*Miriam Seghiri**

Too Big or Not Too Big: Establishing the Minimum Size for a Legal Ad Hoc Corpus

Abstract

A corpus can be described as "[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). 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 (Seghiri 2008). In fact, 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/Piperidis 2002). For this reason, in this paper we will attempt to deal with this key question: we will focus on the complex notion of representativeness and ideal size for ad hoc corpora, from both a theoretical and an applied perspective and we will describe a computer application named ReCor that will be used to verify whether a sample of legal contracts compiled might be considered representative from the quantitative point of view.

1. Introduction

Corpus-driven/based studies rely on the representativeness of each corpus as their true foundation for producing valid results (cf. Biber et al. 1988: 246). However, according to Leech (1991: 2) the assumption of representativeness "must be regarded largely as an act of faith". Actually, as Tognini-Bonelli (2001: 57) stated "at present we have no means of ensuring it, or even evaluating it objectively". Unfortunately, faith and beliefs do not seem to ensure quality… For this reason, in this paper we will attempt to deal with this key question: we will focus on the complex notion of representativeness and ideal size for ad hoc corpora, from both a theoretical and an applied perspective and we will describe a computer application named ReCor, version 2.5, that will be used to verify whether a sample legal ad hoc corpus might be considered representative from the quantitative point of view.

2. The Importance of Being Representative

Thousands of definitions have been provided as to what constitutes a corpus as the followings: "[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 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); "a finite-sized body of machine-readable texts sampled in order to be maximally **representative** of the language variety under consideration" (McEnery/Wilson 2001 [1996]: 24), among others. However, despite the repeated reference to the quality of being *representative* 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: "[t]he definition of representativeness is a crucial point in the creation of a corpus, but is one of the most controversial aspects among spe-

* *Miriam Seghiri, B.A, M.A., PhD.*
Senior Lecturer (with tenure)
Facultad de Filosofía y Letras
Universidad de Málaga
29071-Málaga (Spain)
seghiri@uma.es

cialists, especially as regards the ambiguity inherent in its use due to the intermingling of **quantitative** and **qualitative** connotations" (CORIS/CODIS 2006).

3. Qualitative representativeness

Dealing with the first concept, 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 can be solved by designing a system for gauging the quality of digital information (cf. Seghiri 2006: 89-95). 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 contracts.[1] This corpus will be *monolingual* (Spanish), and diatopically restricted to Spain, due to the large number of countries in which this language is spoken. It will be a *comparable full-text* corpus because it will include complete contracts originally written in Spanish, all of them downloaded from the web, so the corpus will be also *electronic*. Finally, as the corpus will only include travel insurance contracts, it will be *homogenous* in genre and topic.

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 ad hoc 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. On the one hand, the 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. Contracts on this topic have been mainly downloaded from web sites of Spanish insurance companies such as *MAPFRE, Ocaso,* among others. A list of the main insurance companies in Spain can be downloaded from the Spanish Association of Insurance Companies, named Asociación Empresarial del Seguro.[2] 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, for instance. However, according to a great number of analysts (cfr. Radev et al. 2005), Google is the best search engine in terms of the quality of search results. On this point, it is clearly essential to establish descriptors (like *travel insurance* and *contract*) and using Boolean operators (like AND, OR), 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 Spain" (.es) in order to filter pages from other English spoken countries.

Once the Spanish contracts 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 BootCaT,[3] for instance, which allows downloading groups of contracts from a single webpage.

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*[4] or *Concordance*,[5] to

1 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 recognize 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.

2 http://www.unespa.es/frontend/unespa/buscador_guia.php.

3 http://bootcat.sslmit.unibo.it.

4 http://www.lexically.net/wordsmith.

5 http://www.concordancesoftware.co.uk.

name just a few, in accordance with the *clean-text policy* described by Sinclair (1991): "[t]he safest policy is to keep to the text as it is, unprocessed and clean of any other codes".

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 *AbbyFine Reader*[6] 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 topic – travel insurance –, language – Spanish – and formats – original format and plain text –. The texts have been automatically codified (cfr. Figure 1) with the programme Lupas Rename as follows: number (01), language (TO stands for "original text", and ES stands for "Spanish") and genre (CO means "contract").

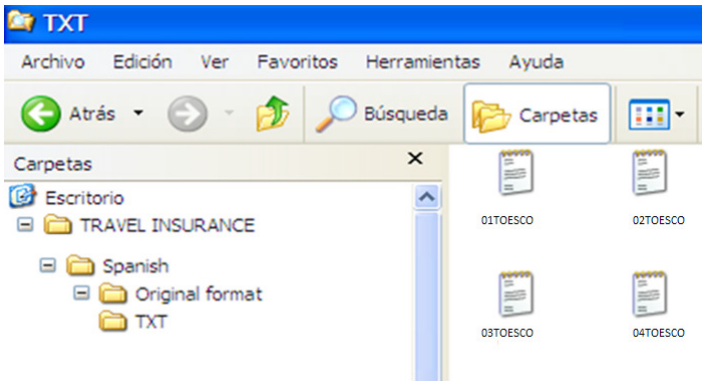


Figure 1. Storage data

In the study now under examination an ad hoc corpus of travel insurance contracts in Spanish was compiled, with 92 documents and 901,869 words (tokens). Quality has been assured through a set of clear design criteria and a compilation protocol divided into four steps. But, the quantity of documents and words (tokens) is enough to cover the terms used in this topic and genre?

