Specialized Corpora for Translators:
A Quantitative Method to Determine Representativeness

by Gloria Corpas Pastor, Ph.D. and Miriam Seghiri, Ph.D.
University of Málaga (Spain)

1. Introduction

Nowadays, there can be no doubt as to the importance or the necessity of using corpora in translation. Equally, given the short deadlines and speed that are now demanded in the translation industry, the virtual corpus has undeniably proved itself a most useful tool. Many authors have explored the possibilities offered by corpora for specialized language teaching and translation (cf. Bernardini and Zanettin, 2000; Corpas, 2001 and 2004, Bowker and Pearson, 2002, to name but a few). Ad-hoc, specialized corpora mined from electronic resources available on the Internet have proved to be a first-class documentary resource, as well as a valuable tool in decision-making and in revision. However, there is a surprising scarcity of studies devoted to analyzing the quality of the corpora that are being used in translation.

There are countless projects of studies based on corpora which rely on the quality and representativeness of each corpus as their foundation for producing valid results. As Biber has pointed out, "the representativeness of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalizability of the results of the research" (Biber et al. 1988: 246). However, despite agreement as to their importance (cf. Biber 1988, 1990, 1993, 1994 and 1995; Atkins, Clear and Ostler 1992; Quirk 1992 or EAGLES 1994, 1996), these concepts continue to be very vague and seemingly no consensus exists:

"Several corpus linguists have raised issues concerning the size and representativeness of specialised corpora as well as the generalizability of their findings. In fact, these are thorny issues which have also been widely debated in the literature on corpus studies in general, and to which there seem to be no easy answers." (Flowerdale, 2004: 18)

So, in this paper we will describe a method to assess the quality of a corpus in terms of representativeness. By using the N-Cor algorithm it is possible to quantify a posteriori, for the first time, the minimum number of documents and words that should be included in a specialized language corpus, in order that it may be considered representative. A computer application has been implemented that automatically determines the representativeness threshold for any given corpus. In the present paper this software will be used with a sample corpus of general conditions in vacation package contracts (English-Spanish) mined from the Internet.

2. Corpus minimum size

The size of the corpus is a decisive factor in determining whether the sample is representative in relation to the needs of the research project.
addition, the general conditions constitute an excellent text type, since concrete communication situation, the sale of vacation packages. In relatively limited as it will be used by a very specific community in a sector for translations of general conditions of vacation packages both for vacation packages in Spanish and English has been used. The collection.

There have been many attempts to set the size, or at least establish a minimum number of texts, from which a specialized corpus may be compiled. Some of the most important are those put forward by Heaps (1978), Young-Mi (1995) and Sánchez Pérez and Cantos Gómez (1997). However, subsequently some of these authors such as Cantos (Yang et al. 2000: 21) recognized some shortcomings in these works, stating that it might be attributed to their preference for Zipf's law. Zipf's law can give us an idea of the breadth of vocabulary used, but it is not limited to a particular or approximate number because this will depend on how the constant is determined (Braun 2005 [1996] and Carrasco Jiménez 2003: 3). Numerous studies have been based on that law, but the conclusions they reach do not specify, even through the use of graphs, the number of texts that are necessary to compile a corpus for a particular specialized field.

A possible solution could be to analyze the lexical density of a corpus in relation to the increase in documentary material included (Corpas Pastor and Seghiri Dominguez, 2006, and Seghiri Dominguez, 2006). In other words, if the ratio between the actual number of different words in a text and the total number of words (types/tokens) is an indicator of lexical density or richness, it may be possible to create a formula that can represent increases in the corpus (C) on a document by document (d) basis: the number of types does not increase in proportion to the number of words the corpus contains, once a certain number of texts has been achieved.

This may make it possible to determine the minimum size of a corpus and the quantity that must be reached for it to begin to be representative. With the help of graphs, it should be possible to establish whether the corpus is representative and approximately how many documents are necessary to achieve this. This theory has become a practical reality in the shape of a software application (ReCor) which enables accurate evaluation of corpus representativeness. Once the question of quality is ensured in terms of corpus design and document selection, this program can be used to determine a posteriori whether the size reached by a given corpus is sufficiently representative of this particular sector of the tourist industry.

For illustrative purposes, a sample corpus composed of general conditions for vacation packages in Spanish and English has been used. The importance of this text type, dealing with vacation packages, is clear because, alongside contracts for time-shares, it is the only type of tourism contract that is covered by substantive communitary legislation. Also, since the Spanish tourist industry is one of the main driving forces behind the Spanish economy, there is a large demand in the tourism sector for translations of general conditions of vacation packages both from Spanish into English and from English into Spanish (cf. ACT, 2005). This the component of general conditions for vacation packages will be relatively limited as it will be used by a very specific community in a concrete communication situation, the sale of vacation packages. In addition, the general conditions constitute an excellent text type, since

3. The software

In order to quantify corpus representatives, a software program has been implemented. ReCor’s interface is simple, intuitive, and user-friendly (cf. Fig. 1). First, an input file may be selected; this could be anything from a particular clause in a policy to the entire corpus. 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 addresses, proper names or even HTML tags, in the case where the corpus has not been cleaned.” Next, three output files are created. The first, “Statistical Analysis,” collates the results from two distinct analyses; first, with the files ordered alphabetically by name and then with the files in random order. The document that appears is structured into five columns which 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). The second output file, “Alphabetical Order,” generates two columns; the first shows the words in alphabetical order with their corresponding number of occurrences appearing in the second column. The same information is shown in the third file, “Frequency,” but this time the words are ordered according to their frequency or rank. The application also allows the user to work with groups of up to ten words (n-grams) and phraseology, as well as allowing numbers to be filtered out.

