Acquiring instrumental sub-competence by building do-it-yourself corpora for business translation

Daniel Gallego Hernández

University of Alicante

Abstract: The aim of this paper is to share our experience in teaching how to build DIY corpora in business translation courses. Business and finance texts have a significant presence on the web and there is free software for Windows that can assist the translator in the different stages of the process of building DIY corpora from web resources. The model we propose in our courses takes into account these two realities and develops some of the sub-competencies that translation competence consists of, especially the instrumental one, which concerns the use of information and communication technologies and documentation resources. This model not only allows to develop translators' instrumental sub-competence but also to introduce for the first time those who are not familiar with command-line interfaces to the basics of corpora building as a translation resource.

1. Introduction: need of developing instrumental sub-competence

Translation and Interpreting degrees in Spain are essentially based on developing a set of sub-competencies that graduates must have acquired during their courses. PACTE translation competence provides a common reference framework for the definition of these sub-competencies: 1) bilingual sub-competence, related to the procedural knowledge (pragmatics, sociolinguistics, textuality, grammar and lexis) required to communicate in two languages; 2) extra-linguistic sub-competence, which comprises bicultural, encyclopaedic and subject

knowledge; 3) knowledge about translation sub-competence, related to the aspects of the profession, the work market, etc.; 4) instrumental sub-competence, which involves procedural knowledge related to the use of documentation resources and information and communication technologies; 5) strategic sub-competence, which is an essential sub-competence that concerns procedural knowledge and guarantees the efficiency of the translation process and creates links between the former sub-competencies; and 6) psycho-physiological components, which include cognitive components (memory, perception, attention, etc.), attitudinal aspects (curiosity, rigour, etc.) and abilities (creativity, analysis, logical reasoning, etc.). PACTE affirms that these sub-competencies establish hierarchies and variations, and interact during the translation process. Strategic sub-competence monitors the process and compensates the other sub-competencies (PACTE, 2008: 106-107).

Translator trainees seem to activate these sub-competencies according not only to the characteristics of each type of translation (literary translation, medical translation, simultaneous interpreting, business translation, etc.) but also to their knowledge and abilities. In the case of business, economic or financial translation, translator trainees usually lack subject-specific knowledge, especially when dealing with highly specialized texts. Therefore they have to make use of both lexicographical resources such as dictionaries or terminological databases and textual resources such as parallel texts, i.e. those texts related to the source text which "provide information on text-type conventions or particularities of field-specific language use: terminological, collocational, phraseological, syntactical, etc." (Sánchez Gijón, 2009: 113).[1] These resources seem to involve the activation of instrumental sub-competence.

As translator trainers in our business translation courses we seek to place emphasis on developing not only their extra-linguistic sub-competence but also their instrumental one. On the one hand, regarding extra-linguistic sub-competence, apart from introducing translator trainees into business, economic or financial subjects and their terminology, we also try to

develop some other aspects related to text types and their location on the web. They may thus foresee where to find parallel texts when translating this kind of texts in the future. On the other hand, regarding instrumental sub-competence, which covers different types of knowledge, such as the use of the translator's workstation (operating systems, word processor, shortcuts, file management, etc.) or lexicographical or linguistic resources such as terminology databases or on-line dictionaries, we also ensure the development of certain aspects specifically related to compensate the lack of subject-specific knowledge we mentioned above. In this respect, we focus on developing skills and abilities related to the use of search engines applied to the retrieval of parallel texts for business translation and the use of different software which can help translator trainees in compiling these texts into their own specific linguistic tool, which may complement traditional translation tools such as dictionaries. In short, our purpose concerns developing those aspects involved in building do-it-yourself corpora which Varantola (2003: 69) summarizes as follows:

> Corpus design and design criteria, search strategies and search word selection, source criticism to assess the reliability of corpus texts, assessment of corpus adequacy and relevancy, software literacy in general, selection of Internet search engines, integrated use of word processing tools and corpus tools.

