Size matters: A quantitative approach to corpus representativeness

Gloria Corpas Pastor & Miriam Seghiri
Universidad de Málaga
gcorpas@uma.es, seghiri@uma.es

1. Introduction

> We should always bear in mind that the assumption of representativeness 'must be regarded largely as an act of faith' (Leech 1991: 2), as at present we have no means of ensuring it, or even evaluating it objectively. (Tognini-Bonelli 2001: 57)

Corpus Linguistics (CL) has not yet come of age. It does not make any difference whether we consider it a full-fledged linguistic discipline (Tognini-Bonelli 2000: 1) or, else, a set of analytical techniques that can be applied to any discipline (McEnery et al. 2006: 7). The truth is that CL is still striving to solve thorny, central issues such as optimum size, balance and representativeness of corpora (of the language as a whole or of some subset of the language).

Corpus-driven/based studies rely on the quality and representativeness of each corpus as their true foundation for producing valid results. This entails deciding on valid external and internal criteria for corpus design and compilation. A basic tenet is that corpus representativeness determines the kinds of research questions that can be addressed and the generalizability of the results obtained (cf. Biber et al. 1988: 246). Unfortunately, faith and beliefs do not seem to ensure quality.

In this paper we will attempt to deal with these key questions. Firstly, we will give a brief description of the R&D projects which

originally have served as the main framework for this research.[1] Secondly, we will focus on the complex notion of corpus representativeness and ideal size, from both a theoretical and an applied perspective. Finally, we will describe a computer application which has been developed as part of the research. This software will be used to verify whether a sample bilingual comparable corpus could be deemed representative.

2. The TuriCor corpus

The TURICOR project (ref. no. BFF2003-04616, 2003-2006) and its follow-up (ref. no. HUM-892, 2006-2009) involve research in the field of translation technology.[2] As a consequence, it has an interdisciplinary character and is geared towards applied research. The ground hypothesis is that the use of corpora and, in particular, virtual corpora compiled from material downloaded from the Internet constitutes a great advance for research in the fields of language engineering and applied linguistics.[3] Furthermore, the benefits and advantages brought about by this advance may also have implications for companies involved in the tourism industry. One of the main aims of the projects is to compile a large multilingual travel and tourism law corpus (in English, French, German, Italian and Spanish) by mining electronic resources available on the Web.

The *Turicor* corpus is a virtual, multilingual macrocorpus of travel and tourism law, made up of different subcorpora, which are

---

[1] The research reported in this paper has been carried out in the framework of R&D projects BFF2003-04616 (Spanish Ministry of Science and Technology/EU ERDF. 2003-2006) and HUM-892 (Andalusian Ministry of Education, Science and Technology. 2006-2009). Gloria Corpas Pastor leads both projects.

[2] For more information on the two projects see the official websites at <http://www.turicor.com> and <http://www.uma.es/hum892>.

[3] In this context an *ad hoc* corpus refers to a virtual, high quality corpus that has been specifically compiled in order to carry out a particular translation project or to document a given translation problem within a project.

parallel and comparable[4], covering five languages – English, French, German, Italian and Spanish. It has been built up using hypertext structures, principally from monolingual and multilingual web pages, as well as other resources on the Internet. This corpus has been created with a view to (a) implementing a multilingual NLG system, (b) compiling specialised multilingual termbanks, and (c) using it as a teaching tool for translation, contrastive rhetoric and comparative law.

The corpus has been diatopically restricted to Germany, Spain, Italy, France, Republic of Ireland and the United Kingdom, by which we mean Wales, England, Scotland, Northern Ireland, the Isle of Man, the Isle of Wight and the Channel Islands. As to the selection of documents for inclusion, the following classification of tourism contracts have been developed: package (cruise), timeshare (maintenance, exchange and ownership transfer or resale), transport (air —charter and low cost, water & sea and road & train), travel insurance, accommodation (hotel and extrahotel), combined services; parking, car rental (with driver and without driver), hotel franchise (management and contingent hiring) and restauration (banqueting and catering).

The *Turicor* corpus currently holds 5,022 documents in German, Spanish, English, French (at an initial stage) and Italian, giving a total of more than twelve million words or to be more precise, 17,139,591 tokens of which half (8,569,791) come from tourism contracts, general conditions and standard forms.[5]

In this paper we will check a collection of general conditions in package holiday contracts[6] in Peninsular Spanish and British and American English as regards representativeness by means of a computer programme which has been developed within the framework of both projects.[7] The Spanish and the British subcorpora have been ex-

[4] The multilingual comparable subcorpus is far larger than the parallel component.

[5] Figures given as of 19 June 2007.

[6] The importance of this text type, dealing with package holidays, is clear because, alongside contracts for time-shares, it is the only type of tourism contract that is covered by substantive legislation. For a wider discussion of this matter, see Almahano Güeto (2002)

[7] The methodology we describe in this paper has been awarded the 2007 Translation Technologies Research Award (Premio de Investigación en Tecnologías de la Tra-

tracted from the *Turicor* corpus, whereas the American subcorpus has been mined from the Internet.[8] This analysis of representativeness is a pilot study that could be applied to any given corpus, or to any component or subcorpus derived from it.

3. The importance of being representative

Countless definitions have been forwarded as to what constitutes a corpus. Examples of the more commonly known definitions follow: '[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' (Francis 1982: 17); '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' (EAGLES 1996*a* y *b*); 'a finite-sized body of machine-readable texts sampled in order to be maximally representative of the language variety under consideration' (McEnery and Wilson 2001: 24).

Despite repeated reference to the quality of being 'representative', constituting a 'sample' and so forth as distinguishing features of corpora as opposed to other kinds of textual collections, there appears to be no consensus amongst the experts on the two crucial questions of quality and quantity.

> The definition of representativeness is a crucial point in the creation of a corpus, but is one of the most controversial aspects among specialists, especially as regards the ambiguity inherent in its use due to the intermingling of quantitative and qualitative connotations (CORIS/CODIS 2006).

---

ducción) by the Translation Technologies Watch (Observatorio de Tecnologías de la Traducción). Further information at the URL: <http://www.uem.es/web/ott>.

[8] The American English subcorpus is described in Corpas Pastor (2006). It has been compiled according to the methodology for virtual corpus compilation described in Seghiri (2006).

Dealing with the first concept, that of quality, the root of the problem here may lie in the low quality of the texts that are included if they come from sources that are insufficiently reliable (Gelbukh et al. 2002: 10). This obstacle has been dealt with by designing, as a central element of the project, a system for gauging the quality of digital information through adopting an evaluation protocol, which has been applied to all the documents that were potential candidates for inclusion in the *Turicor* macrocorpus (Seghiri 2006: 89-95).

Another important issue on the subject of quality concerns the coverage of representative genres and text types in a given corpus. In the case of *Turicor*, genres have been carefully chosen to represent the travel and tourism law domain. They include the main types of contract documents currently produced within the tourism industry (agreements, general conditions and forms), as well as the appropriate rules and regulations in force in the corresponding jurisdictions (cf. section 2 above). The *Turicor* corpus has, therefore, been compiled in accordance with carefully selected diasystematic restrictions.

As to the size of the corpus, our starting point was not to establish figures for the number of documents or the total number of words. Instead, a computer programme has been devised to evaluate the appropriateness of corpus size. Once the question of quality is ensured in terms of corpus design and document selection (external criteria), a programme has been develop to determine whether the size reached by a given component or subcorpus of *Turicor* is sufficiently representative of a particular sector of the tourist industry or of the tourism sector in general (internal criteria).

It should be born in mind that 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 (Lavid 2005). However, as will be demonstrated, even today the concept of representativeness is still surprisingly imprecise, especially if one considers its acceptance as a central characteristic that distinguishes a corpus from any other kind of collection.[9] As Biber, who is one of the most prolific writers on the

[9] There are a surprising number of research projects that whilst endeavouring to compile a 'representative' corpus hardly seem to touch on this concept. Usually, it

subject of corpus representativeness, emphasises, 'a corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language' (Biber et al. 1998: 246). Nevertheless, at the same time Biber remains conscious of the difficulties involved in compiling a corpus that could be defined as 'representative' (Biber et al. 1998: 246-247).