4. Quantitative representativeness

According to Lavid (2005), the size of the corpus is a decisive factor in determining whether the sample is representative in relation to the needs of the translation. However, 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. However, many authors state that there is no general agreement as to what the size of a corpus should ideally be and, "[u]sually, the availability of material in the particular field of study determines the final size of the corpus" (Giouli/Piperidis 2002).

4.1. Zipf's law approach

There have been a great number of papers on the question of quantity as a criterion to reach representativeness as well as suggested formulas for calculating *a priori* the minimum number of words and documents necessary for a specialist corpus to be considered representative (cf. Heaps 1978; Biber 1988, 1990, 1993, 1994 and 1995; Leech 1991; Biber et al. 1998 and Yang et al. 1999 and 2002, amongst others). Most of these formulas are based on Zipf's law. Zipf's law is based on the idea that all texts contain a number of words that are repeated, i.e., the total number of words in any text is referred to as *tokens*, while the total number of distinct words, without counting rep-

6 http://www.abbyyeu.com.

etitions, is known as *types*. If types are divided into tokens, the result will be the frequency of each word in the corpus. 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 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 frequency 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*). According to this information, Zipf's law can be expressed mathematically as follows

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

Figure 2. Zipf's law

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 [1996] 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),[7] Young-Mi (1995) and Sánchez Pérez/Cantos Gómez (1997). However, subsequently some of these authors such as Cantos (cfr. 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.[8] This might be attributed to their preference for Zipf's law".[9]

3.2. Minimum Size Recommendations

It is 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 2004) and where an approximate figure is proposed, many authors appear to take extreme positions. Thus, Sinclair (2004) considers that ideally a corpus should be 'big', although the interpretation of this adjective remains open to debate because no approximate figure is given. McEnery/Wilson (2006 [2000]), Borja-Albi (2000) and Ruiz Antón (2006) suggest that the ideal number of words that any corpus should reach is around a million. Friedbichler/Friedbichler (2000) consider that a figure between "500,000 and 5 million words per language (depending on the target field) will provide sample

7 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 (cf. Seghiri 2006: 399-403).

8 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/Myers (1990).

9 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).

evidence in 97 % of language queries". 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 text with "a few thousand and a few hundred thousand words" are just as useful in the study of languages for specific purposes. Thus, Clear (1994) wrote an article: "I Can't See the Sense in a Large Corpus". 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 (cfr. Murison-Bowie 1993: 50). 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 (1995: 131) 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.

If these principles are applied to the particular case under examination here, it may be stated that the ad hoc corpus on travel insurance contracts 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 an insurance for travelling) and with only one text type being represented (contract), whose frequency in general language use is minimal. In addition, Bravo Gozalo/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 examination has a very specific objective, its size could be even further reduced, taking this consideration into account.

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 ranges of figures that have been suggested differ widely and the proposed calculations are not particularly reliable.[10] In a previous study (cfr. Seghiri, 2006) 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 an algorithm, called N-Cor, that can represent increases in the corpus (C) on a document by document (d) basis, for example:

$$Cn = d1 + d2 + d3 + ... + dn$$

Figure 3. N-Cor Algorithm

Following from this, our starting point is the idea forwarded by Biber (1993) and subsequently endorsed in studies such as those by Sánchez Pérez/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 for the first time the minimum size of a corpus *a posteriori*. With the help of graphs, it should be possible to establish whether the corpus is representative and how many documents and words (tokens) are necessary to achieve this. This theory has become a practical reality in the shape of a software application,

10 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.

named ReCor, which enables accurate evaluation of corpus representativeness[11], as described in the next section. The ReCor programme has been developed on the bases of the N-Cor algorithm (cfr. Figure 3) which was patented in 2010 by the Spanish Patent and Trademark Office.[12]

5. ReCor 2.5

ReCor is a software application which has been designed to facilitate the evaluation of representativeness of corpora in relation to their size. In this study we used version 2.5 of ReCor, 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.

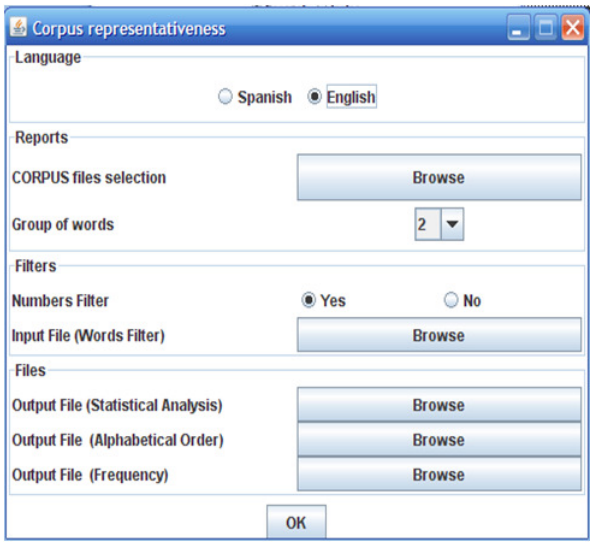


Figure 4: The *ReCor* interface (English version).

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.[13] 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 corpus is representative, and at precisely this point it is possible to see how many texts and words (tokens) 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 Spanish contracts compiled is sufficient to enable us to affirm that our corpus is representative (with 1-gram).

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

12 http://umapatent.uma.es/es/patent/metodo-para-la-determinacion-de-la-representa4b0.