2. Web for corpus: overview of building do-it-yourself corpora

Building DIY corpora (also known as *ad hoc* corpora, virtual corpus, precision corpora or disposable corpora) and its exploitation as a resource for specialised translation has been discussed by several authors (Aston, 1999, 2000; Maia, 2000; Varantola, 2000, 2003; Zanettin, 2002; Corpas Pastor, 2001, 2004a, 2004b; Corpas Pastor & Seghiri, 2009; Sánchez Gijón, 2003, 2004, 2009).

Broadly speaking, the design and compilation of this IT tool involves bringing together within a short time a set of parallel texts whose terminological, phraseological and conceptual

information may help the translator in solving specific translation problems and difficulties. This means finding parallel texts, storing them into the hard disk and having them converted into a readable format by the specific software for extracting from the corpus the information needed to solve each specific problem or difficulty.

In general terms, basic sources from which translators may find parallel texts are CD-Roms, scanned texts (which require OCR processing) and, of course, the Internet. Searches on the Internet for building DIY corpora as explained by these authors may be broadly related to the three types of searches presented by Austermülh (2001: 52-64): institutional searches, thematic searches and keyword searches. With regard to both institutional and thematic searches, Corpas Pastor (2001: 164-165; 2004a: 149-252; 2004b: 141-149) has suggested, in the context of her medical translation courses, different types of sources from which she compiles her *ad hoc* corpus: portals and directories, medical organisations, specialized journals, virtual libraries, bibliographical databases, encyclopaedias, scanned articles, etc. Her intention is to present the corpus to the students as a documentary resource. Focusing on keyword searches, the general protocol for building do-it-yourself corpora basically involves retrieving a set of keywords from the source text which are representative of its topics, translating them into the target language, performing a search on the Internet on the basis of these translated keywords, and downloading the results en masse. It should also be noted that recent studies undertaken by Corpas Pastor & Seghiri (2009: 84), who are interested in translating travel insurance documents, or Sánchez Gijón (2009: 117-118), who is interested in technical translation (in particular, the translation of handbooks of air conditioning), introduce other strategies. These strategies entail retrieving parallel texts on the basis of keywords which are not only representative of the topic but also of the genre or text type the translator is dealing with: e.g. in order to retrieve a set of handbooks of air conditioning the translator can type this search: *"aire acondicionado" "manual de instrucciones" véase* (in this case, *aire acondicionado* is related to

the subject, *manual de instrucciones* is the term used to refer to the genre and *véase* is a typical expression used in this genre). Besides selecting the most suitable keywords in order to retrieve parallel texts, they also recommend to take advantage of the specific functions of the search engine the translator uses (basically narrowing searches and combining multiple terms).

These authors usually propose not only different sources of parallel texts and different tools to retrieve them but also different software to complete the other stages DIY corpus building process consists of. Concerning the downloading of texts, Corpas Pastor (2004*b*: 247) proposes off-line browsers such as *Website Extractor* and *HTTrack Website Copier;* Corpas Pastor & Seghiri (2009) propose free *GNU Wget* (a command line interface program usually for UNIX whose basis has also been used for graphical programs for Windows such as *GWget*); Sánchez Gijón (2004: 186; 2009: 117-118) proposes client-based meta-search applications like *Copernic Pro* and also off-line browsers, such as *Offline Explorer* (both commercial versions). Concerning the conversion of parallel texts, Corpas Pastor & Seghiri (2009) propose different commercial and free software tools such as *Pdf to Word converter 3.0, PDF Converter, Easy PDF to Word Converter.* Finally, concerning text formatting, Sánchez Gijón (2009: 117-118) also proposes *Wordsmith tools* for converting some codes into special characters.