3.1. Some theoretical assumptions

Although the criteria for creating a virtual corpus will depend on specific objectives, a doubt will always remain as to whether the number of texts that have been collected and, closely related to this question, whether the number of words contained in them will be sufficient. In other words, the question turns on determining a minimum quantity from which a collection may be deemed to be representative of the field it is attempting to cover. There have been a great number of papers and research projects on the question of quantity as a criterion to gauge representativeness as well as suggested formulas for calculating the minimum number of words and documents necessary for a specialist corpus to be considered representative (Heaps 1978; Biber 1988, 1990, 1993, 1994 and 1995; Leech 1991; Haan 1992; Zampolli et al. 1994; Lauer 1995*a, b* and *c*; Biber et al. 1998 and Yang et al. 1999 and 2002, amongst others).

It is therefore commonplace to come up against questions over the minimum number of texts that will guarantee that the sample taken is scientifically valid as well as debates over how to specify from what quantity it is possible to decide that the number of texts included, and therefore the number of words, is sufficient (Sanahuja and Silva 2001). On the same point McEnery and Wilson (2001: 32) note that:

> A corpus in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a finite-sized body of machine-readable text, sampled in order to be

---

is noticeable that the availability of material on the particular field the study is dealing with determines the final size of the corpus (Giouli y Piperidis 2002).

maximally representative of the language variety under consideration.

Taking these considerations into account, an attempt has been made, using a computer application designed for such a task, to specify the minimum number of documents, and therefore words, which this 'finite-sized' body of texts should include in order to be considered representative. Thereby, the point at which the addition of further documents becomes unnecessary may be defined. Thus, our starting point is that in an area which is as limited as that being considered here, i.e. package holiday contracts, and whilst allowing that new words will always be included in the corpus (it should be remembered that an infinite number of variables such as numbers, addresses and proper names exist), a point is reached when the addition of more documents will not in practice bring anything new to the collection. In other words, it will not make it more representative because all the different situations, as well as the normal range of terminology in this particular field, have already been covered.

Until recently, a similar method for determining representativeness was carried out by applying Zipf's law.[10] Zipf's law is based on the idea that all texts contain a number of words that are repeated. The total number of words in any text is referred to as tokens, while the total number of distinct words, without counting repetitions, is known as types. If this second quantity, types, is divided by the total number of words in the text, or tokens, the frequency that each word appears in the text may be calculated. Words may thereby be ordered according to their frequency with each word being given a rank. The word with the highest frequency will occupy the first position on the list, or rank one, with the other words following in descending order.

Zipf stated that a relationship existed between the frequency of a word and its rank, so that the higher the rank number of a word the lower its frequency of occurrence in a text, since a higher rank number indicates that the word is further down the list and therefore less frequent. In other words, there is an inverse relationship between fre-

[10] For a historical perspective on how Zipf's law was developed see Moreiro González (2002).

quency and rank, i.e. frequency decreases as rank increases. By using Zipf's law, it is therefore, possible to establish that the number of occurrences of a word or its frequency of occurrence – *f(n)* – is inversely proportional to its number on the list or rank (*n*). Zipf's law can be stated mathematically as follows:

$$f(k; s, N) = \frac{1/k^s}{\sum_{n=1}^{N} 1/n^s}$$

From this law it may be deduced that the words with the highest absolute frequency are those that are 'empty', whilst the least frequent are those that reveal the author's individual style and richness of vocabulary. Words that appear in the middle range in terms of frequency distribution are those that are really representative of the document (Velasco et al. 1999: 35).

Zipf's law can, therefore, 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 and Carrasco Jiménez 2003: 3). Numerous studies have been based on the 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 specialised field (Almahano Güeto 2002: 281).

There have been many attempts to set the size, or at least establish a minimum number of texts, from which a specialised corpus may be compiled. Some of the most important are those put forward by Heaps (1978),[11] Young-Mi (1995) and Sánchez Pérez and Cantos Gómez (1997). However, subsequently some of these authors such as

[11] Indeed, out of this work came the rule known as Heaps' law. Both Zipf's and Heaps' laws are used to grasp the variability of corpora. Heaps' law is an empirical law which examines the relationship between vocabulary size, or in other words, the number of different words (types) and the total number of words in a text (tokens). In this way a sequential increase of vocabulary in relation to text type can be observed. The programme *ReCor* has been validated using this law (Seghiri 2006: 399-403).

Cantos (Yang et al. 2000: 21) recognised some shortcomings in these works, stating that 'Heaps, Young-Mi and Sánchez and Cantos failed by using regression techniques.[12] This might be attributed to their preference for Zipf's law'.[13]

Faced with this situation, other authors have suggested the following solution:

> […] as a result of applying Zipf's law quite a few words have a redundant number of occurrences and make up the greatest part of corpus's volume, while the vast majority of the words have a number of occurrences which is statistically insufficient. The solution to this problem is to use the biggest corpus that humanity has ever created – the Internet (Gelbukh 2002: 7)

It is undeniable that the advantages offered by the Internet have opened up an infinite number of possibilities for linguistic or translation research, indeed programmes such as *WebCorp*[14] have already been developed for this task. However, we concur with Sinclair (2004a) that this approach is not valid because, due to the Internet's inherent peculiarities, it cannot, strictly speaking, be considered a corpus. Examples of these peculiarities include its size, which remains unknown, and its mutability, which means it is continually growing

---

[12] Simple linear and multiple linear are the most usual regression techniques used. The prototype situations that these techniques are applied to consist primarily of a set of subjects or observations in which two variables, X and Y for instance, can be measured. When the value of one of the variables, that of X for example, is known the technique is used to predict the value of this subject in the variable Y. A detailed description of different regression techniques and their applications can be found in Lorch and Myers (1990).

[13] Conscious of these deficiencies, Yang et al. (2000) attempted to overcome them by taking a new approach: a mathematical tool capable of predicting the relationship between linguistic elements in a text (types) and the size of the corpus (tokens). However, at the end of their study, the authors reflected on some of its limitations, "the critical problem is, however, how to determine the value of tolerance error for positive predictions" (Yang et al. 2000: 30).

[14] <http://www.webcorp.org.uk>.

and changing along with its users. In addition, the population it represents is another unknown factor.

3.2. Size recommendations

It equally surprising to observe how, for many authors, no maximum or minimum number of texts, or words, that a corpus should contain seems to exist (Sinclair 2004a) and where an approximate figure is proposed, many authors appear to take extreme positions.

Thus, Biber (1993) or McEnery and Wilson (2006), suggest that the ideal number of words that any corpus should reach is around a million.[15] The same figure is given by other researchers such as Borja Albi (2000) and Ruiz Antón (2006). Others, such as Friedbichler and Friedbichler (2000), consider that a figure between '500,000 and 5 million words per language (depending on the target field) will provide sample evidence in 97% of language queries'. There are also those that go further and end up proposing such 'mottos' as 'there is no text like more text', 'more data is better data' or 'the bigger the corpus the better' (Church and Mercer 1993: 18-19 and Wilkinson 2005: 6). Similarly, Sinclair (2004*b*) considers that ideally a corpus should be 'big', although the interpretation of this adjective remains open to debate because no approximate figure is given.

Although it is the dream of many linguists to have gigantic corpora of more than ten million words at their disposal to enable them to carry out studies on general language (Wilkinson 2005: 6), it has been shown that smaller corpora give optimum results in specialised areas. In fact, an increasing number of researchers, such as Bowker and Pearson (2002: 48), stress that shorter corpora with 'a few thousand and a few hundred thousand words' are just as useful in the study of languages for specific purposes.

Likewise, Leech (1991: 10) had already, at the beginning of the nineties, given four reasons why corpora of a larger size were not necessarily better. Firstly, he pointed out, as we have already seen, that a

[15] Ball (1997 [1996]) allows us to comprehend, with the aid of concrete examples, what a million words actually means in relation to the different sources they have been taken from.

vast collection of texts is not sufficient, in and of itself, to be considered a corpus; rather the texts that form the collection must be representative. Secondly, it is possible that an increase in the size of a corpus is simply due to the inclusion of a huge amount of written texts. In addition, corpora of such a great size usually run up against countless copyright problems, hence the bigger the corpus the more numerous the problems. Finally, Leech mentions the scarcity of programmes that are capable of processing the information contained in these giant corpora. As to this last point, we should remember that these observations were made in 1991 and now the situation is completely different with the existence today of powerful programmes such as *SARA*,[16] *MonoConc*,[17] *MultiTrans*[18] or *WordSmith Tools*,[19] to mention only some of the more popular examples.

