Virtual corpora as documentation resources: Translating travel insurance documents (English-Spanish)*

Gloria Corpas Pastor and Miriam Seghiri
Universidad de Málaga (Spain)

The inclusion of documentation as a core subject in the curriculum of Translation and Interpretation degrees clearly underlines its importance to translators. Training in this discipline is considered essential for a translator given that only sufficient and conscientious work on documentation will allow an adequate translation of a specialised text. The sources of information that may be utilised by the translator are extremely varied, ranging from an oral consultation with an expert to a search using specialised glossaries and dictionaries. However, in the field of translation perhaps the most relevant documentation activity today involves the use of the Internet and, closely related to this, the compilation and management of virtual corpora.

In this chapter, we present a systematic methodology for corpus compilation based on electronic resources available on the Internet. The methodology is illustrated through the creation of a virtual corpus of travel insurance in English and Spanish, whose representativeness is subsequently determined by using a computer programme-called *ReCor* specifically designed for this purpose. Finally, some specific examples of possible uses in direct and inverse translations of this type of document are given.

Key words: Corpus compilation and representativeness, specialized corpora, legal translation.

* The research reported in this paper has been carried out in the framework of the 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).

1. Introduction

Since the tourist industry is one of the principle driving forces behind the Spanish economy,[1] it is hardly surprising that there is a large demand for translations of insurance policies in the tourism sector both from Spanish into English and from English into Spanish (cf. ACT 2005). Although this economic reality could be transitory, the rights of European consumers to demand translations of this type of document under the auspices of European directives[2] on insurance matters and their respective national transpositions[3] should also be taken into account. These directives recognise the right of the party taking out insurance to receive a contract[4] 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. Subsequent directives, such as 2002/92/CE,[5] have also increased demand for translations of all the formal documents that constitute the contract. In the following pages, we shall

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

2. We refer to the *Third EC Directive on Non-Life Insurance (92/49/EEC)* and the *Third EC Directive on Life Assurance (92/96/EEC)*.

3. These transpositions, which are primarily aimed at consumer protection and fostering linguistic plurality in Europe, are given expression, in the case of Spain, in the *Ley 18/1997, de 13 de mayo, de modificaciones del artículo 8 de la Ley de Contrato de Seguro, para garantizar la plena utilización de todas las lenguas oficiales en la redacción de los contratos*, (BOE, 14th May 1997); in the case of the United Kingdom, in *Statutory Instrument 2004, n.º 353. Insurers (Reorganisation and Winding Up) Regulations 2004*; and, finally, in the case of the Republic of Ireland, in the *Insurance Act 2000*.

4. The *policy* (*póliza,* in Spanish) is the document which gives physical form to the insurance contract. In addition, it is where the obligations and rights of both the insurer and the insured person are set out, where the persons or objects that are insured are defined and the guarantees and compensation in the case of damage are established. It also represents the formalisation and culmination of the whole process of contracting the insurance. As a result, in many cases the insurance policy may be referred to as the *contrato* (*contract*) (cf. *Ley 50/1980*; *Insurance Act 2000*; *The Financial Services* and *Markets Act 2000*).

5. We refer specifically to *Directive 2002/92/EC of the European Parliament and of the Council of 9 December 2002 on insurance mediation*. In Article 13 of this directive, under "Information conditions", it is specified that "All information to be provided to customers in accordance with Article 12 shall be communicated: (a) on paper or on any other durable medium available and accessible to the customer; (b) in a clear and accurate manner, comprehensible to the customer;

present a systematic methodology for the creation of a virtual corpus of travel insurance in English and Spanish based on electronic resources available on the Internet. The representativeness of this corpus will subsequently be determined by using a computer programme specifically designed for this purpose.

2. Corpora in translation training

The advantages of using corpora in translation have been shown by various studies (cf. Laviosa 1998; Bowker 2002; Bowker and Pearson 2002; Zanettin et al. 2003, amongst others). Some of the principal advantages of using them are their objectivity, their reusability and multiple usage of a single resource. In addition, they are user-friendly and allow access to and management of huge quantities of information in almost no time. Furthermore, we must consider that the development of our current information society has brought about a demand that did not exist previously for texts written in a variety of languages. Together with economic globalisation, this has resulted in a growing interest[6] in the use of bilingual and multilingual corpora by researchers working in the fields of automatic and assisted translation, language teaching, terminology and specialised language, natural language processing and information recovery as well as, more recently, in training and documentation as applied to translation.

On this last subject, despite the remit of the European project *LETRAC*[7] (*Language Engineering for Translators Curricula*), the use of corpora has only really come to the attention of researchers working in the field of translation training relatively recently. Examples of studies that stand out are: Kenny (2001) on the subject of literary translation based on parallel corpora in German and English;

(c) in an official language of the Member State of the commitment or in any other language agreed by the parties."

6. There has been such a flood of compilers in Europe that we are forced to list only some of the more important examples: *ACL* (*Association for Computational Linguistics*); *ECI* (*European Corpus Initiative*); *LDC* (*Linguistic Data Consortium*); *ICAME* (*International Computer Archive of Modern and Medieval English*); *ACL/DCI* (*Association for Computational Linguistics Data Collection Initiative*) and *ELRA* (*European Language Resources Association*).

7. See <http://www.iai.uni-sb.de/docs/D3.pdf>. In their final report, which was presented to the European Commission DG XII, the LETRAC project stressed the importance of introducing the following elements to the curriculum of translation degrees: applied IT, terminology management programmes, CAT and AT systems, ICTs and linguistic engineering as well as leaving time for publishing programmes, the Internet, controlled languages, project management, translation memories and corpus linguistics.

Corpas Pastor (2001, 2003b, 2004a, b and c) on legal and medical translations based on multilingual corpora compiled from the Internet; and Sánchez-Gijón (2003a: NP) on the subject of virtual *ad hoc* corpora for scientific translations in the English-Spanish language pair. Other examples of studies are: Bernardini and Zanettin (2000); Bowker and Pearson (2002); Zanettin, Bernardini and Stewart (2003) on the possibilities offered by corpora for specialised language teaching. Two studies that deal with the potential use of corpora in language teaching, natural language processing and translation are Aston (2001) and Granger and Petch-Tyson (2003). Finally, in the R&D project described in Corpas Pastor (2003*a*) the corpus was used as a fundamental documentation resource for the translation of legal texts – this new venue of research was further developed some years later by Seghiri (2006).

Both researchers and teachers are in agreement over the importance of corpora in translation training and practice. Some authors have gone even further and specifically indicate *virtual* corpora (cf. Pearson 1998; Bernardini and Zanettin 2000; Corpas Pastor 2001 and 2004a; Zanettin 2002a and b; Sánchez-Gijón 2003a and b) as one of the translator's most important aids when faced with a specialised text. By virtual corpus we refer to a corpus compiled from electronic sources exclusively in order to carry out a specific translation in any direction (direct, inverse or indirect[8]). Its principal objective is to construct a reliable resource quickly and at minimal cost, based on texts mined from the Internet, to satisfy the translator's documentation needs.

Virtual corpora may also be referred to as *ad hoc* (Corpas Pastor 2001: 164; Sánchez-Gijón 2003a: 3), *disposable* (Zanettin 2002a), *do-it-yourself/DIY* (Zanettin 2002a), *domain-specific* (Corpas Pastor 2004a: 226), *web* (Fletcher 2004), *electronic* (Corpas Pastor 2001; Varantola 2003), *ephemeral* (Corpas Pastor 2004a: 226), *precision* (Varantola 1997); and *special purpose* (Jennifer Pearson 1998; Sánchez-Gijón 2003a).

Translators turn to the Internet in search of solutions to information and documentation problems because they are not only translating between languages (for which a good dictionary, whether online or not, would suffice), but also between discourse communities or cultures. In this context, the compilation of corpora and the Internet appear to be two of the most important documentation resources in the practice and research of specialised translation. When facing this

8. A "direct translation" is translation done directly from the original into translator's native language, without an intermediary text; an "inverse translation", also called "other tongue translation (OTT)", is a translation from the translator's native language into another language; finally, an "indirect translation", also denominated "mediated translation", is a translation done via an intermediary translation in a third language, not directly from the original.

kind of assignment, the main problem that translators come up against is that a corpus for the particular speciality is not available for consultation on the Internet or, if one already exists, it often does not cover all the information requirements of the source text. In other words, "one problem with these typically small and domain specific corpora is the limited range of topics and text types for which they are available" (Zanettin 2002a: NP). Faced with this situation, translators have no alternative other than to compile their own virtual corpora for the specific translation that has been commissioned in each case.