It is true that the basic stages described by these authors involves using different resources or tools for building a corpus even if this software has not been specifically developed for this purpose. However, Zanettin's (2002: 246) words have become a reality:

> Hopefully software producers and developers will create professional applications in which the functions of browser and concordancer will be better integrated, and DIY will find their place in the translator's workstation together with other corpus resources and computer assisted tools.

In fact, the general protocol involving extracting from the web a set of parallel texts on the basis of some keywords or seed words has been implemented in specific software which has

been developed to automatically process the whole aspects of building a DIY corpus: commercial *Sketch Engine* (Kilgarriff *et al.*, 2004), *Jaguar* (Nazar *et al.*, 2008), which is currently implemented as a web application, or *BootCaT* (Baroni & Bernardini, 2003), whose graphic user interface is available since March 2010.

3. Our proposal: building a DIY corpus of financial statements

The model of building DIY corpora that we use in our courses has to do with business translation. It focuses on information retrieval rather than netsurfing. It takes into account the significant presence of business and finance texts on the web and consists of four basic stages which allow the use of different software for Windows[2] and help us to develop translator trainees' extra-linguistic and instrumental sub-competencies.

The first stage involves source-text analysis (genre and topic recognition). It requires the developing of both extra-linguistic sub-competence related to business and finance texts, and instrumental sub-competence related to the sources containing parallel target-language texts. The second stage involves browsing the Internet in order to find these texts and retrieving these resources by the use of search engines. In this stage, instrumental sub-competence is activated by the use of search-engine query languages and by the evaluation of resources according to the needs of the translator. The third stage entails downloading these parallel texts en masse. Instrumental sub-competence is also activated since the translator can use download managers or offline browsers. Finally, the fourth stage is related to the conversion of parallel texts to TXT format so that the translator can exploit them with corpus query tools.

Once we have introduced these four main steps, we will develop them supposing that we are dealing with French and Spanish translation of financial statements, particularly the notes to annual accounts. Having discussed the main basic aspects concerning the analysis and location on the Internet of this kind of texts (stage 1), we will show three strategies related to the second

and third stages in order to build a comparable corpus of original financial statements written in the target language. These are three strategies we use to teach in our business translation course. Finally, we will discuss the fourth stage, which may take place while developing the third one.

3.1. Source-text analysis

When dealing with the translation of the notes to the annual accounts of an international company in both directions (French-Spanish and Spanish-French), translator trainees may need not only to build a corpus consisting of financial statements in both French and Spanish but also to know what financial statements are and consist of.

For instance, Google operator *define* and glossaries on the Internet may offer several definitions of what expressions such as *cuentas anuales, estados financieros* or *estados contables* may mean in Spanish, and *états financiers, comptes annuels* or *états comptables* may mean in French: basically those formal accounting documents (balance sheet, income statement, statement of cash flows and their notes) which summarize the financial states of affairs of a society. Differences between consolidated and social accounts should also be explained.

Besides leading our students to perform searches in order to learn about the nature of these documents, we usually place emphasis on two main aspects of this kind of documents which are decisive for the purposes of building DIY corpora: genre and topic aspects of source texts and the location of parallel texts on the Internet.

Regarding the particular case of notes to annual accounts, genre aspects and terminology referring to this text type prove to be a stumbling block for a first global approach to this genre. In fact, this document is usually included in a broader document with other text types or genres such as balance sheets, profit and loss accounts or auditors' reports. This broader document may be called *estados financieros, informe financiero, estados contables* or *cuentas anuales* in