13 It should be noted here that zero sometimes 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.

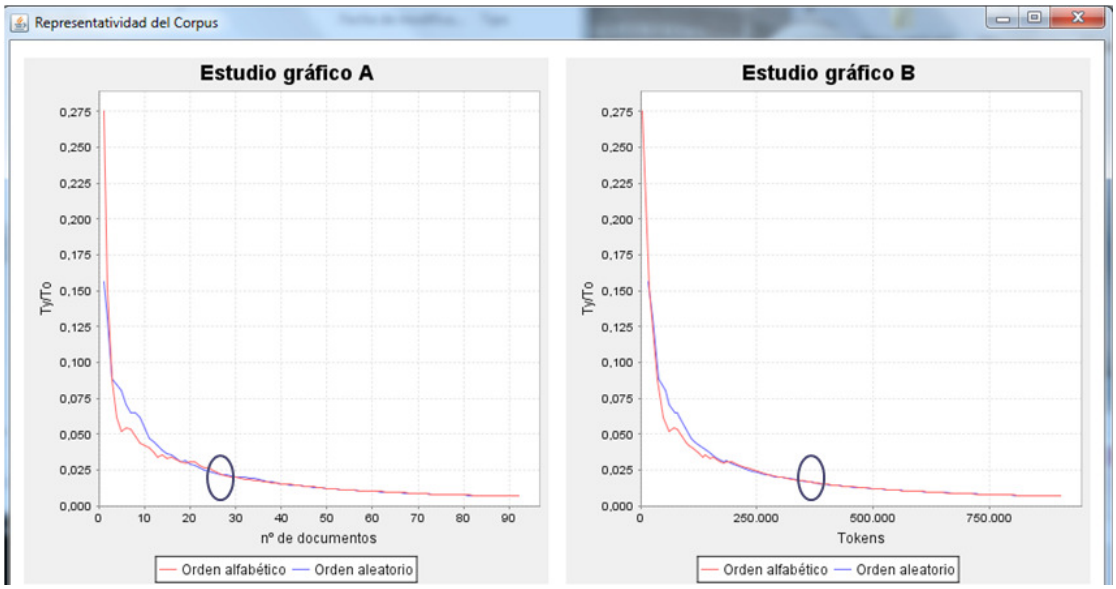


Figure 5. Representativeness of the Spanish corpus (1-gram).

From the data shown in Figure 5 it is possible to deduce, according to Graph A (*Estudio gráfico A*), that the corpus begins to be representative from the point of the inclusion of 25 documents; 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. Graph B (*Estudio gráfico B*) shows the minimum total number of words (tokens) necessary for the corpus to be considered representative, which in this case is 300,000 words approximately (319,494 words exactly, cfr. Figure 7).

We can also check if the corpus is representative from 2 to 10 grams, in order to carry out collocational al phraseological studies. To illustrate this, we will check if the corpus is representative with 2 grams (see Figure 6):

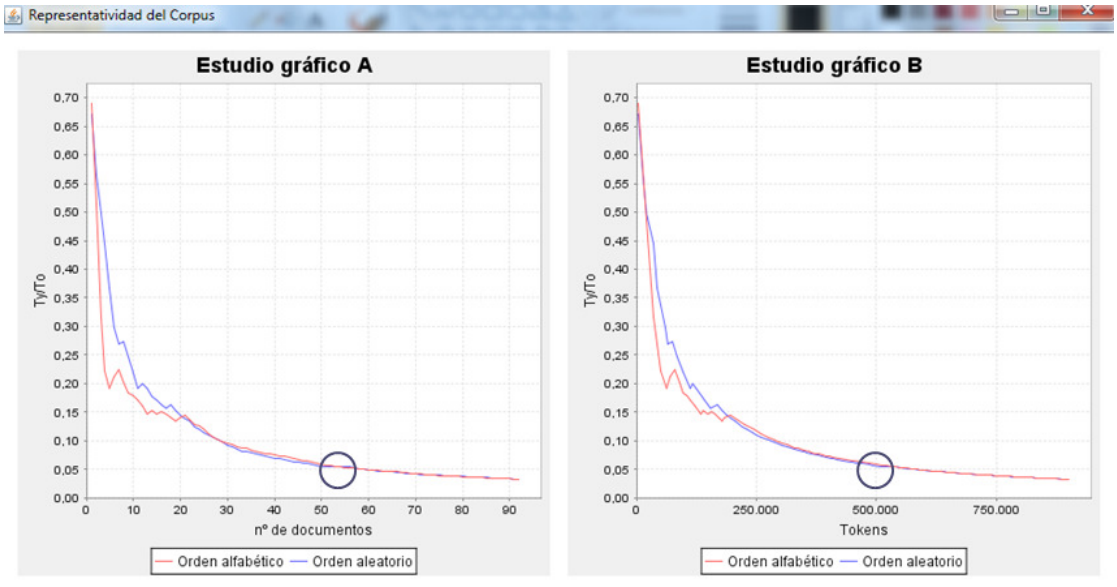
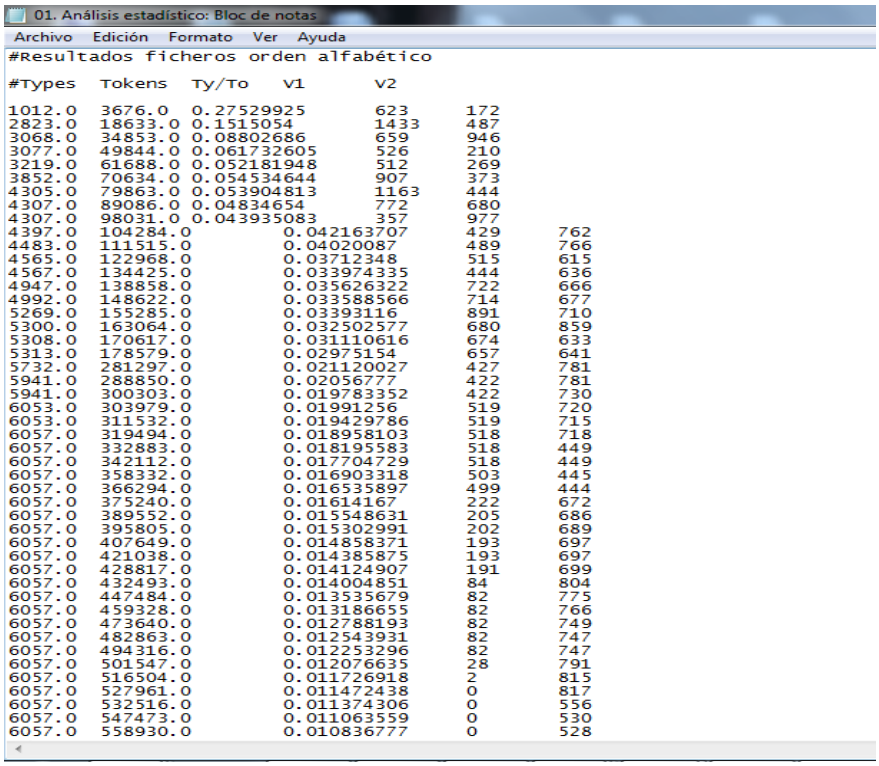