It is also important to take into account that any set of texts does not, in and of itself, constitute a corpus. In order for a collection of texts to be considered a corpus in the strict sense of the term, it must meet a set of clear *design criteria* and abide by a specific *compilation protocol* so that the collection may be deemed representative of the field of specialisation or the particular type of document that is being translated.

3. Guidelines for corpus creation

In this section we will outline the design parameters that the creation of a virtual corpus demands. Following this we will propose a compilation protocol in the form of guidelines. This consists of four distinct phases: (1) locating and accessing resources, (2) downloading data (3) text formatting and (4) data storage.

3.1 Design criteria

Before moving on to deal specifically with how the documentation resources necessary to create a virtual corpus are located, it is essential for the translator-compiler to first of all establish a set of clear design criteria. In this case, the objective is to create a corpus of travel insurance policies in Spanish and English compiled exclusively from tourism law resources available on the Internet. This bilingual corpus must be diatopically restricted due to the large number of countries in which both English and Spanish are official languages. In order to illustrate the methodology put forward, the corpus will be restricted to legislation in force (whether it be communitary, national or from autonomous authorities) and to the formal elements of the contract (principally insurance quotes, proposal forms, certificates of insurance and insurance policies[9]) that have been drawn

9. Another document is the *duplicado de la póliza* (a duplicate of the policy), which is drawn up in writing by the insurer if requested by the person who takes out the insurance, the insured

up in Spain, the Republic of Ireland and the United Kingdom (Scotland, Wales, England and Northern Ireland). In addition, it will be necessary to compile a comparable corpus, made up of two subcorpora, one in Spanish and the other in English, which will include the original texts of the tourism contracts. This will be a textual corpus, i.e. a full-text corpus, since it will include complete texts, and a specialised corpus, in the sense that it includes specific text types dealing with communication between specialists and semi-specialists or laymen.

A travel insurance corpus compiled in accordance with these design criteria will be essentially *unbalanced,*[10] since quality takes priority over quantity (Corpas Pastor 2004a: 236) in this type of virtual corpus which has been compiled *ad hoc*. It is, however, extremely *homogenous* given that it has been created for a specific purpose.

3.2 Compilation protocol

Once the preliminary design parameters have been established the translator-compiler should follow a protocol for the creation of the corpus comprising four stages which will now be described.

3.2.1 *Locating and accessing resources*

The first stage of the protocol consists of locating and accessing information available on the Internet. In order to do this the translator-compiler will have to develop and/or put his/her knowledge of electronic resources into practice.

Once the type of electronic corpus has been designed the question of access to the relevant documents arises. Various possibilities exist for accessing these texts. According to Austermühl (2001: 52 et seq.), there are basically three types of searches that may be carried out on the Internet: *institutional searches*, carried out on the web sites of international organisations and institutions; *thematic searches*, normally carried out using directories and, lastly, *key word searches* using a search engine.

person or the beneficiary. The insurer is obliged to provide a duplicate or copy of the policy if the original is mislaid, the copy must be identical and have the same validity as the original. In addition, there is also a document known as the *boletín de adhesión* (a joining form), a document which gives proof of the insurance and has not been included here because it only applies to life insurance policies.

10. *Unlabaced* because of the distribution of languages on the Internet. According to the "Top Ten Languages Used in the Web (November 2007)" published by Internet World Stats (http://www.internetworldstats.com/stats7.htm), the Spanish language represents 9.0 % of all the Internet users in the world, while English represents 30.1 %.

We shall begin with an *institutional search,*[11] one of the most productive types of search for constructing corpora. This is due not only to the great quantity of documents that these types of institutions, organisations or associations store on the Internet today, but also because they can be assumed to be of a high standard in terms of both quality and reliability because the writers are specialists in the field. This institutional search will be mainly, though not exclusively, carried out from institutional, regulatory and legislative sources. In order to locate legislation the web sites and web pages that follow may be used.

In terms of official organisms and institutions, legislative information can be taken from the headquarters of the *ABI* (*Association of British Insurers*),[12] the *ABTA* (*Association of British Travel Agents*)[13] or the *FSA* (*Financial Services Authority*)[14] for the United Kingdom and Ireland. For Spain, information can be mined from the *Mesa del Turismo,*[15] particularly the section called "legislación general" which includes regulatory laws and laws specifically related to the tourism sector.

Another outstanding web site is that of the *WTO* (*World Tourism Organisation*)[16] which contains one of the principal documentation resources for legislative material, *Lextour.*[17] This is the *WTO*'s database of tourism legislation which has links to web sites, databases, and external servers concerned with tourism legislation set up by parliaments, governmental organisations, universities and professional associations. We have also taken information from other databases to obtain communitary legislation, such as the well respected *Westlaw.*[18] However,

11. On numerous occasions, it may be necessary to perform a key word search to find the names of more organisations to be used in the institutional search. This can usually be performed by introducing descriptors together with Boolean techniques in a search engine such as *Google*. For example, introducing *organismo* OR *turismo, organismo* AND *turismo* OR "*organismo turístico*" will increase the number of names of organisations connected with tourism, whose web sites can then be visited in order to extract information that may be suitable for inclusion in the travel insurance corpus.

12. Available at <http://www.abi.org.uk>.

13. Available at <http://www.abta.com>.

14. Available at <http://www.fsa.gov.uk/consumer>.

15. Available at <http://www.mesadelturismo.com>.

16. Available at <http://www.world-tourism.org>.

17. Available at <http://www.world-tourism.org/doc/S/lextour.htm>.

18. Available at <http://web2.westlaw.com/signon/default.wl?bhcp=1>.

our most significant source has been *EUR-Lex,*[19] the portal to European Union law, which is currently the best database for European Union law.

Practically all the documents involved in the process of making a contract for travel insurance may be found on the web sites of the big insurance companies. In addition, although less frequently, the web sites of numerous online travel agencies contain the texts of their policies, which they sell on from various insurance companies, for their customers' information. Similar rich sources of information are also the web sites of international insurance companies such as *Mondial Assistance*[20] or *Europ Assistance,*[21] British and Irish insurance companies such as *AT Bell Insurance Brokers Ltd,*[22] *Royal and Sun Alliance*[23] or *Lloyds of London;*[24] or Spanish insurance companies, such as *Allianz,*[25] *MAPFRE*[26] or *Ocaso,*[27] to mention only a few of the most representative examples.

The next step is to move on to making *thematic searches*[28] using well known *directories*. In this case, a problem with locating information may arise as a result of the structure of the directories themselves which can even hinder the process of documentation extraction.

Specialist directories stand out as excellent resources for locating communitary, national and autonomous legislation, especially when the resources they contain are also evaluated and commented upon. This is the case for the compilation of the Spanish subcorpus, using the section called "Dret" in the "Indices" of

19. Available at <http://eur-lex.europa.eu>.

20. Available at <http://www.mondial-assistance.com/en/aboutus/homepage.htm>.

21. Available at <http://www.europ-assistance.es/>.

22. Available at <http://www.atbell.co.uk>.

23. Available at <http://www.royalsunalliance.com/royalsun>.

24. Available at <http://www.lloyds.com>.

25. Available at <http://www.allianz.es>.

26. Available at <http://www.mapfre.com/pmapfre/es/index.html>.

27. Available at <http://www.ocaso.es>.

28. As with the institutional search, the thematic search may be complemented by a key word search if it is necessary to augment the names of thematic directories connected to the particular specialisation that is being searched. For example, to locate legal directories we would normally go to *Google* and by using descriptors combined with Boolean operators introduce productive search equations such as *"directorio jurídico"* or *directorio AND jurídico*.

the *Universitat de Barcelona*[29] and the *Universitat Autònoma de Barcelona.*[30] The directories of *The Argus Clearinghouse*[31] and *Search the Law*[32] (particularly the section "Travel") are similarly useful for the English subcorpus.

In general, thematic searches based on indices or directories are the most productive for extracting legislation rather than insurance contracts. In order to do this it is necessary to take a further step and carry out a *key word search*. For this type of search a generic search engine such as *Google* may be used. According to a great number of analysts *Google* is the best search engine in terms of the quality of search results (cf. Radev et al. 2005: 580).

