Descriptors for the finding of travel insurance policies (English)

Once the English policies have been found, the second step is *downloading data*. This stage can be carried out manually although, sometimes, it is possible to automate the task with programmes like *GNU Wget*,[12] for instance, which allows downloading in groups of web pages or batches.

During the third step, *formatting*, the wide variety of formats available on the web needs to be considered: there is a noticeable predilection for HTML (.html) and PDF (.pdf) formats on the Internet, but all these documents have to be converted to an ASCII or plain text format (.txt) in order to be processed by any corpus management tool like *WordSmith Tools*[13] or

http://riuma.uma.es/xmlui/bitstream/handle/10630/2715/16754888.pdf?sequence=1. Last accessed on December 03th, 2011.

8 <http://www.atbell.co.uk>.

9 <http://www.lloyds.com>.

10 <http://www.royalsunalliance.com/royalsun>.

11 Dragomir Radev, Weiguo Fan, Hong Qi, Harris Wu and Amardeep Grewal. *Probabilistic question answering on the web*. New Orleans 2005.

12 <http://www.gnu.org/software/wget/>.

13 <http://www.lexically.net/wordsmith/>.

Concordance,[14] to name just a few, in accordance with the *clean-text policy* described by Sinclair:

> "The safest policy is to keep to the text as it is, unprocessed and clean of any other codes". [15]

The conversion from any format to plain text is as easy as to copy the information and paste it into a plain text document (.txt). For PDF format, *Google* allows the majority of PDF documents to be seen in HTML, thereby permitting the same procedure —copy and paste— to be carried out. When this is not possible, conversion programmes such as *Solid Converter*[16] or *Convert Doc*,[17] for instance, can be used.

The last stage is the *storage* of the data, and it consists of saving the documents that have been downloaded, correctly identifying and arranging them. One possible way of doing this is through the use of files and subfiles depending on the genre (policy), format (original format and plain text) and language (see Fig. 1).

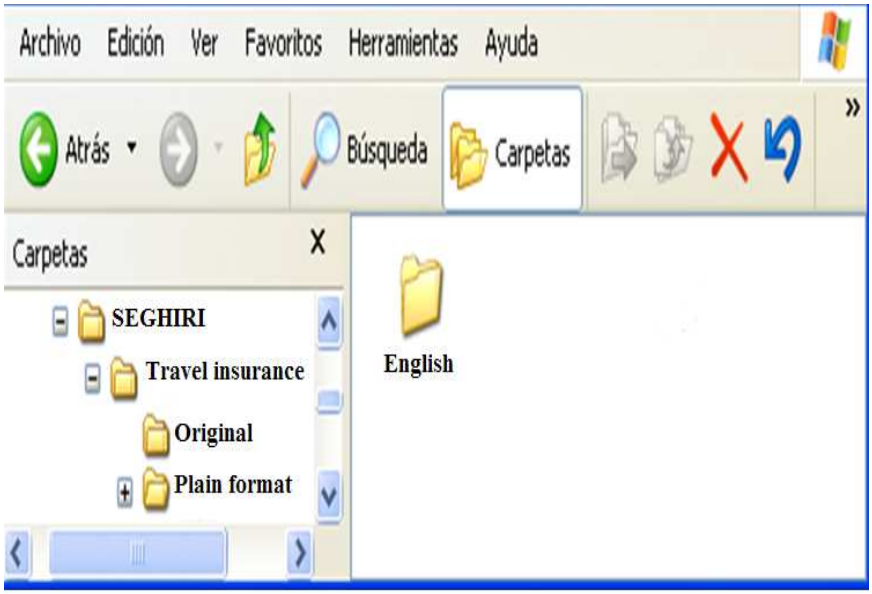

Figure 1: Storage data

[14] <http://www.concordancesoftware.co.uk>.
[15] John Sinclair: *Corpus, Concordance, Collocation*. Oxford 1991.
[16] <http://www.solidpdf.com>.
[17] <http://www.abcdatos.com/programas/programa/z5637.html>.

In the study now under examination an electronic corpus of travel insurance policies in English was compiled, with 176 documents and 1,903,661 words.

3 Establishing corpus representativeness

As Biber et al. said

> "a corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language".[18]

However, the concept of *representativeness* is still surprisingly imprecise considering its acceptance as a central characteristic that distinguishes a corpus from any other kind of collection. In practise, there is no general agreement as to what the size of a corpus should ideally be, and the final size of a corpus tends to reflect the ease or difficulty of acquiring the material.[19]

Nowadays, however, a computer programme, named *ReCor*,[20] enables accurate evaluation of corpus representativeness, because this programme can be used to determine *a posteriori*, for the first time, whether the size reached by a given corpus is sufficiently representative of this particular sector of the tourist industry.