Spanish, and *états financiers, comptes annuels, rapport financier* in French. This kind of macrogenre can also be included in another kind of macrogenre called *informe anual* or *memoria annual* in Spanish, and *rapport annuel, document de référence* in French, which, besides containing the former annual accounts, may also contain other text types such as the *carta del presidente, informe de gobierno corporativo, informe social*, etc. (Pizarro Sánchez, 2009). A terminological problem stemming from a certain degree of vagueness (Gerzymish-Arbogast, 1989) may be added to these typological aspects in the case of Spanish: as Rynne (2001: 34) affirms, the term *memoria* may refer to both notes to the annual accounts *(memoria de cuentas anuales)* and annual report *(memoria anual).* We not only explain to our students the problem of vagueness concerning financial statements but we also suggest different kinds of exercises which may help them, in a global approach to annual accounts, to identify those basic keywords which always appear in each document, e.g. *activo* and *pasivo* in the case of balance sheets, or, in the case of the notes to annual accounts, those keywords resulting from their structure: *principes et méthodes comprables, immobilisations, amortissements*, etc. in French, and *bases de presentación, principios contables, normas de valoración, inmovilizado*, etc., in Spanish. This kind of analysis may help them in the next stage when selecting the keywords representing the text type they need to retrieve.

Regarding the location on the Internet of this kind of texts, we usually suggest to our students two main sources containing financial information related to the source genre: corporate websites where societies make information available to shareholders, specialized press, potential investors, etc., and official registers where annual accounts are filed; Spanish and French websites offering this kind of information are *www.cnmv.es* and *www.info-financiere.fr*. We normally introduce our students to this kind of websites even before developing DIY corpus-related skills (Gallego Hernández, 2010).

3.2. Strategies for retrieving and downloading parallel texts en masse

Once we have identified the genre and the topic we are dealing with and the location on the Internet of those parallel texts which may be included in the DIY corpus, we start with their retrieval. We usually suggest three main strategies for retrieving parallel texts in our courses, which include different ways of locating corporate websites and the use of different kinds of freeware. The first one is the simplest one: we encourage our translator trainees to download the whole financial information from a single corporate website. The second one involves downloading financial information not from one corporate website but from many. This strategy leads not only to the building of a larger corpus but also to the searching of different corporate websites. This may be carried out by surfing the Internet or by retrieving information from the web. The third strategy concerns the creation of one's own list of URLs linking to parallel texts. This may be done by retrieving information from the web itself or the websites of official registers. In this case, we usually suggest using commercial search engines (normally to create the list from a particular website, which may require word processing) or client-based meta-search engines (normally to create the list from the web). These processes and strategies allow the developing of different aspects related to instrumental sub-competence such as the use of different software, the knowledge of query languages and their application to the retrieval of parallel texts, etc. In the next paragraphs we will explore these strategies further.

3.2.1. Downloading financial information from a single corporate website

This strategy entails downloading financial information from a corporate website, especially of a group, by using off-line browsers, which are originally developed to retrieve websites for viewing when not connected to the Internet. For this purpose, knowledge about URL structures (servers, domains, directories, symbols such as # or ?, file extensions, differences between websites and web pages, etc.) and netsurfing are required.

Normally we suggest to our students a starting specific corporate website, particularly the one which has been analysed during the first stage. They have to identify a web page which includes links to the financial information they deem representative of the translation of notes to annual accounts. Generally the content of this kind of websites is clearly identified by their tab menu. One of these tabs is usually called *inversores y accionistas, información para accionistas e inversores, información económico-financiera* in Spanish, and *finance, espace actionnaires, information financière* in French. These particular tabs contain financial information in different text types (annual accounts, annual reports, quarterly reports, statistics, press releases, highlights, etc.) which are normally available at different URLs. The selected web page or pages will be the starting point from which the off-line browser will download texts.