Alongside visits to insurance companies' web sites, key word searches have proved to be (cf. Seghiri 2006) the easiest and quickest way to recover the documents that make up insurance contracts. The best results will be obtained from search engines if knowledge of the facilities they offer is utilised. As well as defining the search appropriately, techniques such as using Boolean operators, truncation and phrase searches should be considered. On this point, it is clearly essential to establish descriptors. A practical example (cf. Tables 1 and 2[33]) is given to illustrate how searches are made to locate the texts that will comprise the corpus. In order to do this, the text types and the field of insurance in which the desired information is to be found (travel insurance) are taken as descriptors and Boolean search techniques are applied using the user friendly interface offered by, for instance, *Google*'s advanced search.[34]

29. Available at <http://www.bib.ub.es/bub/internet.htm>.

30. Available at <http://www.bib.uab.es/internet.htm>.

31. Available at <http://www.clearinghouse.net>.

32. Available at <http://www.search-the-law.com>.

33. In this table only the descriptors that have produced the greatest number of documents for the text type we required in the two specific languages (English and Spanish) are shown. However, it should be pointed out that in reality a vast number of search criteria were used and here we have only shown a sample by way of illustration.

34. In order to mine the Spanish contractual documents, the version of *Google* for Spain (<http://www.google.es>) was used. By selecting the option "páginas de España" it is possible to filter out any documents that come from other Spanish speaking countries. The same procedure may be followed to search for information in English, i.e. the user goes to the version of *Google* for the United Kingdom (<http://www.google.co.uk>) and for Ireland (<http://www.google.ie>) and selects the options "pages from the UK" and "pages from Ireland" respectively in order to avoid the presence of documents that come from other countries. Occasionally, however, this filtering will not be sufficient so that, in addition to searching by country, it may be necessary in cases of doubt as to the origin of a document located by using *Google*, to refer to the domain in order to verify their source. The knowledge that the domains *.es* for Spain, *.uk*

Table 1. Descriptors for the finding of the formal elements of travel insurance contracts (Spanish).

Text type	Descriptors	Search equation
Póliza	Póliza, seguro turístico, asistencia en viaje[35]	póliza AND "seguro turístico" póliza AND "asistencia en viaje"
Solicitud	Solicitud de póliza, seguro turístico, asistencia en viaje	solicitud AND póliza AND "seguro turístico" Solicitud AND póliza AND "asistencia en viaje"
Propuesta	Propuesta, proposición, seguro turístico, asistencia en viaje	póliza AND propuesta OR proposición "seguro turístico" póliza AND propuesta OR proposición "asistencia en viajes"
Carta de Garantía	Carta de garantía, seguro turístico, asistencia en viaje	"carta de garantía" AND "asistencia en viaje" "carta de garantía" AND "seguro turístico"

Table 2. Descriptors for the finding of the formal elements of travel insurance contracts (English)

Text type	Descriptors	Search equation
Policy	Policy, travel insurance	policy AND "travel insurance"
Quote	Quote, travel insurance	Quote AND policy AND "travel insurance"
Proposal Form	Proposal Form, travel insurance	"proposal form" AND policy AND "travel insurance"
Certificate of Insurance	Certificate of Insurance, Insurance Certificate, travel insurance	"certificate of insurance OR "insurance certificate" AND policy

for the United Kingdom and *.ie* for Ireland will therefore be of use. In addition pages in Spanish with the domain *.ar* indicating Argentina, or *.mx* indicating Mexico and pages in English with the domain *.au* indicating Australia or *.us* indicating the United States will be automatically ruled out because they are not appropriate for our corpus.

35. We refer mainly to seguro turístico or travel insurance in accordance with the position taken by Aurioles (cf. Aurioles Martín (2005 [2002]) y and Aurioles Martín et al. (2004) because we believe it to more accurate than the Spanish calque, *asistencia en viaje* of the original English, since travel assistance is only one possible part of travel insurance which may also include coverage for holiday cancellation or medical attention, to cite only some of the most common examples. For a wider perspective on this question see the trilingual (Spanish-English-Italian) classification of travel insurance policies in relation to coverage outlined by Seghiri (2006: 279–281).

The main difficulty with key word searches centres on the choice of the most precise descriptors for the intended search, given that without this a large amount of irrelevant information will be returned. It is up to the translator-compiler to filter out all this "noise" from each of the pages that will be included in the corpus.

3.2.2 *Downloading data*

When the documents have been located and accessed, the next stage is to download the data. Usually, this stage is performed manually, although occasionally it is possible to automate the task when dealing with a group of web pages which have been accessed using the programme *GNU Wget,*[36] which allows downloading in batches.

This downloading phase may be hampered by the inherent structure of the Internet itself. On the one hand, we are faced with a mark-up language or HTML, in other words, the information is organised in hypertext nodes which are often difficult to access. This is usually as a result of the content being inappropriately labelled or because the location of the information is difficult to see on the page. On the other hand, the wide variety of formats that the information may appear in should also now be considered.

3.2.3 *Text formatting*

In the cases of both legislation and contracts related to travel insurance a noticeable predilection for HTML (.html) and PDF (.pdf) exists. The first of these does not involve many problems in terms of conversion since the information may simply be copied and pasted into a text document. *Google* will also allow the majority of PDF documents to be seen in .html format, thereby permitting the same procedure to be carried out. When this is not possible, conversion programmes such as *Solid Converter*[37] may be used. Hence, this third stage of downloading is completed by what might be called *normalisation*, since all the documents will be converted to an ASCII or plain text format. In other words, they are stripped

36. This free software together with its instruction manual may be downloaded from the following web site: <http://www.gnu.org/software/wget/>.

37. A trial version of *Solid Converter* may be downloaded free of charge from <http://www.solidpdf.com>. Given that it is a free trial version, it has a number of limitations: it only functions for a two week period and permits conversion of a maximum of ten pages per document, although it is possible to convert a complete text over a number of operations by specifying a different set of pages each time. There are other free programs available online like *Pdf to Word converter 3.0* (<http://www.geomundos.com/descargas/bajar-pdf-to-word-converter-30_233.html>), *PDF Converter* (<http://www.freepdfconvert.com/convert_pdf_to_source.asp>) or *Easy PDF to Word Converter* (<http://www.pdf-to-html-word.com/ >), for instance.

of the HTML or code of any other kind, in accordance with the *clean-text policy* described by Sinclair (1991: 21).

3.2.4 *Data storage*

The last stage is to store the data. This consists of storing the documents that have been downloaded and correctly identifying and arranging them. One possible way of doing this is through the use of sub-files depending on whether the documents are in their original format or in ASCII format. These sub-files are then subdivided according to the language, text types and text formats of the corpus.

In this study, we have extracted two subcorpora from the multi-lingual *Turicor* corpus of travel and tourism law, which is described and fully documented at the website *http://turicor.com*. The two subcorpora are a bilingual comparable corpus which consists of a Spanish subcorpus with 259 texts[38] (1,837,869 words) and an English subcorpus with 302 documents (3,202,118 words).

4. Determining corpus representativeness

Despite repeated reference by the experts 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 on this crucial issue.

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 (cf. Lavid 2005).

38. On the subject of the legislative documents that form part of the corpus (17 texts in English and 2 texts in Spanish) it is important to point out that travel insurance is not regulated by substantive legislation. Instead it comes under the regulations that apply to all insurance other than life insurance through various communitary directives such as *73/239/EEC*, *73/240/EEC*, *76/580/EEC*, *78/473/ EEC*, *84/641/ EEC*, *87/343/ EEC*, *87/344/ EEC*, *88/357/EEC*, *90/618/EEC*, *92/49/EEC*, *95/26/EEC*, *2000/26/EC*, *2000/64/EC* and *2002/13/EC*. In Spain, travel insurance contracts are also currently regulated by the *Ley 50/1980, de 8 de octubre, de Contrato de Seguro*, [*Act 50/1980, 8th October, Insurance Contracts*] as well as the *Ley 30/1995, de 8 de noviembre, de ordenación y supervisión de los Seguros Privados* [*Act 30/1995, 8th November, Planning and Supervision of Private Insurance*]. In Ireland, insurance contracts are regulated by the *Insurance Act, 2000,* as well as the *European Communities (Non-Life Insurance) Framework Regulations, 1994* (*S.I. No. 359 of 1994*). In the United Kingdom, they are regulated by the *Financial Services and Markets Act 2000* (*Statutory Instrument 2003 N.º 1476*), specifically *Amendment, Nº. 2, Order 2003*. In relation to policies, the central document in this type of agreement, it was possible to include 101 documents (1,000,067 words) in the Spanish policies component and 176 documents (1,903,661 words) in the policies component in English. The remainder of the formal elements of the contract are included in the rest of the corpus.

However, even today 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.[39] As Biber, who is one of the most prolific writers on the 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" (cf. Biber et al. 1998: 246–247).