There is also an increasing number of researchers who support the idea that corpora of a great size are not necessary for work in specific areas. Thus, Clear (1994) wrote an article: 'I Can't See the Sense in a Large Corpus', whose title speaks for itself. Other authors have followed this same line of thought and have emphasised that smaller corpora are extremely useful for sketching out specific areas of a language (Murison-Bowie 1993: 50) or for language teaching (Ghadessy, Henry and Roseberry 2001). Similarly, Fillmore (1992: 35) gives his personal opinion that 'every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way.' Pearson (1998) and Rundell and Stock (1992) have also added two interesting arguments in favour of smaller corpora. The first, forwarded by Pearson (1998: 57), holds that the size of a corpus may be reduced if its design criteria include the intention to represent discourse in a specific scientific community, genre and field; as has been done in the case of *Turicor*. Secondly, Rundell and Stock (1992: 47) state that the final size of a corpus should be in

[16] <http://www.ccl.kuleuven.ac.be/about/ANNO/TOOLS/sara.html>.

[17] *MonoConc* processes comparable texts. A version for dealing with parallel texts also exists, *ParaConc*. For more information see http://www.athel.com/ mono.html>.

[18] For a critique of the *Pro3* version of *MultiTrans*, see Gervais (2003).

[19] For more information and to download *WordSmith Tools*, go to <http://www.lexically.net/wordsmith/>.

proportion to the number of text types that are going to be included, as well as the relative frequency of their appearance in general language. Finally, Baker (2006: 28-29) stresses quality over quantity:

> One consideration when building a specialised corpus in order to investigate the discursive construction of a particular subject is perhaps not so much the size of the corpus, but how often we would expect to find the subject mentioned within it … Therefore, when building a specialised corpus for the purposes of investigating a particular subject or set of subjects, we may want to be more selective in choosing our texts, meaning that the quality or content of the data takes equal or more precedence over issues of quantity.

Taking these points into account, it would seem that the component of general conditions for package holidays now under examination will be relatively limited as it will be used by a very specific community in a concrete communication situation, the sale of package holidays. In addition, the general conditions constitute an excellent text type, since by European Law[20] (cf. *Council Directive of 13 June 1990 on package travel, package holidays and package tours regulations, 90/314/EEC*) they must appear in the brochures that package holiday companies produce for advertising purposes.

Wright and Budin (1997) concur on this point, adding that a corpus of one hundred thousand lexical items is sufficiently large to draw meaningful conclusions because the vocabulary used to deal with a specialised subject is more restricted than that used in non-specialised discourse. Likewise, Ahmad and Rogers (2001: 736) maintain that, 'As a rule of thumb, special-language corpora already start to become useful for key terms of the domain in the tens of thousands of words, rather than the millions of words required for general-language lexicography'.

However, although size is determined in relation to the particular analysis that is intended, even where the interaction under scrutiny is very

[20] There is not American substantive legislation for package travel, package holidays or package tours.

specific, some studies have been carried out using a relatively low number of words and texts due to the fact that they have been compiled, as is usually the case, on the basis of availability of material. 'There is no general agreement as to what the size of a corpus should ideally be. In practice, however, the size of a corpus tends to reflect the ease or difficulty of acquiring the material.' (Giouli and Piperidis 2002).

Whether or not corpora of a reduced size are governed by availability of materials, it must be recognised that they have been, and continue to be, compiled in order to carry out linguistic and translation studies, amongst other kinds of research. Above all, they have proved to be extremely useful tools. In different studies, Haan (1989, 1992) has given a detailed account of the success of a wide variety of analyses based on corpora that contain no more than twenty thousand words. In different linguistic studies carried out using small corpora, Kock (1997 and 2001) also draws the conclusion that these collections (each containing 19 or 20 texts with approximately one hundred thousand occurrences) are more than sufficient, taking into account that 'it is not necessary to have such large corpora if they are homogenous in terms of language register, geographical area and historical time, for instance' (Kock 1997: 292). Biber reduces these figures still further and states that it is possible to represent practically the totality of elements of a specific register with relatively few examples, one thousand words, and a small number of texts belonging to this register, ten to be exact (Biber 1995: 131).[21]

If these principles are applied to the particular case under examination here, it may be stated that the component of general conditions for package holidays has been isolated with the objective of analysing the language used by a very limited community, in a communicative situation that is very specific (the sale of package holidays) and with only one text type being represented (general conditions), whose frequency in general language use is minimal. In addition, Bravo Gozalo and Fernández Nistal (1998: 216) add that size should be in relation to the purpose the corpus is going to be used for. Since the corpus under

[21] To gain a wider perspective on the conclusions reached by this author, see Biber (1988, 1990, 1993, 1994 and 1995).

examination has a very specific objective, its size could be even further reduced, taking this consideration into account.

It is necessary at this point to mention that specialised texts are terminologically far denser than general language texts (Ahmad and Rogers 2001: 726). Documents with a high level of technicality, therefore, display this 'terminological density', or in other words, a high number of units that convey specialised knowledge. Cabré (1999: 89) found that the degree of communicative specialisation conditions the terminological density of a text. Expressive variation when dealing with a concept or situation should also be born in mind. A highly specialised text, therefore, may be characterised as precise, concise and systematic and the terminology it contains tends to be monosemous and univocal. It may, therefore, be concluded that although a specialised corpus does not contain as high a number of words as a general language corpus, it is still possible to obtain satisfactory results as long as the representativeness already described is attained and, additionally, a considerable sample of texts with a high degree of technicality is used.

The fact that no consensus exists as to the number of documents and words that our final collection should include has led us to the conclusion that, before carrying out any kind of analysis, it is essential to ensure that the number of documents and words achieved is sufficient. However, the range of figures that have been suggested differ widely and the proposed calculations are not particularly reliable.[22] In a previous study (Corpas Pastor and Seghiri 2006*a*), we concluded that a possible solution may be to carry out an analysis of lexical density in relation to the increase in documentary material included. 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

[22] On this subject, see the study by Yang et al. (2000: 21) in which reference is made to the shortcomings of studies, which until recently were considered valid, based on Zipf's law.

that can represent increases in the corpus (C) on a document by document (d) basis,[23] for example:

$$C_n = d_1 + d_2 + d_3 + ... + d_n$$

Following from this, our starting point is the idea forwarded by Biber (1993) and subsequently endorsed in studies such as those by Sánchez Pérez and Cantos Gómez (1998) that 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 which enables accurate evaluation of corpus representativeness[24], as described in the next section.

4. The *ReCor* programme

ReCor is a software application which has been designed within the framework of the aforementioned R&D projects so as to facilitate the evaluation of representativeness of corpora in relation to their size. Above all, it is notable for the simplicity of its user interface; this contrasts with the highly mathematical complexities that are typical in this kind of research.

[23] For more information see Corpas Pastor and Seghiri (2007a, 2007b, 2007c and 2008/forthcoming) and Seghiri (2006, 2008a and 2008b).

[24] *ReCor* is an acronym derived from the function it was designed for: (checking) the representativeness of a given corpus.

In this study we used version 2.3 of ReCor. We are currently working on a new version [25] which has an improved capacity for working with multiple and very large files quickly and also allows lexical bundles to be identified on the basis of analysis of n-grams ($n \geq 1$ and $n \leq 10$) of the corpus.

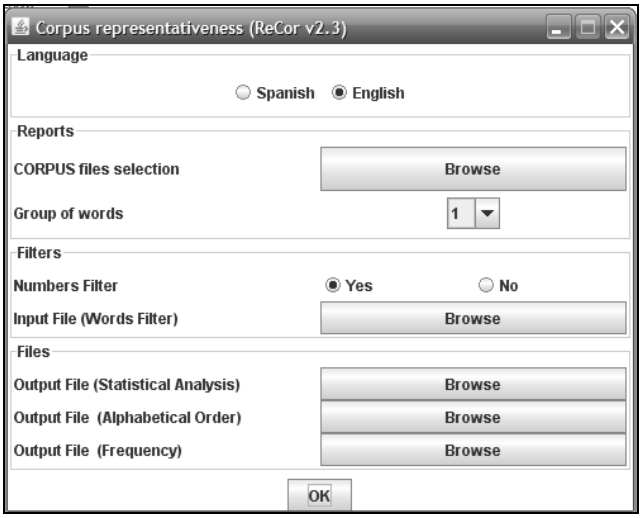
4.1. A userfriendly GUI

ReCor's GUI is simple, intuitive and user-friendly and it is divided into four sections: *Language, Reports, Filters* and *Files*. Firstly, the language may be chosen between English or Spanish. In the second section, *Reports*, an input file –'CORPUS files selection'– may be selected; this could be anything from a particular clause in a policy to the entire *Turicor* corpus. 'Group of words' also allows the user to work with groups of up to ten words (n-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 addresses, proper names or even HTML tags, in the case that the corpus has not been 'cleaned'.