3.1 The ReCor 2.3 interface

The interface of the programme —*ReCor 2.3*— is simple, intuitive and user-friendly (see Figure 2).[21] Firstly, the language may be chosen between

---

[18] See Douglas Biber, Susan Conrad and Randi Reppen: *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge 1998.

[19] Cfr. Voula Giouli and Stelios Piperidis: *Corpora and HLT. Current trends in corpus processing and annotation.* Bulgaria 2002 and CORIS/CODIS: *Progettazione e costruzione di un Corpus di Italiano Scritto.* Bologna 2002. http://corpus.cilta.unibo.it:8080/coris_itaProgett.html. Last accessed on December 04th, 2011.

[20] The programme ReCor and its algoritm (N-Cor), created by Corpas and Seghiri, have been awarded the 2007 Translation Technologies Research Award (Premio de Investigación en Tecnologías de la Traducción) by the Translation Technologies Watch (Observatorio de Tecnologías de la Traducción). Universidad Europea de Madrid. The name of the programme, *ReCor,* is an acronym derived from the function it was designed for: the representativeness of corpora.

[21] The technology and the theoretical presuppositions behind the ReCor programme are explained in detail in Miriam Seghiri: *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad [Compilation of a trilingual corpus of travel insurance contracts (English-Italian-Spanish): evaluation, classification, design and representativeness].* Málaga 2006; Corpas Pastor Gloria and Miriam Seghiri: "Specialized Corpora for Translators: A

English or Spanish. In the second section, *Reports*, an input file –'CORPUS files selection'– may be selected, i.e., the corpus or subcorpus we want to analyse. 'Group of words' allows the user to work with groups of up to ten words or *grams*. In the third section, *Filters*, 'Number Filters' allows numbers to be filtered out. There is also an option, '*Input File (Words Filter)*,' which filters out all those words that the user wants to exclude from the analysis, like proper names or HTML tags, for instance, in the case that the corpus has not been 'cleaned'.

The programme creates three output files: (1) 'Statistical Analysis,' collates the results from two distinct analyses, firstly, with the files ordered alphabetically by name and, secondly, with the files in random order. The document that appears is structured into five columns that show the number of types, the number of tokens, the ratio between the number of different words and the total number of words (types/tokens), the number of words that appear only once (V1) and the number of words that appear only twice (V2); (2) 'Alphabetical Order,' generates two columns: the first shows the words in alphabetical order with their corresponding number of occurrences appearing in the second column; (3) 'Frequency' shows the same information but this time the words are ordered according to their frequency in the corpus.

Quantitative Method to Determine Representativeness". *Translation Journal* 11 (3). 2007b. http://translationjournal.net/journal/41corpus.htm. Last accessed on December 05th, 2011; Corpas Pastor Gloria and Miriam Seghiri: "Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor". *Procesamiento del Lenguaje Natural* 39. 165-172. Seville 2007b. http://www.sepln.org/revistaSEPLN/ revista/39/20.pdf. Last accessed on December 05th, 2011; Corpas Pastor Gloria and Miriam Seghiri: "Fuentes de información electrónicas para la compilación de un corpus virtual bilingüe de seguros turísticos". In *La traducción del futuro: Mediación lingüística y cultural en el siglo XXI (III Congreso de AIETI).* Vigo 2007c. http://www.uma.es/hum892. Last accessed on December 05th, 2011; Corpas Pastor Gloria and Miriam Seghiri: *El concepto de representatividad en lingüística de corpus: Aproximaciones teóricas y consecuencias para la traducción*. Málaga 2011/forthcoming.