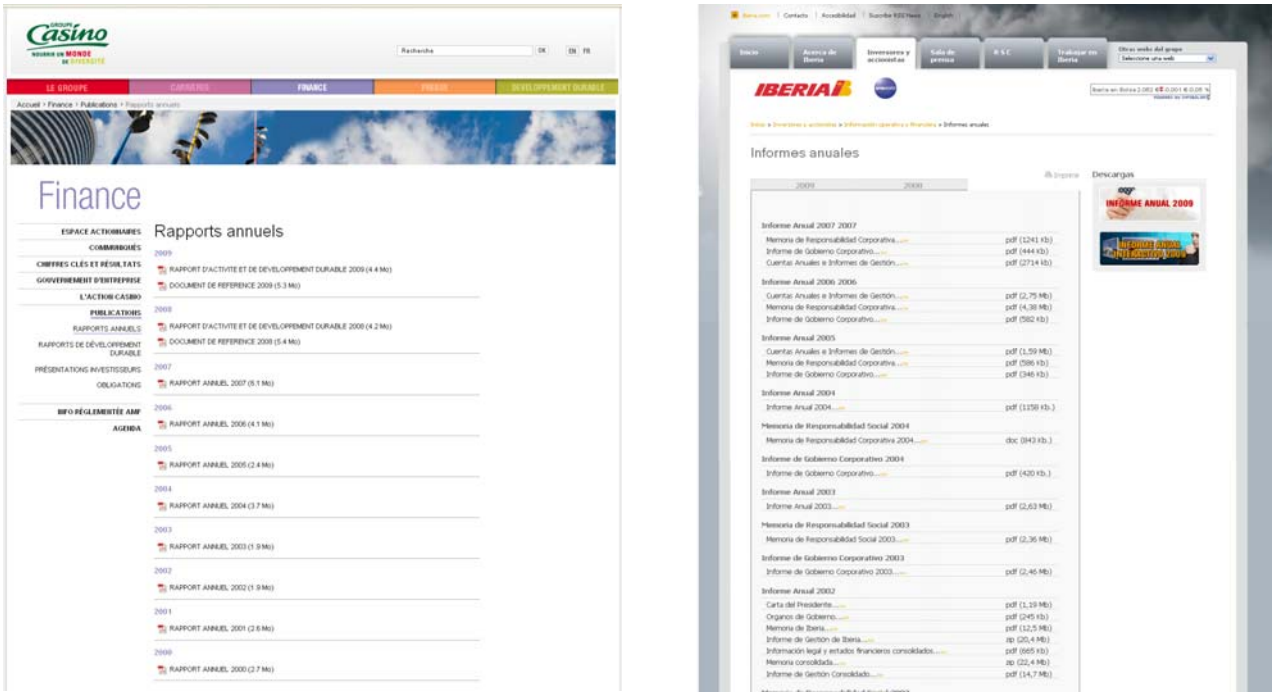

Figure 1. Corporate Web Pages Containing Financial Information

Once translator trainees have decided which web page they are downloading the parallel texts from, we proceed to explain the main functions of off-line browsers (limits, depth, scan rules, etc.). In our courses we usually deal with *Getleft*, which is the most valuable free software

programme in Spanish according to the team of softonic.com (the software download portal we normally use in our courses), and *HTtrack*, which, besides being suggested by some of the former authors, is the most downloaded free Spanish software programme by softonic.com users. Nevertheless we also encourage our students to look for other software by performing query searches such as *related:www.httrack.com* or by navigating in *www.softonic.com* or other similar portals and reading the comments of their teams and collaborators so that they get used to read comments about those applications they are about to install in their computers. At the end of this stage, translator trainees should have saved into their hard disks a copy of a part of the corporate website related to financial information.

3.2.2. Downloading financial information from different corporate websites while navigating the Internet

The second strategy we usually explain in our courses involves downloading financial information while surfing the web. It also concerns new kinds of software and information retrieval. In this case, instead of downloading financial information from a single corporate website using off-line browsers, we lead our students to locate different corporate websites and to download parallel texts from them. For this purpose information retrieval and query language knowledge (differences between search engines, meta-search engines, main operators and their application to the retrieval of parallel texts, etc.) are required to get direct access to those corporate web pages containing financial information. The main idea is to use the keywords that identify the name of those tabs of corporate websites containing the required parallel texts. As we normally use Google to carry out our searches, and as we know that these keywords are usually in the links of these websites, we may combine these keywords with operators such as *inanchor* or *intitle*, which retrieve pages that have the keyword somewhere in their URLs and pages that have the keyword somewhere in their titles, respectively. Other field operators may

also be used to reduce noise in the results such as *site:com,* which retrieves pages from commercial websites. The results of this kind of search equation should be a list of different corporate web pages linking directly to parallel texts. Once they get their results, translator trainees can now navigate from website to website on the basis of these results (at this point we normally insist on using the middle mouse button in order to open web pages in a new tab and not to lose the information of Google results page).