Three output files are created. The first, '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 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 in other words, by their rank.

[25] The new version of ReCor will be soon available on line.

Figure 1: The *ReCor* interface (English version).



4.2. Graphical representation

The programme illustrates the level of representativeness of a corpus in a simple graph form, which shows lines that grow exponentially at first and then stabilise as they approach zero.[26]

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 programme double checks to verify that the order in which the texts are introduced does not have repercussions for the representativeness of the corpus. Both operations show an exponential decrease as the number of texts selected increase. However, at the point where both the red and blue lines stabilise, it is possible to state that the cor-

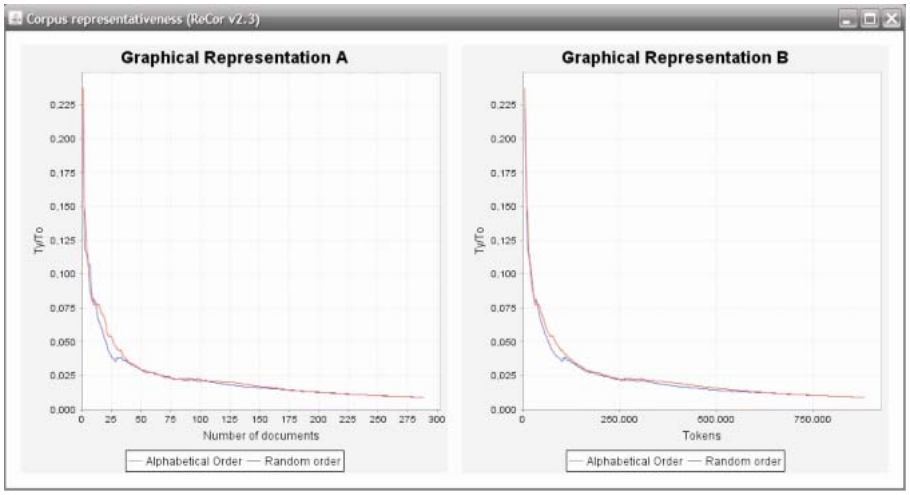
[26] It should be noted here that zero is unachievable because of the existence in the text of variables that are impossible to control such as addresses, proper names or numbers, to name only some of the more frequently encountered.

pus 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[27] of general conditions for package holidays that have been compiled in the two languages involved –English (British and American varieties) and Spanish– is sufficient to enable us to affirm that our component is representative. See Figures 2, 3 and 4 below which show the representativeness of the languages involved.

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



[27] The *Turicor* component for general conditions in Spanish contains 288 documents (820,015 words), whereas the British English component is composed of 234 documents (775,120 words). The American English subcorpus comprises 216 documents and 500,213 words.

Figure 3: Representativeness of the British English component (1-gram).

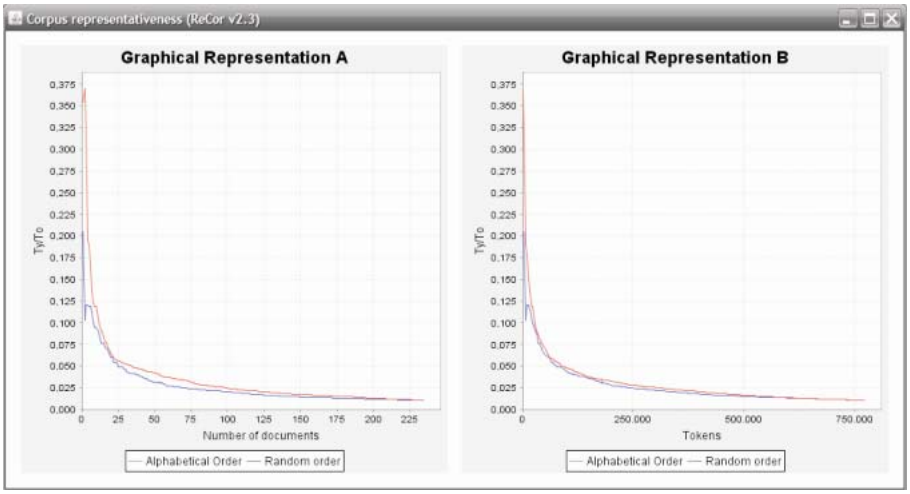
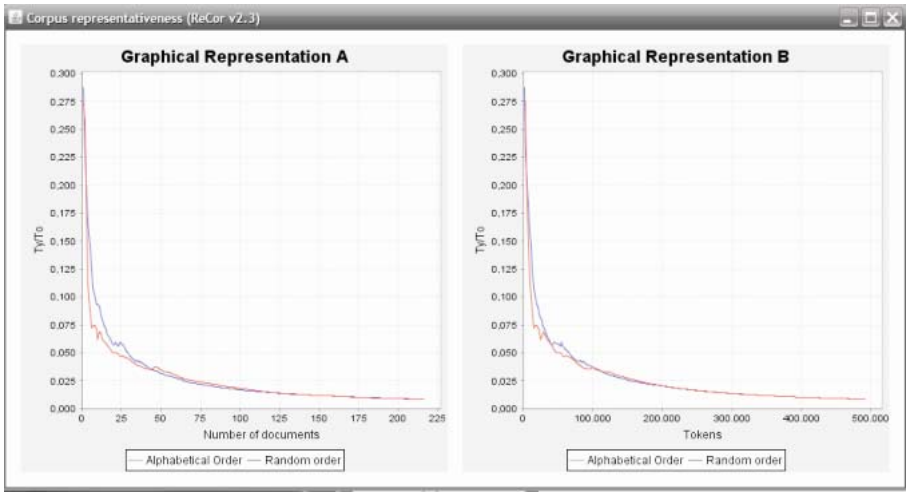


Figure 4: Representativeness of the American English component (1-gram).



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 225 docu-

ments; since the curve hardly varies either before or after this number, in other words this is the point where the lines stabilise 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,[28] all documents always contain an infinite 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 component in British English becomes representative from the point where 180 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.

In the case of Figure 4, it is possible to affirm from Graph A that the component in American English becomes representative with 150. Graph B shows that the minimum total number of words necessary for the corpus to be considered representative is 350,000 words.

A comparison of the two sets of graphs in Figures 2, 3 and 4 shows that the American English documents reach the point of representativeness long before the British English documents, firstly, and the Spanish documents secondly: 150 documents and 350,000 words in American English as against 180 documents and 600,000 words in British English and 225 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): 200 documents and 400,000 words in American English (Figure 7) as against 225 documents and 750,000 words in British English (Figure 6) and 250 documents and 800,000 words in Spanish (Figure 5).

It merits mentioning that neither corpora reach representativeness from 3-grams onwards, as the lines do not in practice stabilise. This means that more data are needed at this point in order to establish a representative threshold beyond bigrams.

[28] This filter also removes Roman numbers.

Figure 5: Representativeness of the Spanish component (2-gram).

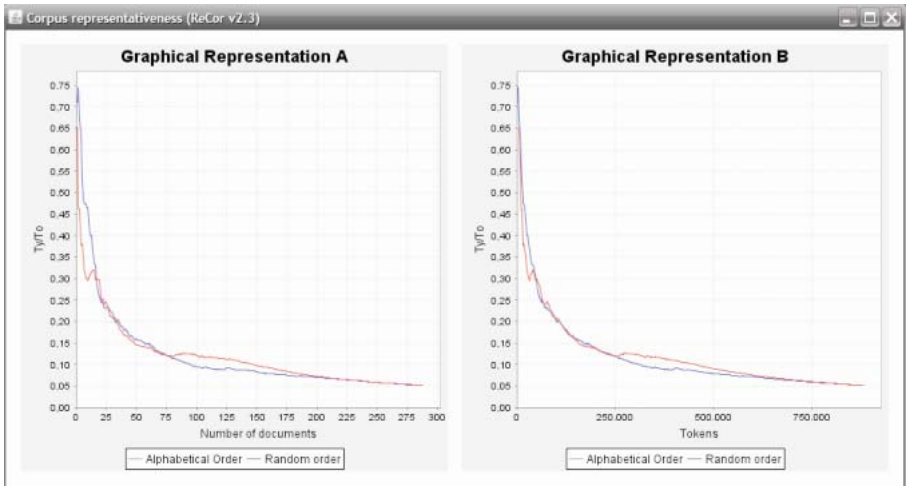


Figure 6: Representativeness of the British English component (2-gram).