It is therefore commonplace to come up against questions over the minimum number of texts needed to guarantee that a sample is scientifically valid, as well as debates over how to specify a sufficient number of texts and number of words for a corpus (Sanahuja and Silva 2001).

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),[40] Young-Mi (1995) and Sánchez Pérez and Cantos Gómez (1997). However, subsequently, some of these authors, such as Cantos (Yang et al. 2000: 21), recognised some shortcomings in these works, suggesting that they might be attributed to the use of Zipf's law.[41] Zipf's law[42] can give us an idea of the breadth of vocabulary used, but it is not limited to a particular or approximate number because this will depend on how the constant is determined (Braun 2005 [1996] and Carrasco Jiménez 2003: 3).

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

40. 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).

41. 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).

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

Numerous studies have been based on the law, but the conclusions they reach do not specify, not 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).

A possible solution could be to analyse the lexical density of a corpus 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 that can represent lexical density as the corpus increases on a document by document basis: once a certain number of texts have been included, the number of types does not increase in proportion to the number of words the corpus contains.

This formula may make it possible to determine the minimum size that a corpus must reach for it to begin to be representative. With the help of graphs, it should be possible to establish whether the corpus is representative and approximately how many documents are necessary to achieve this. This theory has become a practical reality in the shape of a software application, *ReCor,*[43] which enables accurate evaluation of corpus representativeness.

It should be made clear that the method for evaluating the homogeneity of a very specialised corpus assumes that the target population is known and available to the researcher. This clearly involves careful design of the corpus in terms of components, text types to be included, diasystematic limits (diaphasic, diastratic, diachronic and diatopic), as well as type of corpus (comparable, parallel, etc.), number and status of languages, text documentation for DTDs and headers, *inter alia*.

Once the question of quality is ensured in terms of corpus design and document selection, this programme can be used to determine *a posteriori* whether the size reached by a given corpus is sufficiently representative of this particular sector of the tourist industry. For further information, the technology and the theoretical presuppositions behind the ReCor Programme are explained in detail in Seghiri (2006), Corpas Pastor and Seghiri (2006a, 2006b, 2007a, 2007b and forthcoming).

4.1 The *ReCor* interface

ReCor's interface is simple, intuitive and user-friendly (see Figure 1). Firstly, an input file may be selected; this could be anything from a particular clause in a policy

43. *ReCor* is an acronym derived from the function it was designed for: the representativeness of corpora.

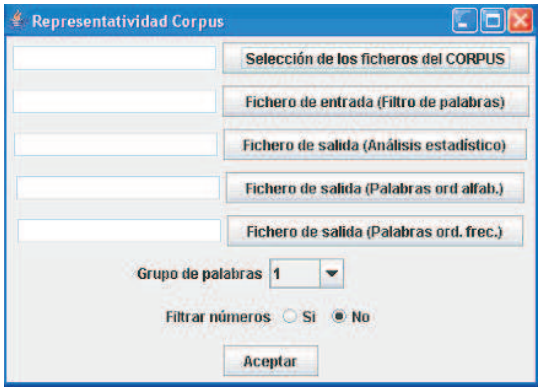

Figure 1. The ReCor interface

to the entire corpus. There is also an option: "*Filtro de entrada*", 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". Next, three output files are created. The first, "*Análisis estadístico*" or 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, "*Palabras ord. alfa.*", 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, "*Palabras ord. frec.*", but this time the words are ordered according to their frequency, or in other words, by their rank. The application also allows the user to work with groups of up to ten words (n-grams)[44] and phraseology, as well as allowing numbers to be filtered out.

44. In this study we used the 2.1 version of *ReCor*. We are currently working on a new version (ReCor 3.0) which has an improved capacity for working with multiple and very large files quickly and also allows phraseological units to be identified on the basis of analysis of n-grams ($n \geq 1$ and $n \leq 10$) of the corpus.

4.2 Graphical representation of data

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.[45]

In the first presentation of the corpus generated by the programme in graph form – *Estudio gráfico A* – the number of files selected is shown on the horizontal axis, while the vertical axis shows the type/token ratio. The results of two different operations are shown, one with the files ordered alphabetically (the red line), and the other with the files introduced at random (the blue line). In this way the programme double-checks to verify that the order in which the texts are introduced does not have repercussions on the representativeness of the corpus. Both operations show an exponential decrease as the number of texts selected increases. However, at the point where both the red and blue lines stabilise, it is possible to state that the corpus is representative, and at precisely this point it is possible to see approximately how many texts will produce this result.

At the same time another graph is generated – *Estudio gráfico B* – 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 travel insurance documents that have been assembled in the two languages involved – English and Spanish – is sufficient to enable us to affirm that our corpus is representative. See Figures 2 and 3 below which show the representativeness of the two languages involved.

The results generated by *ReCor* allow us to conclude that the Spanish subcorpus of travel insurance (cf. Figure 2) can be considered representative from 140 documents and 1 million words onwards, whereas the English subcorpus needs almost double the number of documents (275) and words (2.5 million) in order to reach representativeness (cf. Figure 3). The results remain largely the same even when the analysis is performed on a two-word basis (2-grams). In other words, the English subcorpus of travel insurance (cf. Figure 5) must contain twice the total number of documents and tokens that are necessary for the Spanish subcorpus to be deemed representative (cf. Figure 4).

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

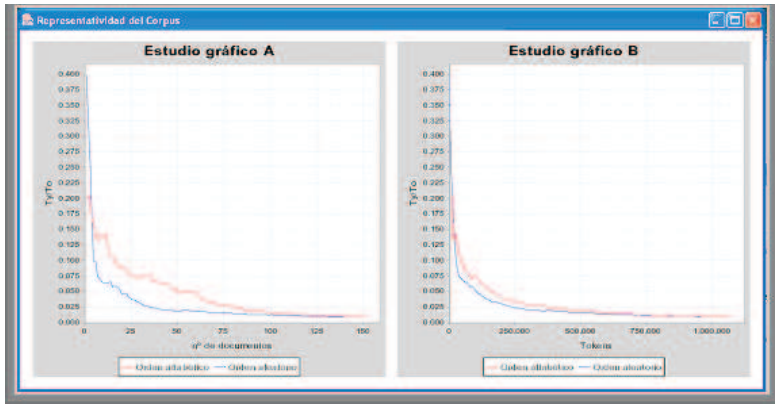

Figure 2. Representativeness of the Spanish travel insurance subcorpus (1-gram)

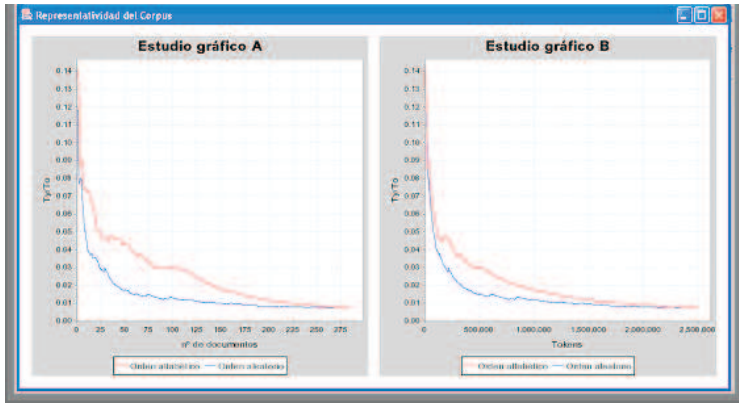

Figure 3. Representativeness of the English travel insurance subcorpus (1-gram)

Furthermore, the quantitative data produced by *ReCor* permits us to conclude that, despite the absence of substantive legislation on insurance in the tourism industry in either of the legal systems involved, Spanish travel insurance documents tend to be more homogenous than the English text forms. In other words, it is possible to infer that that the Spanish documents present super-, macro- and microstructures that are very similar to each other in addition to using a narrower terminological range.