Regarding the downloading of the information translator trainees are identifying while navigating from website to website, we normally use another kind of free download manager which is not an off-line browser. In particular, we suggest *JDownloader,* which is written in the Java programming language and was originally developed to download files from One-Click-Hosters like Megaupload or Rapidshare. Through its linkgrabber, users can automatically download all the files linked to the URL appearing in a text they have copied to the clipboard (ctrl. c). This function allows translator trainees to copy one by one the URLs hidden in the names of the linked files from the web pages containing links to parallel texts (context menu function) and then let *JDownloader* automatically download them, or even to copy the whole source code (ctrl. u, Mozilla) of the web pages and then let *JDownloader* grab the links and download the parallel texts.

3.2.3. Downloading financial information from a list of URLs

The third strategy we develop in our business translation courses involves retrieving parallel texts by the use of search engines. In this case translator trainees do not search a single corporate website from which they download parallel texts using an off-line browser nor do they use search engines in order to search those web pages from corporate websites containing financial information. We suggest them to generate a list of URLs linking to parallel texts directly from Google results web page. We will discuss two different examples of generating

this kind of lists. The first one concerns the use of commercial search engines like Google and entails retrieving parallel texts from official registers websites where annual accounts are filed, i.e. *www.cnmv.es,* in the case of Spanish, and *www.info-financiere.fr,* in the case of French. The second one involves the use of client-based meta-search engines like free *Copernic Agent Basic*, as similarly suggested by Sánchez Gijón (2009: 117-118), and consists in retrieving parallel texts from the web itself.

In the case of retrieving information from this kind of official websites, we encourage again netsurfing exercises so that translator trainees become familiar with the structure of these websites and identify those directories containing the parallel texts they are looking for. Once they have identified these directories, they can retrieve the texts by using the *site* operator. In the case of *www.cnmv.es*, the search equation may be *site:http://www.cnmv.es/audita*, which just retrieves financial reports written in Spanish. In the case of *www.info-financiere.fr,* the search equation may be *site:http://www.info-financiere.fr/upload/,* which retrieves different text types containing financial information. In this case, the search may be narrowed by including the operator *intitle:"rapport financier"*, so that Google retrieves texts whose titles include the keyword *rapport financier* from the website *www.info-financiere.fr*. The next figure shows Google's results for both searches:

Figure 2. Google Results for *intitle:"rapport financier" site:http://www.info-financiere.fr/upload/* and *site:http://www.cnmv.es/audita*

There are two possibilities to download these results depending on whether translator trainees use *HTtrack* or *JDownloader*. In the case of using *HTtrack*, which, besides allowing the retrieval of a website, also makes it possible to download those files linked to a URL list (this is not the case for *Getleft*), translator trainees should generate a cleaned list of URLs from the results web page. For this purpose, among the different ways of generating this kind of list, the source code text of this web page may be edited with a word processor by using search and replace functions. In this case, we may suggest different strategies based on search and replace functions depending on whether the word processor has recognized hyperlinks when copying and pasting them or not: if it has, the strategy may entail selecting the whole text which has not

a hyperlink format, cutting it and pasting it in a new document; if the word processor has not recognized hyperlinks, the strategy may involve replacing the keywords *href="* and *"* (which normally define URLs in HTML) by a paragraph break in order to separate URLs from the rest of the code, and then sorting the text in alphabetical order so that we can clean the code and obtain the list of URLs which *HTtrack* will recognize. At this state of the process, we also introduce our students to using regular expressions y search and replace functions and macros in MS Word.