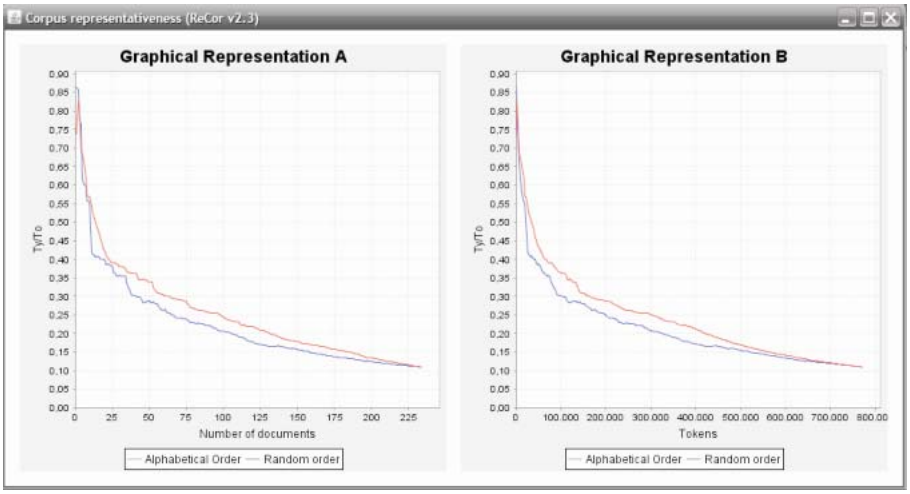


Figure 7: Representativeness of the American English component (2-gram).

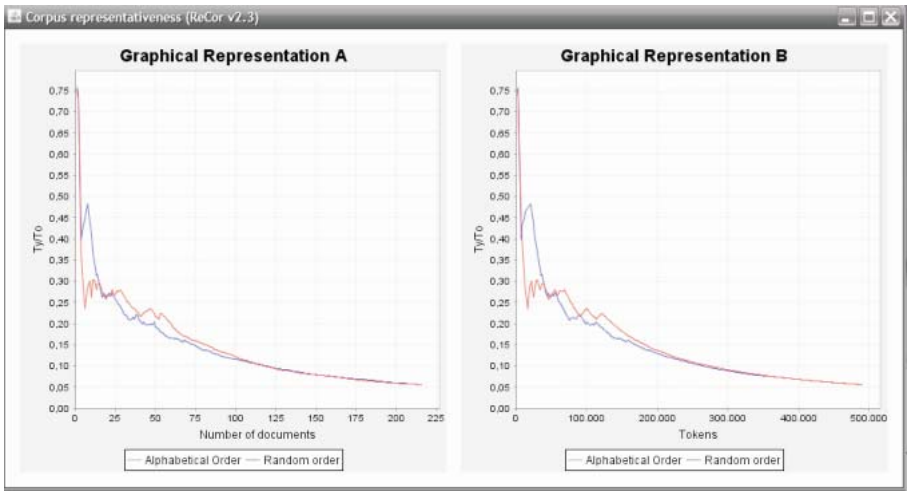


Figure 8: Representativeness of the Spanish component (3-gram).

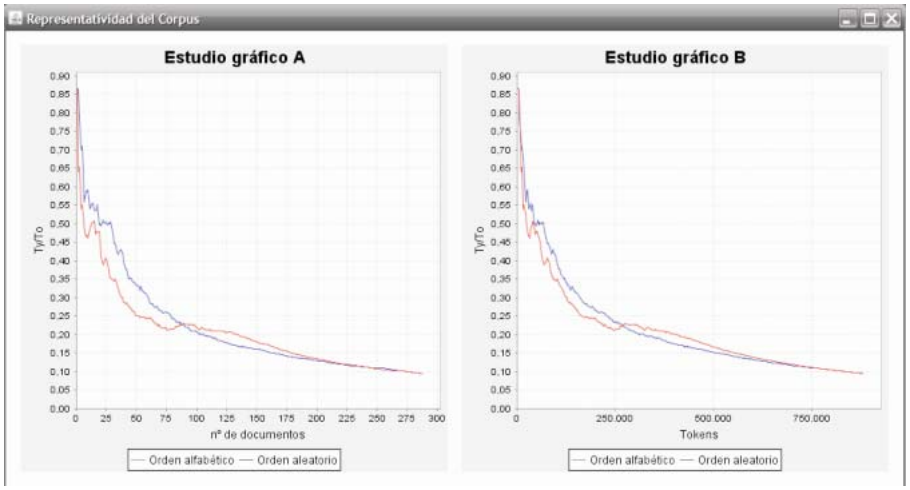


Figure 9: Representativeness of the British English component (3-gram).

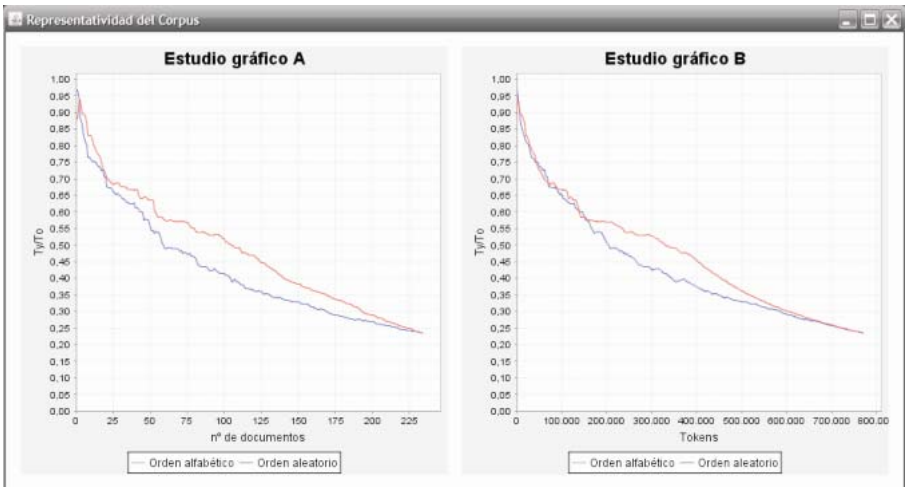
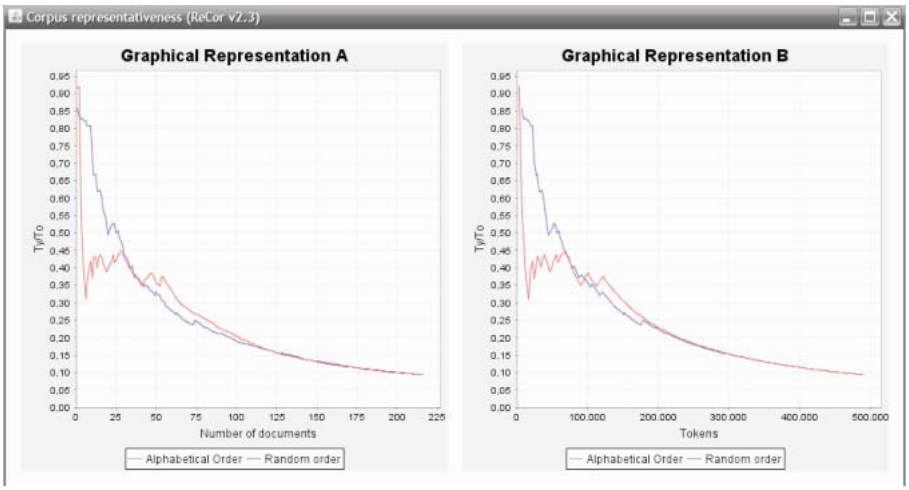


Figure 10: Representativeness of the American English component (3-gram).