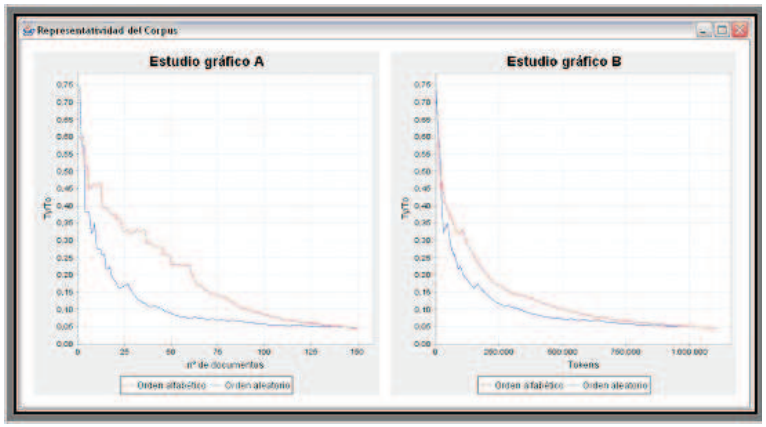

Figure 4. Representativeness of the Spanish travel insurance subcorpus (2-grams)

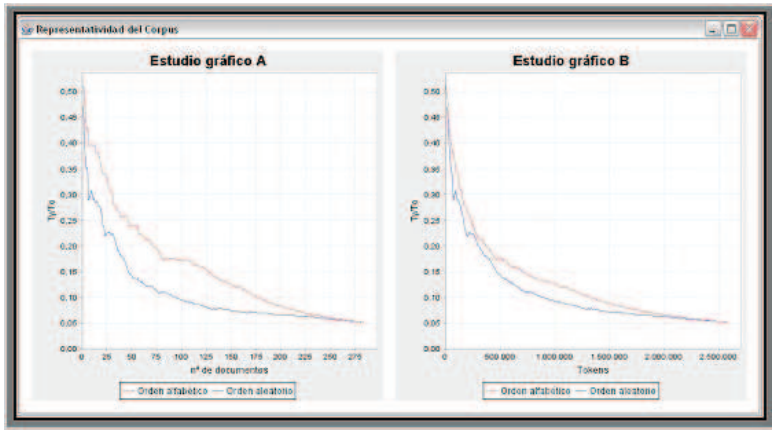

Figure 5. Representativeness of the English travel insurance subcorpus (2-grams)

5. Using the corpus to translate

A well-constructed virtual corpus facilitates diverse studies on translation as both product and process. Furthermore, one of the most promising uses of corpora is in translation teaching and learning to translate. Representative virtual corpora

provide translators (trainers, trainees and professionals) with a first-rate documentation resource for rendering source texts (STs) into the target language.

In addition, the compilation of a virtual corpus calls for a thorough understanding of electronic resources, search skills and data mining techniques from the Internet, thereby promoting the development of the translator-compiler's heuristic sub-competence. Moreover, when a corpus has been appropriately designed and implemented, we can assume that the compiler has carried out a preliminary evaluation of information resources, in order to ensure the overall quality of the textual collection. Evaluation and selection of the documents to be included in a given corpus will usually speed up the translation and/or revision process. As a result, translators can devote extra time to decision-making and problem-solving and focus on these more demanding tasks, instead of repeatedly reviewing the reference material. Hence, using corpora as an aid may also enhance potential users' overall competence as translators.

5.1 Source text samples

Comparable corpora are particularly useful for meeting translators' information needs. In the following subsections we will illustrate the value of corpora for finding information on terminology, phraseology, concepts and discourse for direct and inverse translation of an extract from a travel insurance policy. In order to do this, we have selected two extracts from travel insurance policies, one in English and the other in Spanish as source text (ST) samples.

Extract 1 (ST):[46]

> Important
>
> This is your travel insurance policy. It contains details of cover, conditions and exclusions relating to each insured person and is the basis on which all claims will be settled.

46. The extract comes from a travel insurance policy from the British insurance company *Direct Travel Insurance*: <http://www.direct-travel.co.uk/FAQ/Wordings/policywording010506.pdf>.

Extract 2 (ST):[47]

> CONDICIONES GENERALES
>
> Artículo Preliminar.-El Contrato de Seguro.-El presente Contrato de Seguro se rige por lo dispuesto en la Ley 50/1980, de 8 de octubre, de Contrato de seguro, en la Ley 30/1995, de 8 de Noviembre, de Ordenación y Supervisión de los Seguros Privados.

5.2 Documentation needs

Even two short ST fragments like those chosen in 5.1 offer abundant evidence to argue in favour of the use of comparable corpora in the actual translation process. We are mainly concerned with the terminological and phraseological needs of translators, the extraction of conceptual or domain information, and the comparison of textual and discourse features in the source and target languages.

5.2.1 *Terminology and Phraseology*

The first problem that a translator may come up against is how to translate the term *travel insurance policy* (cf. Extract 1). On this point it should be noted that the term *seguro turístico* has a long tradition in our legal system since the publication in 1964 of the *Spanish Presidential Decree 3304/64 on insurance contracts for foreign tourists*. However, this all changed when the text of the *Council Directive 84/641/EEC of 10 December 1984 amending, particularly as regards tourist assistance, the First Directive (73/239/EEC) on the co-ordination of laws, regulations and administrative provisions relating to the taking-up and pursuit of the business of direct insurance other than life assurance* was transposed to the Spanish legal system through the *Ministerial Order of 27 January 1988 which describes coverage of assistance while travelling as part of private insurance*. This ministerial order employed the term *travel assistance* which was translated into Spanish with the officially accepted neological calque *asistencia en viaje*. Since then, this neological calque from international/Euro English has been incorporated into the Spanish legal system and has supplanted the original *seguro turístico*, which is much more correct given that travel assistance is only one possible part of travel insurance coverage. Other aspects which may be covered include coverage for

47. The extract comes from a travel insurance policy from *Agrupación Astes, Seguro Turístico* published on the web site of the travel agents, *Condor Vacaciones S.A*: <http://www.special-tours.com/ficheros/Seguro_Europa_ES.pdf>.

cancellation of the holiday or medical assistance, to mention only some of the most frequent.

The Spanish corpus also contains two synonyms for the term *travel insurance*: *seguro turístico* and *seguro de asistencia en viaje*, although the frequency with which they appear varies.

As may be seen, *seguro turístico* (cf. Figure 6) produces only 15 concordances,[48] as compared with 26 for *seguro de asistencia en viaje* (cf. Figure 7). It should be pointed out that *asistencia en viaje* appears 107 times. This clearly demonstrates the preference in Spanish for the English calque when drawing up this type of document as well as the influence of English as the lingua franca par excellence (often referred to as "international legal English") and its impact on legislation in the field of travel insurance in peninsular Spanish.

Similar problems arise for translators when faced with translating *El Contrato de Seguro* (cf. Extract 2) into English as there appears to be two possibilities: *assurance contract* or *insurance contract*. A search for *contract* in the corpus reveals a preference in English for *contract of insurance* (cf. Figure 8). In addition, when it appears in this particular position in the text, a fixed expression (*This is your contract of insurance*) can be identified which should be reproduced in translation.

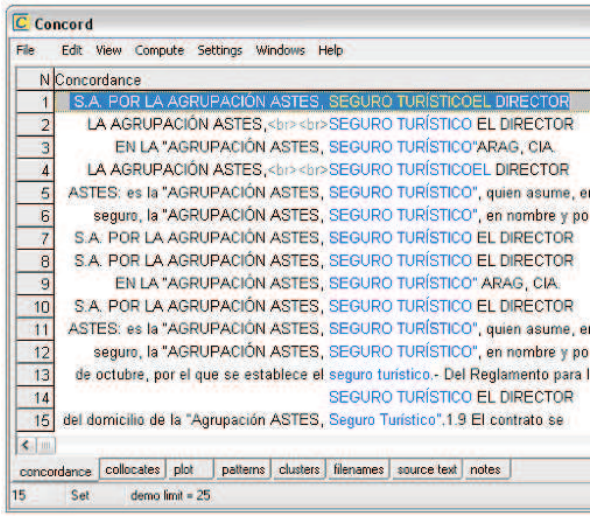

Figure 6. Concordances for 'seguro turístico'

48. The analysis of concordances was carried out using WordSmith Tools 4.0.

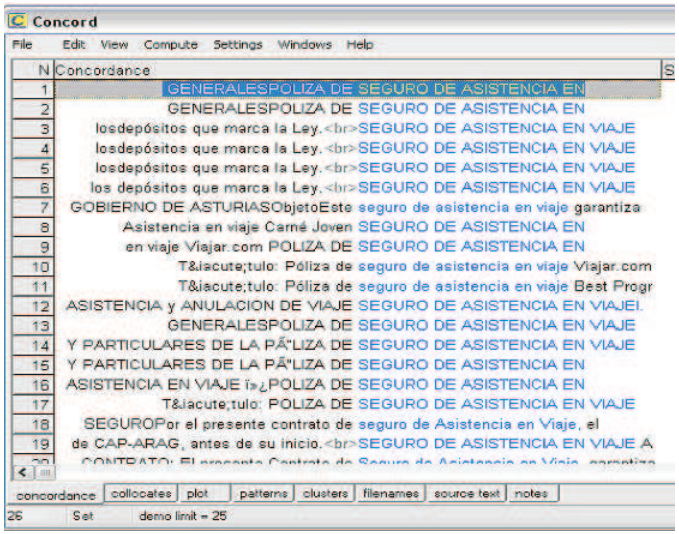

Figure 7. Concordances for 'seguro de asistencia en viaje'