Figure 3: MS Word Find and Replace

In the case of using *JDownloader*, the process of downloading the results is easier: by just copying the source code, the download manager will recognize its URLs and automatically download them.

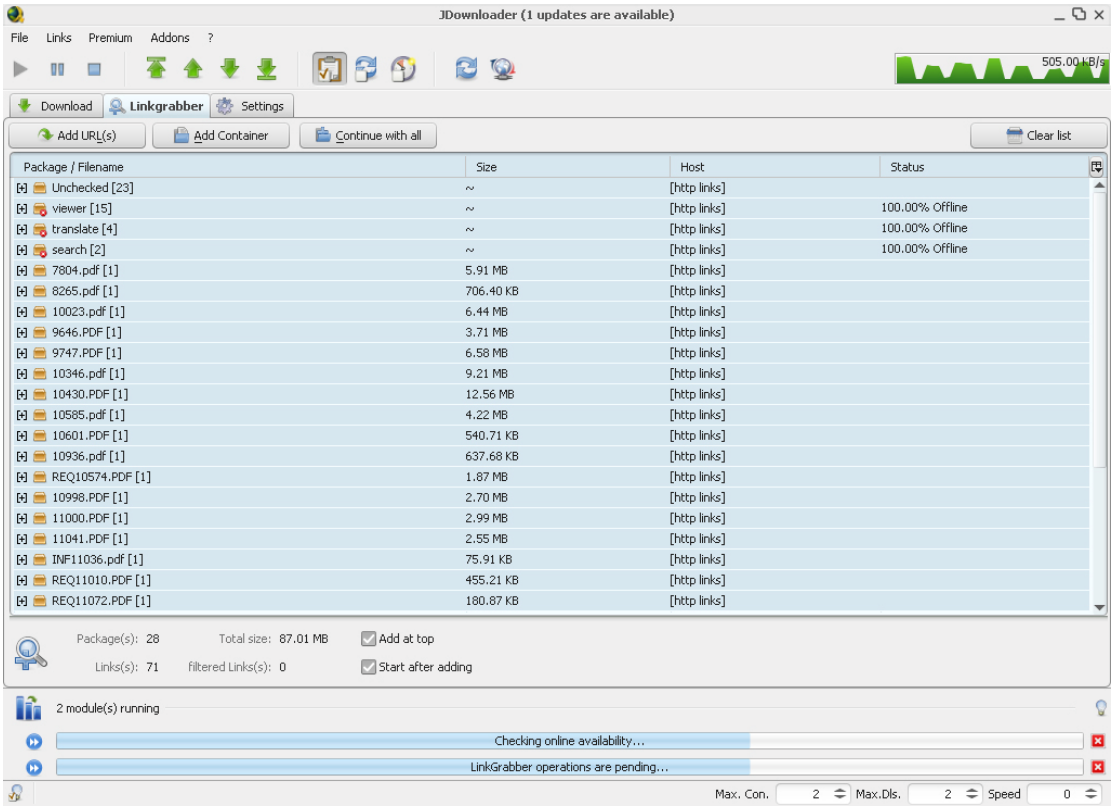

Figure 4. JDownloader Linkgrabber

Regarding the information retrieval from the web, it is possible to search parallel texts by using keywords that identify both genre and topic. In the case of financial statements, the keywords representing the genre usually appear in the title of these documents. The *title* operator may then be used: *intitle:"rapport financier" OR intitle:"comptes annuels"*, in the case of French, and *intitle:"cuentas anuales" OR intitle:"estados financieros"*, in the case of Spanish. Other keywords which usually appear in the notes to annual accounts and which we normally have studied during the source text analysis may be added to these search equations: e.g. *"immobilisations corporelles"*, *"méthode comptable"*, etc., in the case of French, or *"inmovilizado material"*, *"criterios contables"*, etc., in the case of Spanish. Finally, it is also possible to add another operator to the search equation in order to filter the results to a format in which financial statements are usually represented: *ext:pdf* or *filetype:pdf* in the case of Google.