Figure 1: The ReCor interface.

3.1. Graphical representation of data

The program illustrates the level of representativeness of a corpus in a simple graph form, which shows lines that grow exponentially at first and then stabilize as they approach zero. It should be noted here that zero (0) is unachievable because of the existence in the text of variables that are impossible to control such as addresses, proper names or numbers, to name only some of those more frequently encountered.

In the first presentation of the corpus in graph form that the programme generates—Graphical Representation A—the number of files selected is shown on the horizontal axis, while the vertical axis shows the types/tokens 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 program
double-checks to verify that the order in which the texts are introduced does not have repercussions on the representativeness of the corpus. Both operations show an exponential decrease as the number of texts selected increases. However, at the point where both the red and blue lines stabilize, 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.

At the same time another graph—Graphical Representation B—is generated in which the number of tokens is shown on the horizontal axis. This graph can be used to determine the total number of words that should be set for the minimum size of the collection.

Once these steps have been taken, it is possible to check whether the number of general conditions of a travel package that have been compiled in the two languages involved—English and Spanish—is sufficient to enable us to affirm that our sample corpus is representative. See Figures 2 and 3 below which show the representativeness of the two languages involved.

Figure 2: Representativeness of the Spanish subcorpus (1-gram).

From the data shown in Figure 2 it is possible to deduce that, according to Graph A, the component of general conditions in Spanish begins to be representative from the point of the inclusion of 200 documents; since the curve hardly varies either before or after this number, in other words this is the point where the lines stabilize and are closest to zero. As mentioned above, in practice zero is unattainable because, despite having chosen ReCor's option to filter out numbers as well as using the word filter, all documents always contain a number of variables which are impossible to control (for example, proper names or addresses, to mention only some of the more frequent examples). Graph B shows the minimum total number of words (tokens) necessary for the corpus to be considered representative, which in this case is 750,000 words.
In the case of Figure 3, from Graph A it is possible to assert that the English subcorpus becomes representative from the point where 175 documents are included. In addition, according to the data generated by ReCor shown in Graph B, the figure for the total number of words necessary in order to claim representativeness is around 600,000 words.

A comparison of the two sets of graphs in Figures 2 and 3 shows that despite the fact that a similar number of general conditions have been found on the Internet for both languages—279 texts in Spanish and 240 in English—the English documents reach the point of representativeness long before the Spanish documents: 175 documents and 600,000 words in English against 200 documents and 750,000 words in Spanish.

The results remain largely the same even when the analysis is performed on a two-word basis (2-grams): 225 documents and 750,000 words in English (cf. Figure 5) as against 250 documents and 800,000 words in Spanish (cf. Figure 4).

Figure 4: Representativeness of the Spanish subcorpus (2-grams).

Figure 5: Representativeness of the English subcorpus (2-grams).

From this it may therefore be deduced that, despite the fact that the legal systems involved in the study all have substantive legislation on the subject of vacation packages, the English general conditions tend to be more homogeneous than those in Spanish. In other words, it is possible to infer that the general conditions in English present super-, macro- and microstructures that are very similar to each other and use a narrower terminological range.

Despite these quantitative differences, however, it is not possible to determine \textit{a priori} the exact total number of words or documents that should be included in specialized language corpora (which in general tend to be smaller) in order that they may be considered representative. This
is because, as has been illustrated, size will be determined according to the language and text types, as well as the restrictions of a particular specialized field or diatopic limitations.

6. Conclusion

Now, for the first time, corpus representativeness can be measured a posteriori by means of the N-Cor algorithm. ReCor is a computer application based on the N-Cor algorithm that calculates the minimum number of documents and words that should be included in specialized language corpora, in order that they may be considered representative. It should be pointed out that it is not possible to establish the minimum number of documents for a given corpus a priori, as the size will depend on the language and genres involved, as well as on the restrictions of a particular specialized field and any other diasytematic limitations. This new quantitative method will make exciting future research for collocational and phraseological studies on corpus representativeness possible.

References


1 The research reported in this paper has been carried out in the framework of R&D Project for Excelence La contratación turística electrónica multilingüe como mediación intercultural: aspectos legales, traductológicos y terminológicos [Multi-lingual tourism e-contracts: legal, translational and terminological aspects]. Funding source: Andalusian Ministry of Education, Science and Technology. Ref. no. HUM-892 (2006-2009).

2 The methodology we describe in this paper has 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). Further information at the URL: http://www.uem.es/web/ott. The ReCor program (version 3.0) will be soon available at: http://www.recorweb.com.

3 A systematic methodology for corpus compilation based on electronic resources available on the Internet is described in Corpas (2002) and Seghiri (2006).

4 There are a surprising number of research projects that, whilst endeavoring to compile a "representative" corpus, hardly seem to touch on this concept. Usually, it is noticeable that the availability of material in the particular field of study determines the final size of the corpus (Giouli y Piperidis, 2002).

5 ReCor is an acronym derived from the function it was designed for: the representation of corpora.

6 Tourism is responsible for a huge volume of business in the international economy with Europe occupying a privileged position at the top of the world scale. In 2006 Europe generated $6,466.2 billion in this sector, equivalent to 10.3% of the world’s gross domestic product (GDP), forecast to rise to 11% by 2011, accounting for 8.7% of total employment (WTTC, 2006a). Also see studies by the WTTC concerning the United Kingdom (2006b), Ireland (2006c) and Spain (2006d) for a more detailed analysis of the figures for these countries in this sector.