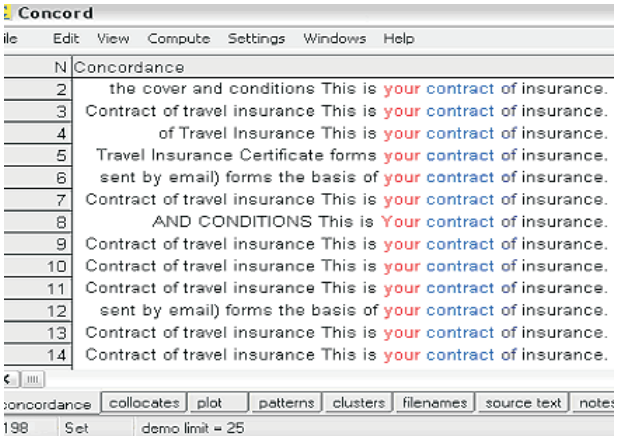

Figure 8. Concordances for 'contract'

The next problem that could arise for the translator is how to translate the English *cover, conditions and exclusions* (cf. Extract 1) into Spanish. A search in the Spanish corpus for the literal translation *condiciones, coberturas y exclusiones* shows only one concordance. On this point it is important to remember that legal

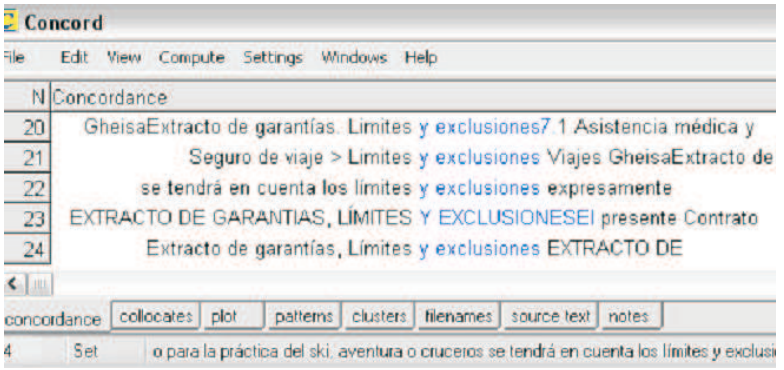

Figure 9. Concordances for 'exclusiones'

Figure 10. Concordances for 'conditions'

language is characterised not only by its precision, but also by its formulaic and extremely conservative style. The translator should be aware of the abundance of verbose and often redundant phraseological units and other fixed expressions and the archaic or conventional forms that these texts contain, often with the sole purpose of making them appear more grandiose. Finally, the Spanish corpus revealed that the term *exclusiones* is always found as part of the phraseological unit *límites y exclusiones* (or, else, as *garantías, límites y exclusiones*), as can be inferred by the results presented by the program when writing *exclusiones* (cf. Figure 9).

A similar problem may be encountered by the translator when translating *CONDICIONES GENERALES* (cf. Extract 2) into English. A search in the corpus for *conditions* shows that in English the construction *General Terms and Conditions* (cf. Figure 10) with capital letters is preferred in most cases.

5.2.2 *Conceptual information*

In English the policies always refer to the *insured person* (cf. Extract 1), whereas the Spanish legal system recognises various figures. As a result, it may be beneficial to distinguish between the *asegurado* (the insured person), the *tomador* (the person who takes out the insurance) and the *beneficiario* (the beneficiary of the insurance). The *asegurado* is the person (either physical or legal) who is exposed to a particular risk, either to his person or his property or assets. In other words, the *asegurado* is the subject of the contract whether in his person (in the case of life insurance or pensions for example) or his property (in the case of house insurance or insurance against fire amongst others). The *tomador* is the person who takes out the insurance and pays the premiums, but may not necessarily be the beneficiary. The *beneficiario* is the person specified in the policy as the recipient of the assistance or compensation covered by the insurance.

The corpus may therefore also be used to clarify concepts and, as a result, identify which person is being referred to in Spanish. Hence, a search in the corpus based on the expression *insured person* (cf. Figure 11) shows definitions such as "Insured person, you, your – each person who an insurance premium has been paid for as shown on the policy schedule".

It may, therefore, be concluded that the English term *insured person* should be translated as *Asegurado* with a capital letter as illustrated by the information shown from the corpus (cf. Figure 12). The option *persona asegurada*, with 20 occurrences, may be ruled out in favour of *Asegurado* or *Asegurados* with 5,692 and 646 occurrences respectively.

In the case of the Spanish fragment (cf. Extract 2), the main problem is rooted in the difficulties of rendering the legislation in translation: *Ley 50/1980, de 8 de octubre, de Contrato de seguro, en la Ley 30/1995, de 8 de Noviembre, de Ordenación y Supervisión de los Seguros Privados.* Here it may be helpful to remember that although there is no substantive communitary legislation on the subject of travel insurance, the contract may be subject to the national regulations of the countries that the parties making the agreement come from. If the customer wants an adaptation of the translation to the British legal system, the translator can use the corpus to find the information necessary to perform this task. The results of a search in the English subcorpus (cf. Figure 13) for *law* (*legislation* was also searched, but produced no occurrences) show a substantial difference from the way that legislation is expressed in Spanish. Whereas in Spanish there is much

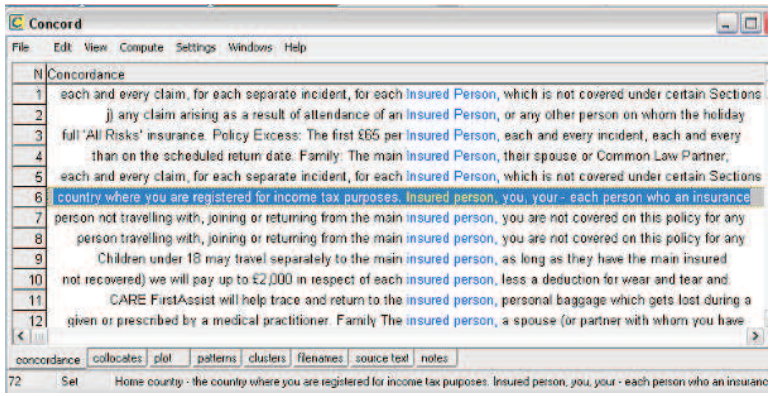

Figure 11. Definition of 'insured person'

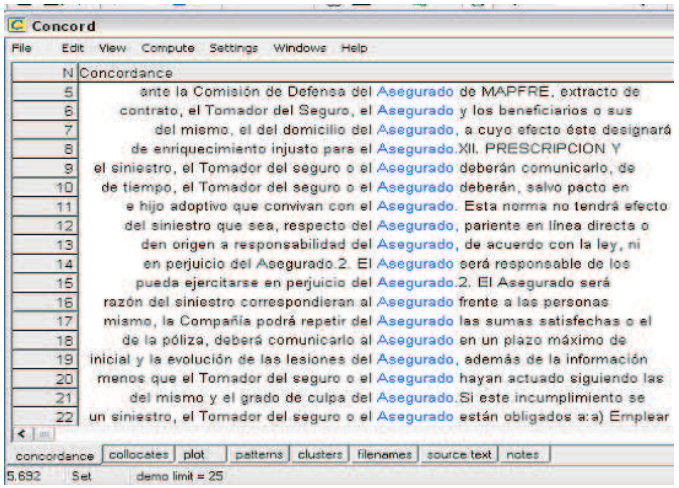

Figure 12. Concordances for 'asegurado'

more precision, in English a more generic means of expression is preferred, with reference made solely to *English Law* and no mention of the specific regulations that apply. In addition, on the subject of legislation, it may be seen that in English the opening formula, *Law applicable*, does not coincide with the Spanish *Artículo preliminar*. This question will be dealt with in the following section (cf. 5.2.3).

Figure 13. Concordances for 'law'

5.2.3 *Textual conventions*

Finally, the preliminary documentation work involves carrying out searches focusing on the typology of the text to be translated. In this case our intention was to find typical opening formulas in the travel insurance policies in Spanish equivalent to the English *Important* (cf. Extract 1). We therefore searched for concordances in Spanish based on *Importante*. The results show that the typical opening formula for this section in Spanish is not *Importante* but *MUY IMPORTANTE* with the whole sequence in capital letters (cf. Figure 14).