From these results it may therefore be deduced that the American English general conditions tend to be more homogenous than those in British English and, mainly, those in Spanish. In other words, it is possible to infer that that the general conditions in American English

present super-, macro- and microstructures that are very similar to each other as well as using a narrower terminological range. There could be several plausible explanations for that. One possible reason could be that the three corpora were not lemmatized. Since Spanish is a very flexive language, it could be inferred that it reaches representativeness later on due to its richer inflectional morphology. Perhaps the three corpora should have been lemmatized for more accurate results.

However, a comparative analysis of German and Italian (two languages with rich flexional systems) shows that both package subcorpora reach representativeness even sooner than their English counterparts. The German subcorpus[29] reaches representativeness at the point of including 90 documents (250,000 words) for 1-grams, and 160 documents (385,000 words) for 2-grams. As to Italian, representativeness for 1-grams is attained at 115 documents (200,000 words), and at 175 documents (400,000 words) for 2-grams. Both components almost reach representativeness on a 3-gram basis at slightly more than 170 documents (390,000 words) and 200 (470,000 words), respectively. Clearly other factors rather than inflexional richness must be at stake.

5. Concluding remarks

So far academics have failed to resolve the questions of corpus representativeness and ideal size:

> However, despite the care given to selection, and despite the sheer size of these corpora, it is generally accepted that true representativity is illusory. We seem doomed to build larger and larger corpora at the risk of losing the wood for the trees. (Williams 2002: 44)

[29] The German component of the *Turicor* corpus contains 173 documents and 400,504 words. The Italian component is composed of 230 documents (550,035 words).

Yet, these controversial issues are of paramount importance when designing and evaluating specialised corpora. In practice such an evident lack of consensus poses insurmountable problems, as corpus compilation is by no means a mere question of beliefs. Yet, it is not possible to determine *a priori* the exact total number of words or documents that should be included in specialised language corpora (which in general tend to be smaller) in order that they may be considered representative. This is because, as it has been illustrated, size will be determined according to the language and text types as well as the restrictions of a particular specialised field, diatopic limitations, plus other functionally oriented criteria.

In this paper we have described a data-driven approach to evaluating corpus representativeness. No preconceived ideas or fixed figures have been used as starting points. Instead, a double approach to corpus building has been adopted, based on two arguments. Firstly, corpus representativeness can be obtained by establishing coherent diasystematic limits and carefully selecting textual genres for inclusion. These could be considered as external selection criteria to be established from the outset in order to ensure corpus representativeness and quality. Secondly, internal selection criteria can not be established on an *a priori* basis. The number of tokens and/or documents a specialised corpus should contain may vary in relation to the languages, domains and textual genres involved, as well as to the objectives set for a specific analysis (i.e., a corpus should provide enough evidence for the researchers' purposes and aims).

Corpus size is, thus, a valid internal criterium. However, it is only possible to determine that the corpus is of an adequate size after it has actually been compiled, (or, alternatively, during the compilation process as a quality control device), or even during analysis. The results obtained in this study support the assumption that English packages (both in the American and British varieties) tend to be more homogeneous than their Peninsular Spanish counterparts. And, consequently, corpus representativeness can be reached sooner in English than in Spanish. Further research should be carried out in order to find plausible explanations for those cross-linguistic and cross-cultural differences (e.g., language-specific features, culturally-bound textual constrains, differences in the legal systems involved) and whether those differences affect other types of tourism contracts or, even, other

types of legal documents. Achieving representativeness from internal criteria will, no doubt, provide 'food for thought' to academics within CL and will help it come of age eventually.

6. References

Ahmad, Khurshid and Rogers, Margaret. 2001. "Corpus linguistics and terminology extraction". In *Handbook of Terminology Management,* S. E. Wright and G. Budin (eds.). Amsterdam/Philadelphia: John Benjamins. 725-760.

Almahano Güeto, Inmaculada. 2002. *El contrato de viaje combinado en alemán y español: las condiciones generales. Un estudio basado en corpus.* Tesis doctoral inédita. Málaga: Universidad de Málaga.

Atkins, Sue, Clear, Jeremy and Ostler, Nicholas. 1992. "Corpus design criteria". *Literary and Linguistic Computing* 7 (1): 1-16.

Baker, Paul. 2006. *Using Corpora in Discourse Analysis.* London/ New York: Continuum.

Ball, Catherine N. 1997 [1996]. "Concordances and corpora" http://www.georgetown.edu/faculty/ballc/corpora/tutorial.html. Visited May 2008.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1990. "Methodological Issues Regarding Corpus-based Analyses of Linguistic Variations". *Literary and Linguistic Computing* 5: 257-269.

Biber, Douglas. 1993. "Representativeness in Corpus Design". *Literary and Linguistic Computing* 8 (4): 243-257. [Also: In *Current Issues in Computational Linguistics*, A. Zampolli, N. Calzolari and M. Palmer (eds.). Dordrecht and Pisa: Kluwer and Giardini. 377-408].

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison.* Cambridge: Cambridge University Press.

Biber, Douglas, Conrad, Susan and Reppen, Randi. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Borja Albi, Anabel. 2000. *El texto jurídico inglés y su traducción al español*. Barcelona: Ariel.

Bowker, Lynne and Pearson, Jennifer. 2002. *Working with Specialised Language: A Practical Guide to Using Corpora*. London: Routledge.

Braun, Eliezer. 2005 [1996]. "El caos ordena la lingüística. La ley de Zipf". In *Caos fractales y cosas raras*, E. Braun (ed). Mexico D.F.: Fondo de Cultura Económica. http://omega.ilce.edu.mx:3000/sites/ciencia/volumen3/ciencia3/150/htm/caos.htm. Visited May 2008.

Bravo Gozalo, José M. and Fernández Nistal, Purificación. 1998. "La lingüística del corpus, las nuevas tecnologías de la información y los Estudios de Traducción en la década de 1990". In *La traducción: orientaciones lingüísticas y culturales,* P. Fernández Nistal and J. M. Bravo Gozalo (eds.). Valladolid: Universidad de Valladolid. 205-257.

Cabré Castellví, M. Teresa. 1999. *Terminología: Representación y comunicación. Una teoría de base comunicativa y otros artículos.* Barcelona: Universidad Pompeu Fabra.

Carrasco Jiménez, Rafael Carlos. 2003. *La ley de Zipf en la Biblioteca Miguel de Cervantes.* Alicante: Universidad de Alicante. http://www.dlsi.ua.es/asignaturas/aa/Zipf.pdf. Visited May 2008.

Church, Kenneth Ward and Mercer, Robert L. 1993. "Introduction to the special issue on computational linguistics using large corpora". *Computational Linguistics* 19 (1): 1-24.

Clear, Jeremy H. 1994. "I can't see the sense in a large corpus". In *Papers in Computational Lexicography: COMPLEX '94*, F. Kiefer, G. Kiss and J. Pajzs (eds.). Budapest: Research Institute for Linguistics and Hungarian Academy of Sciences. 33-48.

CORIS/CODIS. 2006. "Progettazione e costruzione di un Corpus di Italiano Scritto". http://corpus.cilta.unibo.it:8080/coris_itaProgett.html. Visited May 2008.

Corpas Pastor, Gloria. 2002. "Traducir con corpus: de la teoría a la práctica". In *Texto, terminología y traducción,* J. García Palacios and M. T. Fuentes (eds.). Salamanca: Almar. 189-226.

Corpas Pastor, Gloria. 2003. "Tourism and travel law: Electronic resources for a corpus-based multilingual generation project". *Revista europea de derecho de la navegación marítima y aeronáutica* XIX: 2807-2818.

Corpas Pastor, Gloria. 2006. "Tourism and travel law in United States, Great Britain and Spain: A multifaceted approach to the compilation of a corpus of tourism contracts ("packages") from the WWW". Unpublished paper. European Law Research Center. Harvard Law School.

Corpas Pastor, Gloria and Seghiri, Miriam. 2006a. *El concepto de representatividad en la Lingüística del Corpus: aproximaciones teóricas y metodológicas.* Technical document BFF2003-04616 MCYT/TI-DT-2006-1.