Figure 6. Representativeness of the Spanish corpus (2-grams).

From the data shown in Figure 6 it is possible to state that, according to Graph A (*Estudio gráfico A*), the corpus begins to be representative (with 2-grams) from the point of the inclusion of

52 documents; since the curve hardly varies either before or after this number. Graph B (*Estudio gráfico B*) shows the minimum total number of words (tokens) necessary for the corpus to be considered representative, which in this case is 500,000 words approximately (527,108 words exactly, cfr. Figure 8).

At the same time, three output files are created (in plain text and excel). The first output file, *Statistical Analysis*, shows the results from two distinct analyses; firstly, with the files *ordered alphabetically* by name (see Figure 7 for 1-gram and Figure 9 for 2 grams) and secondly with the files in random order (see Figure 8 for 1-gram and Figure 10 for 2 grams). 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 at least one more time, i.e. one type plus one token (V1) and the number of words that appear at least twice, i.e., one type plus two tokens (V2).



01. Análisis estadístico: Bloc de notas

Archivo Edición Formato Ver Ayuda

#Resultados ficheros orden alfabético

#Types	Tokens	Ty/To	V1	V2
1012.0	3676.0	0.27529925	623	172
2823.0	18633.0	0.1515054	1433	487
3068.0	34853.0	0.08802686	659	946
3077.0	49844.0	0.061732605	526	210
3219.0	61688.0	0.052181948	512	269
3852.0	70634.0	0.054534644	907	373
4305.0	79863.0	0.053904813	1163	444
4307.0	89086.0	0.04834654	772	680
4307.0	98031.0	0.043935083	357	977
4397.0	104284.0	0.042163707	429	762
4483.0	111515.0	0.04020087	489	766
4565.0	122968.0	0.03712348	515	615
4567.0	134425.0	0.033974335	444	636
4947.0	138858.0	0.035626322	722	666
4992.0	148622.0	0.033588566	714	677
5269.0	155285.0	0.03393116	891	710
5300.0	163064.0	0.032502577	680	859
5308.0	170617.0	0.031110616	674	633
5313.0	178579.0	0.02975154	657	641
5732.0	281297.0	0.021120027	427	781
5941.0	288850.0	0.02056777	422	781
5941.0	300303.0	0.019783352	422	730
6053.0	303979.0	0.01991256	519	720
6053.0	311532.0	0.019429786	519	715
6057.0	319494.0	0.018958103	518	718
6057.0	332883.0	0.018195583	518	449
6057.0	342112.0	0.017704729	518	449
6057.0	358332.0	0.016903318	503	445
6057.0	366294.0	0.016535897	499	444
6057.0	375240.0	0.01614167	222	672
6057.0	389552.0	0.015548631	205	686
6057.0	395805.0	0.015302991	202	689
6057.0	407649.0	0.014858371	193	697
6057.0	421038.0	0.014385875	193	697
6057.0	428817.0	0.014124907	191	699
6057.0	432493.0	0.014004851	84	804
6057.0	447484.0	0.013535679	82	775
6057.0	459328.0	0.013186655	82	766
6057.0	473640.0	0.012788193	82	749
6057.0	482863.0	0.012543931	82	747
6057.0	494316.0	0.012253296	82	747
6057.0	501547.0	0.012076635	28	791
6057.0	516504.0	0.011726918	2	815
6057.0	527961.0	0.011472438	0	817
6057.0	532516.0	0.011374306	0	556
6057.0	547473.0	0.011063559	0	530
6057.0	558930.0	0.010836777	0	528

Figure 7. Statistical analysis ordered alphabetically (1-gram).