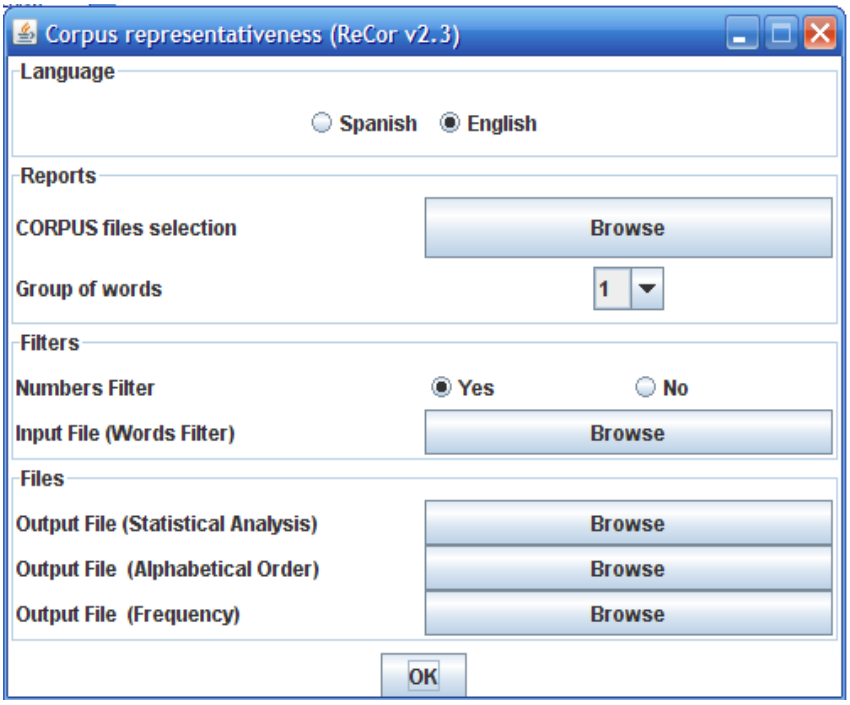

Figure 2: The *ReCor 2.3* interface (English version).

3.2 Corpus analysis

The programme illustrates the level of representativeness of a corpus in a simple graph form, which shows lines that grow at first and then stabilise as they approach zero.[22] In *Estudio gráfico A* (on the left) the number of files selected is shown on the horizontal axis, while the vertical axis shows the type/token ratio. The results of two different operations are shown, one with the files ordered alphabetically (the red line), and the other with the files introduced at random (the blue line). In this way the programme double-checks to verify that the order in which the texts are introduced does not have repercussions for the representativeness of the corpus. At the point where both the red and blue lines stabilise, it is possible to state that the corpus is representative, and at precisely this point it is possible to see approximately how many texts will produce this result. This graph can be used to determine the total number of *documents* that should be set for the

[22] It should be noted here that 0 (=zero) is unachievable because of the existence in the text of variables that are impossible to control such as addresses or proper names to name only some of the more frequently encountered.

minimum size of the collection. At the same time, *Estudio gráfico B* (on the right), can be used to determine the total number of *words* needed for the minimum size of the corpus. Figure 3 shows the representativeness of the English corpus.

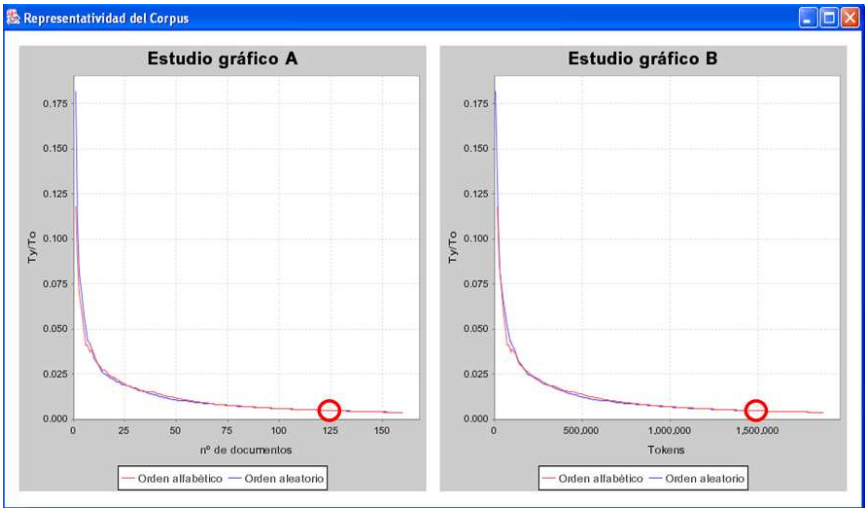

Figure 3. Representativeness of the English corpus of travel insurance policies (1- gram)

The results generated by *ReCor* 2.3 let us to conclude that the English corpus (cfr. Fig. 3) can be considered representative from 125 texts and 1,5 million words.

4 Conclusion

The methodology behind corpus compilation is not always clear and often the availability of documents on the Internet is the crucial criterion which determines the size of the collection of texts.[23] For this reason, we have presented a systematic methodology for corpus compilation divided into four steps: finding data, downloading, normalization and storage. At first, a set of clear design criteria needs to be established. Once the corpus has been designed, compiled and finally created, corpus representativeness needs to be measured *a posteriori*. It is not possible to establish the minimum number of documents for a given corpus *a priori* because the size will depend on the language and genre involved. For this reason, we have used *ReCor* 2.3, a

[23] Cfr. Voula Giouli and Stelios Piperidis: *Corpora and HLT. Current trends in corpus processing and annotation*. Bulgaria 2002 and CORIS/CODIS: *Progettazione e costruzione di un Corpus di Italiano Scritto*. Bologna 2002.

computer programme that calculates, once the corpus has been compiled or during the compilation process, the minimum number of documents and words that should be included in specialised language corpora in order that they may be considered representative. Representative electronic comparable corpora, created in accordance with the protocol outlined in this study, are extremely useful for the study of any field of specialisation. The corpus compiled is ready to carry out different studies from a monolingual and monocultural perspective as well as from the point of view of translation.