Corpas Pastor, Gloria and Seghiri, Miriam. 2006b. "Recursos documentales para la traducción de seguros turísticos en el par de lenguas inglés-español". In *Investigación y traducción: una mirada al presente en la labor investigadora y en el ejercicio de la profesión de la licenciatura Traducción e Interpretación*, E. Postigo Pinazo (ed.). Málaga: Universidad de Málaga. 313-353.

Corpas Pastor, Gloria and Seghiri, Miriam. 2007a. "Specialized corpora for translators: A quantitative method to determine representativeness". *Translation Journal* 11 (3). http://translationjournal.net/journal/41corpus.htm. Visited May 2008.

Corpas Pastor, Gloria and Seghiri, Miriam. 2007b. "Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor". *Procesamiento del Lenguaje Natural* 39: 165-172. http://www.sepln.org/revistaSEPLN/revista/39/20.pdf. Visited May 2008.

Corpas Pastor, Gloria and Seghiri, Miriam. 2007c. "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)*. http://www.uma.es/hum892. Visited May 2008.

Corpas Pastor, Gloria and Seghiri, Miriam. 2008/forthcoming. *El concepto de representatividad en lingüística de corpus: Aproximaciones teóricas y consecuencias para la traducción.* Málaga: Servicio de Publicaciones de la Universidad.

Council Directive of 13 June 1990 on package travel, package holidays and package tours regulations, 90/314/EEC.

EAGLES. 1994. "Corpus typology: A framework for classification". EAGLES Document 080294. 1-18.

EAGLES. 1996a. "Text corpora Working Group reading Guide". EAGLES Document EAG-TCWG-FR-2. http://www.ilc.cnr.it/EAGLES/corpintr/corpintr.html. Visited May 2008.

EAGLES. 1996b. "Preliminary recommendations on corpus typology". EAGLES Document EAG-TCWG-CTYP/P. http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html. Visited May 2008.

Fillmore, Charles J.1992. "'Corpus linguistics' or 'Computer-aided armchair linguistics'". In *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991,* J. Svartvik (ed.). Berlin/New York: Mouton de Gruyter. 35-60.

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

Friedbichler, Ingrid and Friedbichler, Michael. 2000. "The potential of domain-specific target-language corpora for the translator's workbench". In *I corpora nella didattica della traduzione. Corpus Use and Learning to Translate,* S. Bernardini and F. Zanettin (eds.). Bologna: CLUEB. http://www.sslmit.unibo.it/cultpaps/fried.htm. Visited May 2008.

Gelbukh, Alexander, Sidorov, Grigori and Hernández, Liliana Chanona. 2002. "Corpus virtual, virtual: Un diccionario grande de contextos de palabras españolas compilado a través de Internet". In *Multilingual Information Access and Natural Language Processing, International Workshop (November 12) at IBERA-MIA-2002, VII Iberoamerican Conference on Artificial Intelligence*, J. Gonzalo, A. Peñas and A. Ferrández (eds.). Seville: IBERAMIA, ELSNET and RITOS-2. 7–14. http://nlp.uned.es/ia-mlia/iberamia2002/papers/mlia08.pdf. Visited May 2008.

Gervais, Daniel. 2003. "MultiTrans Pro 3". In *Entornos informáticos de la traducción profesional. Las memorias de traducción*, G. Corpas Pastor and M. J. Varela Salinas (eds.). Granada: Atrio. 139-154.

Ghadessy, Mohsen, Henry, Alex and Roseberry, Robert L. (eds.) 2001. *Small Corpus Studies and ELT: Theory and Practice.* Amsterdam/Philadelphia: John Benjamins.

Giouli, Voula and Piperidis, Stelios. 2002. *Corpora and HLT. Current Trends In Corpus Processing And Annotation.* Bulgaria: Insitute for Language and Speech Processing. http://www.larflast.bas.bg/balric/eng_files/corpora1.php. Visited May 2008.

Haan, Pieter de. 1989. *Postmodifying Clauses In The English Noun Phrase. A Corpus-Based Study*. Amsterdam: Rodopi.

Haan, Pieter de. 1992. "The optimum corpus sample size?". In *New Dimensions In English Language Corpora. Methodology, Results, Software Development*, G. Leitner (ed.). Berlin/NewYork: Mouton de Gruyter. 3-19.

Heaps, Harold Stanley. 1978. *Information Retrieval: Computational and Theoretical Aspects.* New York: Academic Press.

Jeong, Young-Mi. 1995. "Statistical characteristics of Korean vocabulary and its application" *Lexicographic Study* 5 (6): 134-163.

Johansson, Stig, Leech, Geoffrey N. and Goodluck, Helen. 1978. *Manual of Information to Accompany the Lancaster-Oslo/ Bergen Corpus of British English, For Use With Digital Computers.* Oslo: University. http://khnt.hit.uib.no/icame/manuals/lob /INDEX.HTM. Visited May 2008.

Kock, Josse de. 1997. "Gramática y corpus: los pronombres demostrativos". *Revista de filología románica* 14 (1): 291-298. http://www.ucm.es/BUCM/revistas/fll/0212999x/articulos/RFRM9797120291A.PDF. Visited May 2008.

Kock, Josse de. 2001. "Un corpus informatizado para la enseñanza de la lengua española. Punto de partido y término". *Hispanica Polonorum 3*. 60-86.

Lauer, Mark. 1995a. "How much is enough? Data requirements for statistical NLP". *Proceedings of the 2nd Conference of the Pacific Association for Computational Linguistics*. Brisbane: Pacific Association for Computational Linguistics. http://arxiv.org/PS_cache/cmp-lg/pdf/9509/9509001.pdf. Visited May 2008.

Lauer, Mark. 1995b. "Corpus statistics meet the noun compound: some empirical results". *Proceedings of the 33rd Annual Meet-*

ing of the Association for Computational Linguistics. Boston: Association for Computational Linguistics. 47-54. http://arxiv.org/PS_cache/cmp-lg/pdf/9504/9504033.pdf. Visited May 2008.

Lauer, Mark. 1995c. "Conserving fuel in statistical language learning: Predicting data requirements". In *Eighth Australian Joint Conference on Artificial Intelligence,* X. Yao (ed.). Canberra: University College, the University of New South Wales, Australian Defence Force Academy. http://arxiv.org/PS_cache/cmp-lg/pdf/9509/9509002.pdf. Visited May 2008.

Lavid López, Julia. 2005. *Lenguaje y nuevas tecnologías: nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*. Madrid: Cátedra.

Laviosa, Sara. (ed). 1998. "L'approche basée sur le corpus / The corpus-based approach". *Meta* 43 (4).

Laviosa, Sara. 1997. "How comparable can 'comparable corpora' be?". *Target* 9 (2): 289-319.

Leech, Geoffrey. 1991. "The state of the art in corpus linguistics". *English Corpus Linguistics*, K. Aijmer and B. Altenberg (eds.). London: Longman. 8-29. http://ling.kgw.tu-berlin.de/corpus/art.htm. Visited May 2008.

Leech, Geoffrey. 1992. "Corpora and theories of linguistics performance". *Directions in Corpus Linguistics: Proceedings of Nobel Symposium*, J. Svartvik (ed.). Berlin: Mouton de Gruyter. 105-122.

Lorch, Robert F**.** and Myers, Jerome L. 1990. "Regression analyses of repeated measures data in cognitive research". *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16: 149-157.

Maggi, Romano and Trujillo Pérez, Juan Diego. 2006. *ReCor (v. 1.0): Diseño e implementación de una aplicación informática para determinar la representatividad de un corpus.* Technical document BFF2003-04616 MCYT/TI-DT-2006-2. 1-18.

McEnery, Anthony and Wilson Andrew. 2001 [1996]. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery, Anthony and Wilson Andrew. 2006 [2000]. "ICT4LT Module 3.4. Corpus linguistics". http://www.ict4lt.org/en/en_mod3-4.htm. Visited May 2008.

McEnery, Tony; Xiao, Richard and Tono, Yukio. (eds). 2006. *Corpus-Based Language Studies. An Advanced Resource Book.* London and New York: Routledge.

Moreiro González, José Antonio. 2002. "Aplicaciones al análisis automático del contenido provenientes de la teoría matemática de la información". *Anales de documentación* 5: 273-286. http://www.um.es/fccd/anales/ad05/ad0515.pdf. Visited May 2008.