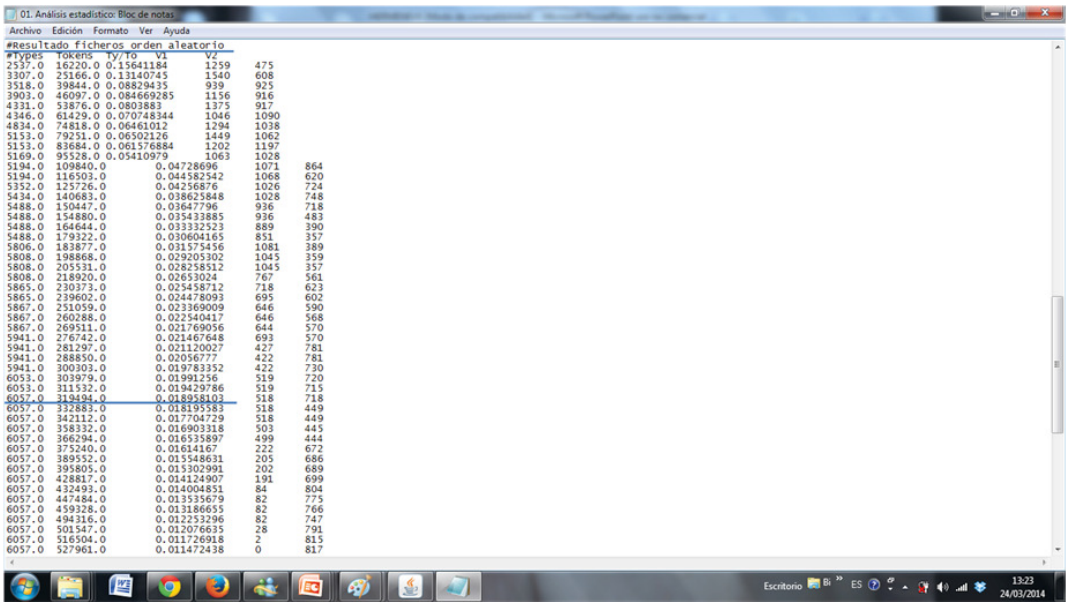


Figure 8. Statistical analysis at random (1-gram).

We can see (cfr. Figure 7 & 8) that with 6,057 types and 319,494 tokens the corpus grows in size (i.e. tokens) but not in grams (i.e. types), so we can confirm that as no 1-gram-types are entering in our corpus, the minimum size has been reached. As for 2 grams, the ReCor programme creates the following statistical analysis:

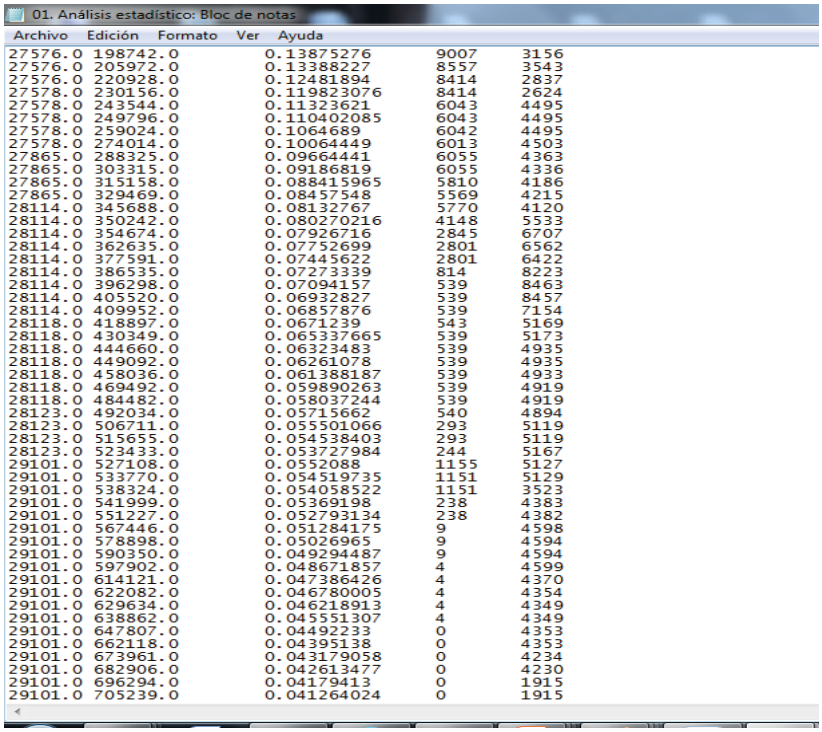


Figure 9. Statistical analysis ordered alphabetically (2-grams).