In the case of the Spanish text (cf. Extract 2), the typical opening formula consists of a preliminary article (*Artículo Preliminar*) which contains references to the relevant legislation. However, the corpus shows that the English convention has its own opening formula in travel insurance policies, *Law applicable*, which, furthermore, generally appears in the last paragraph of the policy and therefore constitutes a closing formula rather than the opening formula found in Spanish.

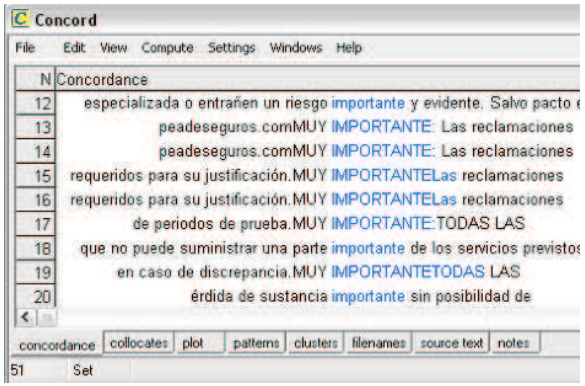

Figure 14. Concordances for 'importante'

5.3 Target text samples

Once all the necessary information has been gathered from the travel insurance corpus, the translator is in a position to offer a translation of both extracts. It is essential to take into account all the points that have been outlined so far given their importance when it comes to segmenting and reorganising the information in the target text (TT). The following are suggested translations of Extracts 1 and 2.

Extract 1 (TT):

> MUY IMPORTANTE
> Esta es su póliza de asistencia en viaje. En ella se incluyen las garantías, límites y exclusiones de los Asegurados y a partir de las cuales podrá efectuarse cualquier reclamación.

Extract 2 (TT):

> General Terms and Conditions
> This is your travel insurance contract.
> Law applicable: This policy is subject to Spanish law.

6. Conclusion

We would like to begin our concluding remarks by quoting Zanettin (2002a: NP):

> Recent research in translation studies has stressed the contribution which corpora of electronic texts can bring to translators. By using appropriate software translators can look up words in a matter of seconds, and highlight patterns by sorting contexts around search words. If a corpus is appropriately designed, it can provide reliable evidence of authentic linguistic behaviour and text-structuring conventions by highlighting recurrent patterns. Terminological and collocational information can be especially useful.

As we have seen, it is possible to meet a large part of the translator's documentation needs through the compilation and/or management of comparable virtual corpora. As a result, translators gain a great deal through becoming both corpus compilers and users. The heuristic tasks necessary in selecting systems to be used for mining the information, as well as the parallel task of finding the information that will be taken from the Internet, are an authentic exercise in applied documentation. Simultaneously, this leads to the development of documentation competence and, as a result, linguistic-textual competence for the translator.

At the same time, a well planned virtual corpus that complies with appropriate design criteria and which is representative in terms of the type of target text that is required may contribute to the development of translators' overall competence. The preparatory tasks involved in selecting and evaluating information sources lead to obvious savings in terms of time and effort that allow the translator to focus on other issues that require more attention, such as taking decisions or evaluating different translation options.

In this article we have focused on the use of virtual corpora as the documentation resource par excellence in specialist translation training. However, the methodology behind corpus compilation is not always very clear and all too often the availability of documents on the Internet is the crucial criterion which determines the size of the collection of texts. As a result, if the collection of texts is to qualify as a "corpus" and be considered as representative of a particular field, it is essential that it conforms to clear design parameters that are set out from the beginning followed by a specific compilation protocol. This protocol is divided into four distinct phases: (a) locating and accessing resources; (b) downloading data; (c) text formatting; and (d) data storage.

Corpus representativeness may also be measured *a posteriori* using *ReCor*, a computer programme that calculates the minimum number of documents and words that should be included in specialised language corpora, in order that they

may be considered representative. It should be pointed out that it is not possible to establish the minimum number of documents for a given corpus *a priori*, as the size will depend on the language and text types involved, as well as on the restrictions of a particular specialised field and any other diasystematic limitations.

Virtual comparable corpora, constructed in accordance with the protocol outlined in this study, are extremely useful for the study of discourse within the field of specialisation under examination, the way this discourse manifests itself in the respective documents as well as the forms these texts take in practice. This utility may be seen from a monolingual and monocultural perspective as well as from the point of view of translation, comparison and interlinguistic and intercultural mediation. As a result, the virtual corpus may be viewed as a highly effective tool in specialised translation training since it promotes autonomous processes of teaching-learning by establishing appropriate mechanisms for specialisation and diversification for the translator. In addition, it encourages the study of texts that students have translated with the objective of correcting and validating translation assignments, as well as many other possible uses that are still to be discovered.

References

ACT. 2005. *Primer estudio de mercado de los servicios de traducción profesional en España de la Asociación de Empresas de Traducción (ACT)*. Madrid: ACT.

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

Aston, G. (ed.). 2001. *Learning with Corpora*. Bolonia: CLUEB.

Aurioles Martín, A. 2005 [2002]. *Introducción al Derecho Turístico (Derecho Privado del Turismo)*. Madrid: Tecnos.

Aurioles Martín, A., Benavides Velasco, P. G. and González Fernández, M. B. 2004. *Contratación Turística*. Technical document BFF2003-04616 MCYT/TI-DT-2004-1. 1–12. <http://turicor.com/privada/documentos/TI-DT-2004-1.pdf>. [14/03/2007].

Austermühl, F. 2001. *Electronic Tools for Translators*. Manchester: St. Jerome.

Bernardini, S. and Zanettin, F. (eds). 2000. *I corpora nella didattica della traduzione. Corpus Use and Learning to Translate*. Bolonia: CLUEB.

Biber, D., Conrad, S. and Reppen, R. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bowker, L. 2002. *Computer-Aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press.

Bowker, L. and Pearson, J. 2002. *Working with Specialized Language: A practical guide to using corpora.* London: Routledge.

Braun, E. 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> [14/03/2007].

Carrasco Jiménez, R. C. 2003. *La ley de Zipf en la Biblioteca Miguel de Cervantes*. Alicante: Universidad de Alicante. <http://www.dlsi.ua.es/asignaturas/aa/Zipf.pdf> [14/03/2007].

CORIS/CODIS. 2006. "Progettazione e costruzione di un Corpus di Italiano Scritto." *CORIS/CODIS*. Bologna: CILTA. <http://corpus.cilta.unibo.it:8080/coris_itaProgett.html> [14/03/2007].

Corpas Pastor, G. 2001. "Compilación de un corpus *ad hoc* para la enseñanza de la traducción inversa especializada." *Trans: Revista de Traductología* 5: 155–184.

Corpas Pastor, G. (ed.) 2003a. *Recursos documentales y técnicos para la traducción del discurso jurídico (español, alemán, inglés, italiano, árabe)*. Granada: Comares.

Corpas Pastor, G. 2003b. "Diseño de un tipologizador para la traducción jurídica: Del corpus al prototipo textual." In *Recursos documentales y técnicos para la traducción del discurso jurídico (español, alemán, inglés, italiano, árabe)*, G. Corpas Pastor (ed.), 33–58. Granada: Comares.

Corpas Pastor, G. 2004a. "Localización de recursos y compilación de corpus vía Internet: Aplicaciones para la didáctica de la traducción médica especializada." In *Manual de documentación y terminología para la traducción especializada,* C. Gonzalo García and V. García Yebra (eds), 223–257. Madrid: Arco/Libros.

Corpas Pastor, G. 2004b. "The Turicor Project: Work in Progress." *Revista Europea de Derecho de la Navegación Marítima y Arenonáutica* xx: 1–14. <http://turicor.com/pdf/corpas2004b.pdf> [14/03/2007].

Corpas Pastor, G. 2004c. "La traducción de textos médicos especializados a través de recursos electrónicos y corpus virtuales." In *Las palabras del traductor. Actas del II Congreso Internacional «El español, lengua de traducción», 20 y 21 de mayo, Toledo 2004,* L. González and P. Hernúñez (eds), 137–164. Brussels: Comisión Europea/ESLETRA. <http://www.turicor.com/pdf/corpas2004c.pdf> [14/03/2007].

Corpas Pastor, G. and Seghiri, M. 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, G. and Seghiri, M. 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.

Corpas Pastor, G. and Seghiri, M. 2007a. "Specialized Corpora for Translators: A Quantitative Method to Determine Representativeness." Translation Journal 11 (3). < http://translationjournal.net/journal/41corpus.htm> [14/03/2007].