Moreno Sevilla, Juan Diego. 2006. *Generador de documentos XML conforme al estándar TEI para el tratamiento automatizado de un corpus*. Málaga: Universidad de Málaga.

Murison-Bowie, Simon. 1993. *MicroConcord Manual. An Introduction To The Practices And Principles Of Concordancing In Language Teaching.* Oxford: Oxford University Press.

Pearson, Jennifer. 1998. *Terms in Context, Studies in Corpus Linguistics.* Amsterdam/Philadelphia: John Benjamins Publishing.

Pérez Hernández, M. Chantal. 2002a. "Terminografía basada en corpus: principios teóricos y metodológicos". In *Investigar en terminología,* P. Faber and C. Jiménez (eds.). Granada: Comares. 127-166.

Pérez Hernández, M. Chantal. 2002b. "Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento". *Estudios de Lingüística Española (ELiEs)* 18. http://elies.rediris.es/elies18/. Visited May 2008.

Quirk, Randolph. 1992. "On corpus principles and design". In *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991,* J. Svartvik (ed.). Berlin/ New-York: Mouton de Gruyter. 457- 469.

Ruiz Antón, Juan Carlos. 2006. "Corpus y otros recursos lingüísticos". 1-21. http://www.trad.uji.es/asignatura/obtener.php?letra=1&codigo=42&fichero=1098353728142. Visited May 2008.

Rundell, Michael and Stock, Penni. 1992. "The corpus revolution". *English Today* 8 (4): 45-51.

Sanahuja, Sonia and Silva, Ana. 2001. "Muestreo teórico y estudios del discurso. Una propuesta teórico-metodológica para la generación de categorías significativas en el campo del Análisis del Discurso". *El estudio del discurso: Metodología multidisciplinaria. II Coloquio Nacional de Investigadores en Estudios del Discurso. La Plata, 6 al 8 de septiembre de 2001.* Buenos Aires: Asociación Latinoamericana de Estudios del Discurso and Universidad Nacional del Centro de la Provincia de Buenos Aires. http://www.sai.com.ar/KUCORIA/discurso.html. Visited May 2008.

Sánchez Pérez, Aquilino and Cantos Gómez, Pascual. 1997. "Predictability of word forms (types) and lemmas in linguistic corpora. A case study based on the analysis of the CUMBRE corpus: An 8-million-word corpus of contemporary Spanish". *International Journal of Corpus Linguistics* 2 (2): 259-280.

Sánchez Pérez, Aquilino and Cantos Gómez, Pascual. 1998. "El ritmo incremental de palabras nuevas en los repertorios de textos. Estudio experimental y comparativo basado en dos corpus lingüísticos equivalentes de cuatro millones de palabras, de las lenguas inglesa y española y en cinco autores de ambas lenguas". *Atlantis* XIX (2): 205-223. http://dialnet.unirioja.es/servlet/fichero_articulo?articulo=637978&orden=38848. Visited May 2008.

Sánchez-Gijón, Pilar. 2002. "Aplicaciones de la lingüística de corpus a la práctica de la traducción. Complemento de la traducción asistida por ordenador". *Terminologie et Traduction* 2. http://europa.eu.int/comm/translation/bulletins/puntoycoma/79/pyc7910_es.htm. Visited May 2008.

Sánchez-Gijón, Pilar. 2003b. *Els documents digitals especialitzats: utilització de la lingüística de corpus com a front de recursos per a la traducció.* PhD. Dissertation. Barcelona: Universitat Autònoma de Barcelona. ISBN: 84-688-3918-3.

Sánchez-Gijón, Pilar. 2004. *L'ús de corpus en la traducció especialitzada: compilació de corpus ad hoc i extracció de recursos terminològics.* Barcelona: IULA. http://www.tdx.cesca.es/TESIS_UAB/AVAILABLE/TDX-0123104-173209/. Visited May 2008.

Seghiri, Miriam. 2006. *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]*. PhD Dissertation. Málaga: Universidad de Málaga. ISBN: 978-84-690-5775-9. http://www.sci.uma.es/bbldoc/tesisuma/16754888.pdf. Visited May 2008.

Seghiri, Miriam. 2008a/forthcoming. "Virtual corpus: A systematic methodology for compilation». *Proceedings of the American Association for Applied Linguistics Annual Conference (AAAL 2008, Washington, D.C.).*

Seghiri, Miriam. 2008b/forthcoming. "Using corpora in translation training: a step-by-step approach". *Proceedings of the American Translation and Interpreting Studies Association Conference (ATISA 2008, University of Texas at El Paso).*

Sinclair, John M. (ed.) 1987. *Looking Up: an Account of the COBUILD Project in Lexical Computing*. London: Collins.

Sinclair, John M. (ed). 1987. *Collins COBUILD English Language Dictionary*. London: Collins.

Sinclair, John M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John M. 2004a. "Corpus and text: Basic principles". In *Developing Linguistic Corpora: A Guide to Good Practice,* M. Wynne (ed.). Oxford: Oxford University Press. http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter1. htm. Visited May 2008.

Sinclair, John M. 2004b. *Trust the Text: Language, Corpus and Discourse*. Routledge, an imprint of Taylor & Francis Books.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work.*Amsterdam/ Philadelphia: John Benjamins

Velasco, Manuel; Díaz, Irene; Lloréns, Juan; Amescua, Antonio and Martínez, Vicente. 1999. "Algoritmo de filtrado multi-término para la obtención de relaciones jerárquicas en la construcción automática de un tesauro de descriptores". *Revista Española de Documentación Científica* 22 (1): 34-49. http://bddoc.csic.es:8080/basisbwdocs_rdisoc/rev0001/1999_vol22-1/1999_vol22-1_pp34-49.htm. Visited May 2008.

Wilkinson, Michael. 2005. “Compiling a specialised corpus to be used as a translation aid”. *Translation Journal* 9 (3) : 1-6. http://www.joensuu.fi/hallinto/jopke/dokumentit/Wilkinson.doc. Visited May 2008.

Williams, Geoffrey. 2002. “In search of representativity in specialised corpora”. *International Journal of Corpus Linguistics* 7 (1): 43-64.

Wright, Sue Ellen and Budin, Gerhard. 1997. *Handbook of Terminology Management.* Amsterdam/Philadelphia: John Benjamins.

Yang, Dan-Hee; Cantos Gómez, Pascual and Song, Mansuk. 2000. “An algorithm for predicting the relationship between lemmas and corpus size”. *ETRI Journal* 22 (2): 20-31. http://etrij.etri.re.kr/Cyber/servlet/GetFile?fileid=SPF-1042453354988. Visited May 2008.

Yang, Dan-Hee; Lee, I. and Cantos Gómez, Pascual. 2002. “On the corpus size needed for compiling a comprehensive computational lexicon by automatic lexical acquisition”. *Computers and the Humanities* 36 (2). 171-190. http://www.springerlink.com/media/f83etnxtxp4xdyvnyd9l/contributions/g/y/g/6/gyg6pa77k51cuaqr.pdf. Visited May 2008.

Yang, Dan-Hee; Lim, Soojong and Song, Mansuk. 1999. “The estimate of the corpus size for solving data sparseness”. *Journal of KISS* 26 (4): 568-583.

Zampolli, Antonio; Calzolari, Nicoletta and Palmer, Martha. (eds). 1994. *Current Issues in Computational Linguistics: In Honour of Don Walker.* Dordrecht/ Pisa: Kluwer and Giardini.

Zanettin, Federico. 1998. “Bilingual corpora and the training of translators”. *Meta* 43 (4): 616-630. http://www.erudit.org/revue/meta/1998/v43/n4/004638ar.pdf. Visited May 2008.

Zanettin, Federico. 2002a. “DIY Corpora: The WWW and the Translator”. In *Training the Language Services Provider for the New Millennium,* B. Maia; J. Haller and M. Urlrych (eds.) Porto: Facultade de Letras, Universidade do Porto. http://www.federicozanettin.net/DIYcorpora.htm. Visited May 2008.

Zanettin, Federico. 2002b. “CEXI. Designing an English Italian translational corpus”. in *Teaching and Learning by Doing Corpus*

Analysis, B. Ketteman and G. Marko (eds). Amsterdam: Rodopi. 329-343.

Zanettin, Federico; Bernardini, Silvia; and Stewart, Dominic. (eds). 2003. *Corpora in Translator Education.* Manchester: St. Jerome.