01. Análisis estadístico: Bloc de notas

Archivo Edición Formato Ver Ayuda

27865.0	303315.0	0.09186819	6055	4336
27865.0	315158.0	0.088415965	5810	4186
27865.0	329469.0	0.08457548	5569	4215
28114.0	345688.0	0.08132767	5770	4120
28114.0	350242.0	0.080270216	4148	5533
28114.0	354674.0	0.07926716	2845	6707
28114.0	362635.0	0.07752699	2801	6562
28114.0	377591.0	0.07445622	2801	6422
28114.0	386535.0	0.07273339	814	8223
28114.0	396298.0	0.07094157	539	8463
28114.0	405520.0	0.06932827	539	8457
28114.0	409952.0	0.06857876	539	7154
28118.0	418897.0	0.0671239	543	5169
28118.0	430349.0	0.065337665	539	5173
28118.0	444660.0	0.06323483	539	4935
28118.0	455485.0	0.058037244	539	4919
28123.0	467035.0	0.05715662	540	4894
28123.0	506711.0	0.055501066	293	5119
29101.0	527108.0	0.0552088	1155	5127
29101.0	533770.0	0.054519735	1151	5129
29101.0	538324.0	0.054058522	1151	3523
29101.0	541999.0	0.05369198	238	4383
29101.0	551227.0	0.052793134	238	4382
29101.0	567446.0	0.051284175	9	4598
29101.0	578898.0	0.05026965	9	4594
29101.0	590350.0	0.049294487	9	4594
29101.0	597902.0	0.048671857	4	4599
29101.0	614121.0	0.047386426	4	4370
29101.0	622082.0	0.046780005	4	4354
29101.0	629634.0	0.046218913	4	4349
29101.0	638862.0	0.045551307	4	4349
29101.0	647807.0	0.04492233	0	4353
29101.0	662118.0	0.04395138	0	4353
29101.0	673961.0	0.043179058	0	4234
29101.0	682906.0	0.042613477	0	4230
29101.0	696294.0	0.04179413	0	1915
29101.0	705239.0	0.041264024	0	1915

Figure 10. Statistical analysis at random (2-grams).

According to Figures 9 and 10, the corpus with 2-grams is quantitative representative with 29,101 types and 527,108 tokens as no 2-gram-types are entering in the corpus.

The second output file, 'Alphabetical Order,' generates two columns; the first column shows the words in alphabetical order with their corresponding number of occurrences appearing in the second one (cfr. Figures 11 & 12).

```
#Fichero ordenado por palabra
#Palabra          Frecuencia

          336
a         18372
a07       20
ab        4
abajo     16
abandono          44
abdomen 24
abierto 8
abintestato       8
ablación          24
abogado 28
abogados          4
abolición         12
abona     16
abonada 44
abonadas          36
abonado 4
abonados          12
abonar    60
abonarse          4
abonará 256
abone     28
abono     84
abonos    8
abortos 4
abril     156
absoluta          236
absoluto          20
abuelos 96
acaba     12
acabados          24
acaecido          60
acaecidos         228
```

Figure 11. Types ordered alphabetically (1-gram).

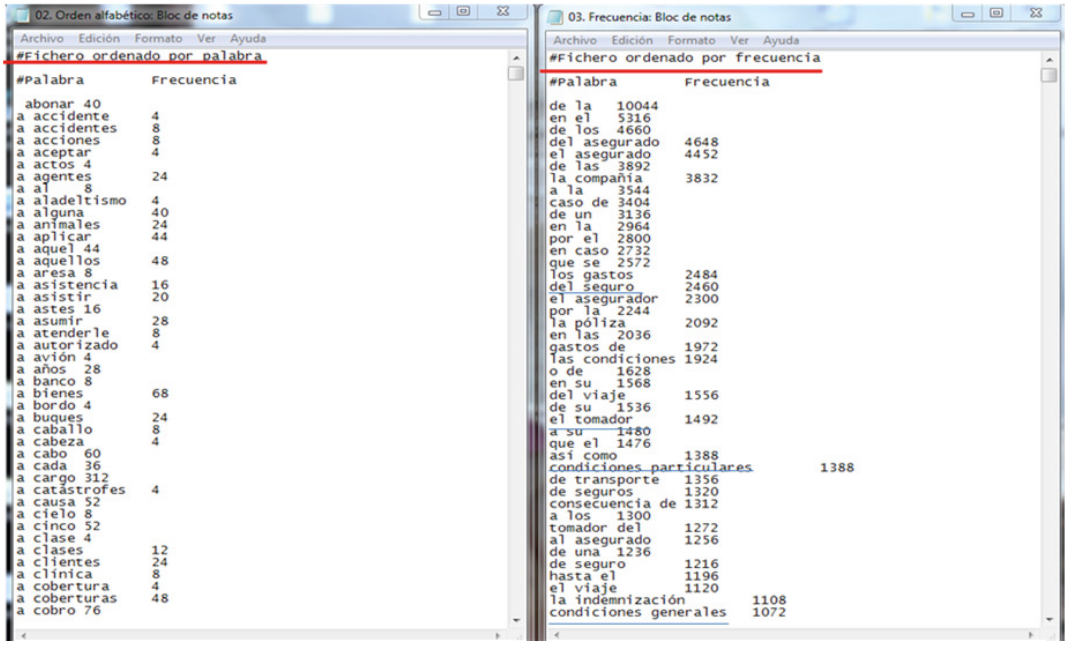


Figure 12. Types ordered alphabetically (2-gram).

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. From this list 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 (see Figures 13 and 14).