Corpas Pastor, G. and Seghiri, M. 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> [14/03/2007].

Corpas Pastor, G. and Seghiri, M. 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 73/240/EEC of 24 July 1973 abolishing restrictions on freedom of establishment in the business of direct insurance other than life assurance.

Council Directive 76/580/EEC of 29 June 1976 amending Directive 73/239/EEC on the coordination of laws, regulations and administrative provisions relating to the taking up and pursuit of the business of direct insurance other than life assurance.

Council Directive 78/473/EEC of 30 May 1978 on the coordination of laws, regulations and administrative provisions relating to Community co-insurance.

Council Directive 84/641/EEC of 10 December 1984 amending, particularly as regards tourist assistance, the First Directive (73/239/EEC) on the coordination of laws, regulations and administrative provisions relating to the taking-up and pursuit of the business of direct insurance other than life assurance.

Council Directive 87/343/EEC of 22 June 1987 amending, as regards credit insurance and suretyship insurance, First Directive 73/239/EEC on the coordination of laws, regulations and administrative provisions relating to the taking-up and pursuit of the business of direct insurance other than life assurance.

Council Directive 87/344/EEC of 22 June 1987 on the coordination of laws, regulations and administrative provisions relating to legal expenses insurance.

Council Directive 90/618/EEC of 8 November 1990, amending, particularly as regards motor vehicle liability insurance, first Council Directive 73/239/EEC and second Council Directive 88/357/EEC on the coordination of laws, regulations and administrative provisions relating to direct insurance other than life assurance.

Council Directive 92/49/EEC of 18 June 1992 on the coordination of laws, regulations and administrative provisions relating to direct insurance other than life assurance and amending Directives 73/239/EEC and 88/357/EEC (third non-life insurance Directive).

Council Directive 92/96/EEC of 10 November 1992 on the coordination of laws, regulations and administrative provisions relating to direct life assurance and amending Directives 79/267/EEC and 90/619/EEC (third life assurance Directive).

Directive 2000/26/EC of the European Parliament and of the Council of 16 May 2000 on the approximation of the laws of the Member States relating to insurance against civil liability in respect of the use of motor vehicles and amending Council Directives 73/239/EEC and 88/357/EEC.

Directive 2000/64/EC of the European Parliament and of the Council of 7 November 2000 amending Council Directives 85/611/EEC, 92/49/EEC, 92/96/EEC and 93/22/EEC as regards exchange of information with third countries.

Directive 2002/13/EC of the European Parliament and of the Council of 5 March 2002 amending Council Directive 73/239/EEC as regards the solvency margin requirements for non-life insurance undertakings.

Directive 2002/92/EC of the European Parliament and of the Council of 9 December 2002 on insurance mediation.

European Parliament and Council Directive 95/26/EC of 29 June 1995 amending Directives 77/780/EEC and 89/646/EEC in the field of credit institutions, Directives 73/239/EEC and 92/49/EEC in the field of non- life insurance, Directives 79/267/EEC and 92/96/EEC in the field of life assurance, Directive 93/22/EEC in the field of investment firms and Directive 85/611/EEC in the field of undertakings for collective investment in transferable securities (Ucits), with a view to reinforcing prudential supervision.

First Council Directive 73/239/EEC of 24 July 1973 on the coordination of laws, regulations and administrative provisions relating to the taking-up and pursuit of the business of direct insurance other than life assurance.

Fletcher, W. H. 2004. "Facilitating the Compilation and Dissemination of Ad-Hoc Web Corpora." In *The Fith International Conference on Teaching and Language Corpora,* G. Aston, S. Bernardini and D. Stewart (eds), 1–18. Amsterdam: Benjamins. <http://www.kwicfind-

er.com/Facilitating_Compilation_and_Dissemination_of_Ad-Hoc_Web_Corpora.pdf> [14/03/2007].

Giouli, V. and Piperidis, S. 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> [14/03/2007].

Granger, S. and Petch-Tyson, S. (ed.). 2003. *Extending the Scope of Corpus-Based Research: New Applications, New Challenges*. Amsterdam and Atlanta: Rodopi.

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

Insurance Act 2000.

Kenny, D. 2001. *Lexis and Creativity in Translation. A Corpus-based Study*. Manchester: St. Jerome.

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

Laviosa, S. (ed.). 1998. *L'approche basée sur le corpus / The Corpus-based Approach*, *Meta* 43 (4).

Ley 18/1997, de 13 de mayo, de modificaciones del artículo 8 de la Ley de Contrato de Seguro, para garantizar la plena utilización de todas las lenguas oficiales en la redacción de los contratos. BOE. 0115 de 14 de mayo de 1997.

Ley 30/1995, de 8 de noviembre, de ordenación y supervisión de los Seguros Privados.

Ley 50/1980, de 8 de octubre, del Contrato de Seguro.

Ley 50/1980, de 8 de octubre, del Contrato de Seguro.

Moreiro González, J. A. 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> [14/03/2007].

Orden Ministerial de 27 de enero de 1988 por la que se califica la cobertura de las prestaciones de asistencia en viaje como operación de seguro privado.

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

Radev, D., Fan, W., Qi, H., Wu, H. and Grewal, A. 2005. "Probabilistic question answering on the web." *Journal of the American Society for Information Science and Technology (JASIST)* 56 (6): 571–583. <http://filebox.vt.edu/users/wfan/paper/www/www.pdf> [14/03/2007].

Sanahuja, S. and Silva, A. 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> [14/03/2007].

Sánchez-Gijón, P. 2003a. "És la web pública la nova biblioteca del traductor?" *Tradumàtica: Traducció i tecnologies de la informació i la comunicació* 2: 1–7. <http://www.bib.uab.es/pub/tradumatica/15787559n2a7.pdf> [14/03/2007].

Sánchez-Gijón, P. 2003b. *Els documents digitals especialitzats: utilització de la lingüística de corpus com a font de recursos per a la traducció.* PhD Thesis. Barcelona: Universidad Autónoma de Barcelona.

Sánchez Pérez, A. and Cantos Gómez, P. 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.

Second Council Directive 88/357/EEC of 22 June 1988 on the coordination of laws, regulations and administrative provisions relating to direct insurance other than life assurance and laying down provisions to facilitate the effective exercise of freedom to provide services and amending Directive 73/239/EEC.

Seghiri, M. 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 Thesis. Málaga: Universidad de Málaga.

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

The Financial Services and Markets Act 2000 (Regulated Activities).

The Insurers (Reorganisation and Winding Up) Regulations 2004.

Varantola, K. 1997. "Translators, dictionaries and text corpora." In *I corpora nella didattica della traduzione*, S. Bernardini and F. Zanettin (eds), 117–133. Bologna: CLUEB.

WTTC. 2006a. *World Travel and Tourism climbing to new heights. The 2006 Travel & Tourism Economic Research.* London: World Travel & Tourism Council. <http://www.wttc.org/2006TSA/pdf/World.pdf> [14/03/2007].

WTTC. 2006b. *United Kingdom Travel and Tourism climbing to new heights. The 2006 Travel & Tourism Economic Research.* London: World Travel & Tourism Council. <http://www.wttc.org/2006TSA/pdf/United%20Kingdom.pdf> [14/03/2007].

WTTC. 2006c. *Ireland Travel and Tourism climbing to new heights. The 2006 Travel & Tourism Economic Research.* London: World Travel & Tourism Council. <http://www.wttc.org/2006TSA/pdf/Ireland.pdf> [14/03/2007].

WTTC. 2006d. *Italy Travel and Tourism climbing to new heights. The 2006 Travel & Tourism Economic Research.* London: World Travel & Tourism Council. <http://www.wttc.org/2006TSA/pdf/Italy.pdf> [14/03/2007].

WTTC. 2006e. *Spain Travel and Tourism climbing to new heights. The 2006 Travel & Tourism Economic Research.* London: World Travel & Tourism Council. <http://www.wttc.org/2006TSA/pdf/Spain.pdf> [14/03/2007].

Yang, D., Cantos Gómez, P. and Song, M. 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> [14/03/2007].

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

Zanettin, F. 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> [14/03/2007].

Zanettin, F. 2002b. "CEXI. Designing an English Italian Translational Corpus." In *Teaching and Learning by Doing Corpus Analysis,* B. Ketteman and G. Marko (eds), 329–343. Amsterdam: Rodopi.

Zanettin, F., Bernardini S. and Stewart, D. (eds). 2003. *Corpora in translator education*. Manchester: St. Jerome.