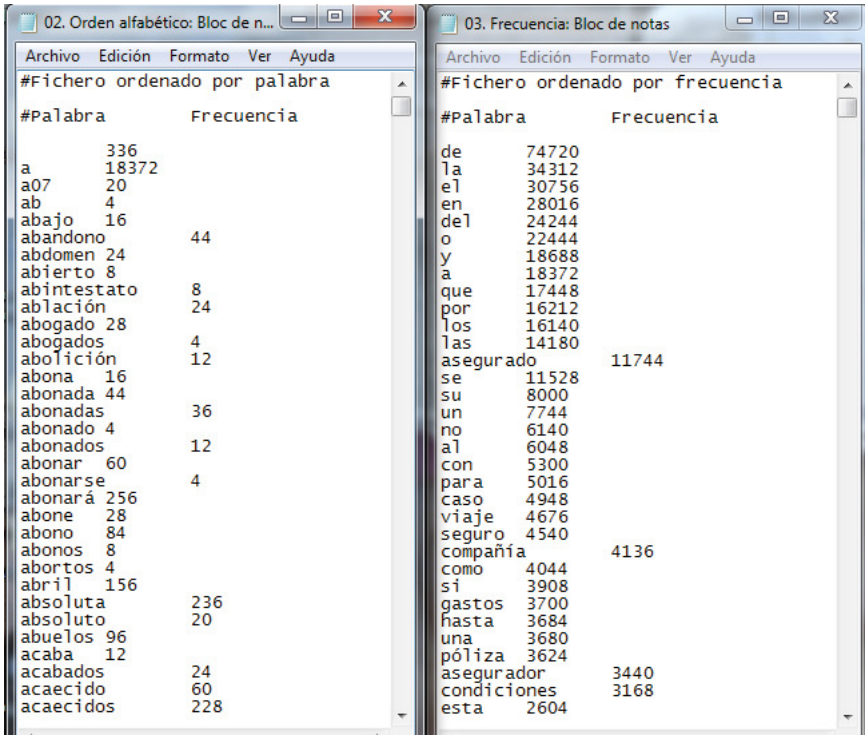


Figure 13. Types ordered by frequency (1-gram).

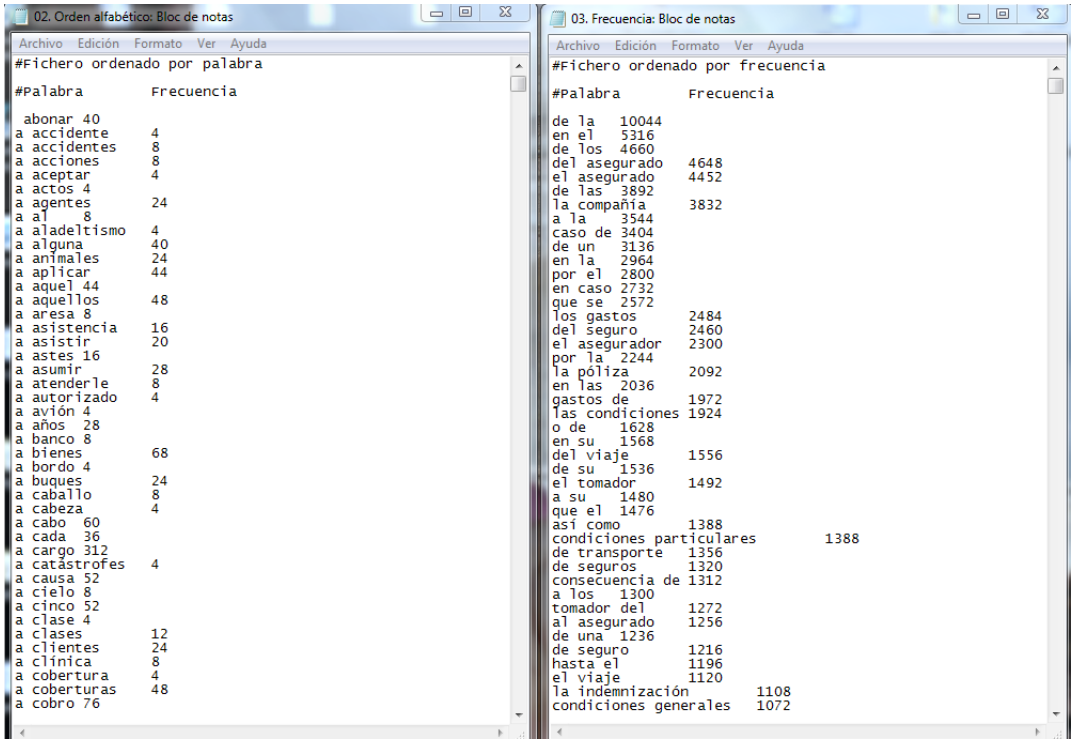


Figure 14. Types ordered by frequency (2-grams).

6. Conclusions

Nowadays it is not possible to determine *a priori* the exact total number of words (tokens) or documents that should be included in specialised ad hoc corpora in order that they may be considered representative. However, in this paper we have described a corpus-driven approach to evaluating corpus size *a posteriori*. In order to achieve this, a double approach to corpus building has been adopted, based on two arguments: firstly, a *qualitative* approach has been followed where a set of clear design criteria and a compilation protocol in four steps are needed in order to ensure corpus representativeness according to quality. Secondly, a *quantitative* approach has been adopted based on the N-Cor algorithm and the ReCor programme. The ReCor programme 2.5. allows to determine that the corpus is of an adequate size of documents and words (tokens) after it has actually been compiled (or even during analysis), i.e. *a posteriori*. As no new types are entering in the corpus, the minimum size has been proved to be reached. This methodology has been illustrated trough the compilation of an ad hoc corpus of travel insurance contracts in Spanish; however, this methodology can be used to compile any ad hoc corpus, in any language and covering any topic and genre.

7. References

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

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

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

Biber, Douglas 1993: Representativeness in Corpus Design. In *Literary and Linguistic Computing* 8 (4), 243-257.

Biber, Douglas 1994: Representativeness in Corpus Design. In Zampolli, Antonio/Calzolari, Nicoletta/Palmer, Martha (eds.), *Current Issues in Computational Linguistics: In Honour of Don Walker*. Dordrech/Pisa: Kluwer & Giardini, 377-408.

Biber, Douglas 1995: *Dimensions of Register Variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, Douglas/Conrad, Susan/Reppen, Randi 1998 (eds): *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, Lynn/Pearson, Jennifer 2002: *Working with Specialized Language: A practical guide to using corpora.* London: Routledge.

Braun, Eliezer. 2005 [1996]: El caos ordena la lingüística. La ley de Zipf. In Braun, Eliécer (ed.), *Caos fractales y cosas raras.* México D. F.: Fondo de Cultura Económica.

Bravo Gozalo, José María/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 Fernández Nistal, Purificación/Bravo Gozalo, José María (eds.), *La traducción: orientaciones lingüísticas y culturales.* Valladolid: Universidad de Valladolid, 205-257.

Carrasco Jiménez, Rafael C. 2003: *La ley de Zipf en la Biblioteca Miguel de Cervantes*. Alicante: Universidad de Alicante.

Clear, Jeremy H 1994: I Can't See the Sense in a Large Corpus. In Ferenc Kiefer, Gábor Kiss/Pajzs, Júlia (eds), *Papers in Computational Lexicography: COMPLEX '94.* Budapest: Research Institute for Linguistics y Hungarian Academy of Sciences, 33-48.

CORIS/CODIS 2006. Progettazione e costruzione di un CORpus di Italiano Scritto. In *CORIS/CODIS*. Bolonia: CILTA.

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

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

Friedbichler, Ingrid/Friedbichler, Michael 2000: The Potential of Domain-Specific Target-Language Corpora for the Translator's Workbench. In Bernardini, Silvia/Zanettin, Federico (eds.), *I corpora nella didattica della traduzione. Corpus Use and Learning to Translate*. Bolonia: CLUEB.

Gelbukh, Alexander/Sidorov, Grigori/Hernández, Liliana Chanona 2002: Corpus virtual, virtual: Un diccionario grande de contextos de palabras españolas compilado a través de Internet. In Gonzalo, Julio/Peñas, Anselmo/Ferrández, Antonio (eds.), *Multilingual Information Access and Natural Language Processing, International Workshop (November 12) at IBERAMIA-2002, VII Iberoamerican Conference on Artificial Intelligence*. Sevilla: IBERAMIA, ELSNET y RITOS-2. 7-14.

Giouli, Voula/Piperidis, Stelios 2002: *Corpora and HLT. Current trends in corpus processing and annotation*. Bulgaria: Institute for Language and Speech Processing.

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 Leitner, Gerhard (ed.), *New dimensions in English language corpora. Methodology, results, software development*. Berlin/New York: Mouton de Gruyter, 3-19.

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

Kock, Josse de 1997: Gramática y corpus: los pronombres demostrativos. In *Revista de filología románica* 14 (1), 291-298.

Kock, Josse de 2001: Divergencias y convergencias en dos corpus de registros de lengua diferentes. In *Verba: Anuario Galego de Filoloxía* 28, 31-58.

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

Leech, Geoffrey 1991: The state of the art incorpus linguistics. In Aijmer, Karin/Altenberg, Bengt. (eds.), *English Corpus Linguistics*. London: Longman, 8-29.

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

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

McEnery, Anthony/Wilson, Andrew 2006 [2000]: ICT4LT Module 3.4. Corpus linguistics [online]. <http://www.ict4lt.org/en/en_mod3-4.htm>.

Murison-Bowie, Simon 1993: *MicroConcord Manual. An introduction to the practices and principles of concordancing in language teaching.* Oxford: Oxford University Press.

Ruiz Antón, Juan C. 2006: *Corpus y otros recursos lingüísticos*. Castellón: UJI.

Sánchez Pérez, Aquilino/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. In *International Journal of Corpus Linguistics* 2 (2), 259-280.

Sánchez Pérez, Aquilino/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. In *Atlantis* XIX(2), 205-223.

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

Seghiri, Miriam 2008: Creating virtual corpora step by step. In *Researching and Teaching specialized languages: New contexts, new challenges. VII Annual Conference of the European Association of Languages for Specific Purposes (AELFE).* Murcia: Universidad de Murcia, 435-449.

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

Sinclair, John M. 2004: Corpus and Text: Basic Principles. In Wynne, Martin (ed.), *Developing Linguistic Corpora: a Guide to Good Practice.* Oxford: Universidad de Oxford.

Tognini-Bonelli, Elena 2001: *Corpus linguitics at work*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Wilkinson, Michael 2005: Compiling a specialized corpus to be used as a translation aid. In *Translation Journal* 9 (3),1-6.

Yang, Dan-Hee/Lee, Ik-Hwan/Cantos Gómez, Pascual 2002: On the Corpus Size Needed for Compiling a Comprehensive Computational Lexicon by Automatic Lexical Acquisition. In *Computers and the Humanities* 36 (2), 171-190.

Yang, Dan-Hee/Cantos Gómez, Pascual/Song, Mansuk 2000: An Algorithm for Predicting the Relationship between Lemmas and Corpus Size. In *ETRI Journal* 22 (2), 20-31.

Yang, Dan-Hee/Lim, Su-Jong/Song, Mansuk 1999: The Estimate of the Corpus Size for Solving Data Sparseness. In *Journal of KISS* 26 (4), 568-583.

Young-Mi Jeong 1995: Statistical Characteristics of Korean Vocabulary and Its Application. In *Lexicographic Study* 5 (6), 134-163.