When using client-based meta-search engines, translator trainees should be aware of which search engines support the software they are using. In the case of *Copernic Agent Basic*, Google is not supported. Therefore operators like *ext:pdf* should be translated to those query languages understood by other supported search engines: e.g. *originurlextension:pdf* in the case of Yahoo!, which is supported by *Copernic Agent Basic*. The use of free *Copernic Agent Basic* has the advantage that the user may create a cleaned list of URLs from the results. Then the whole files may be downloaded using download managers or off-line browsers which include the function of getting individual files from an URL list, such as *HTtrack*.

3.3. Conversion to TXT

The final stage involves the conversion of retrieved parallel texts to TXT format once translator trainees have downloaded them to their hard disk or even while downloading them. This stage is especially necessary in the case of financial statements which appear in PDF format and are not readable by corpus query tools. Furthermore, PDF may present documents whose texts have not been recognized. This makes the copy and paste function impossible. There is freeware and shareware for TXT conversion. In our courses, we normally use two applications: *Ultra Document to Text Converter* and *Omniformat*, which also OCRs PDFs. Particularly, the second one includes a watch folder which automatically processes those files placed into it. Therefore, when bringing to this folder the default path or another location where download managers or off-line browsers save files, it may be possible to convert parallel texts while downloading them.

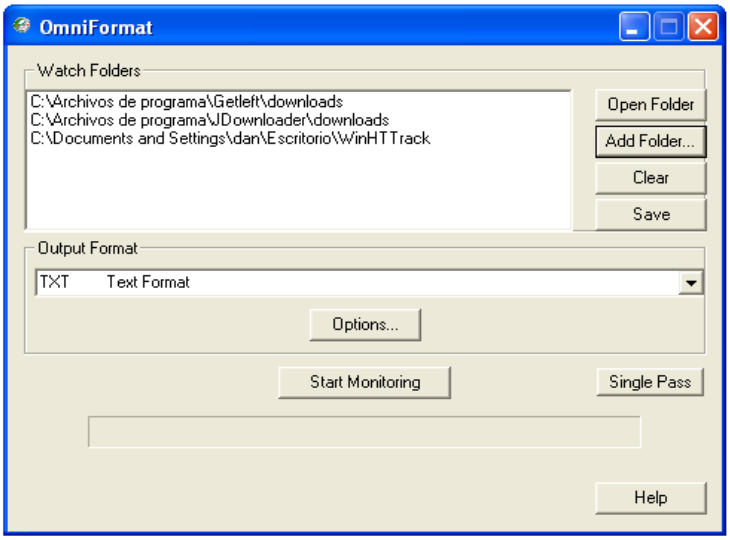

Figure 5. Omniformat Watch Folder

As there seems to be no consensus on the use of this kind of programs and their performances, we also encourage our students to search and evaluate different software to convert PDFs into text-only files in batches or to OCR PDFs so that they can form their own view.

3.4. Optional simple cleaning procedures and file management

Building DIY corpora involves bringing together within a short time a set of on-line parallel texts. This is basically why, once translator trainees have converted their files to TXT, they may start translating the source text. Nevertheless, as translator trainers we may go further in developing their instrumental sub-competence by focusing on simple cleaning procedures and file management.

Concerning simple cleaning procedures, if converters have not cleaned texts up in the previous stage, we may insist on search and replace functions in order to join possible broken paragraphs, remove end-of-line hyphens, remove any text from header and footer, etc. For this

purpose, we usually suggest free *Nodesoft Search and Replace*, which allows searching a lot of text files (basically TXT and HTML) and replacing some of the content.

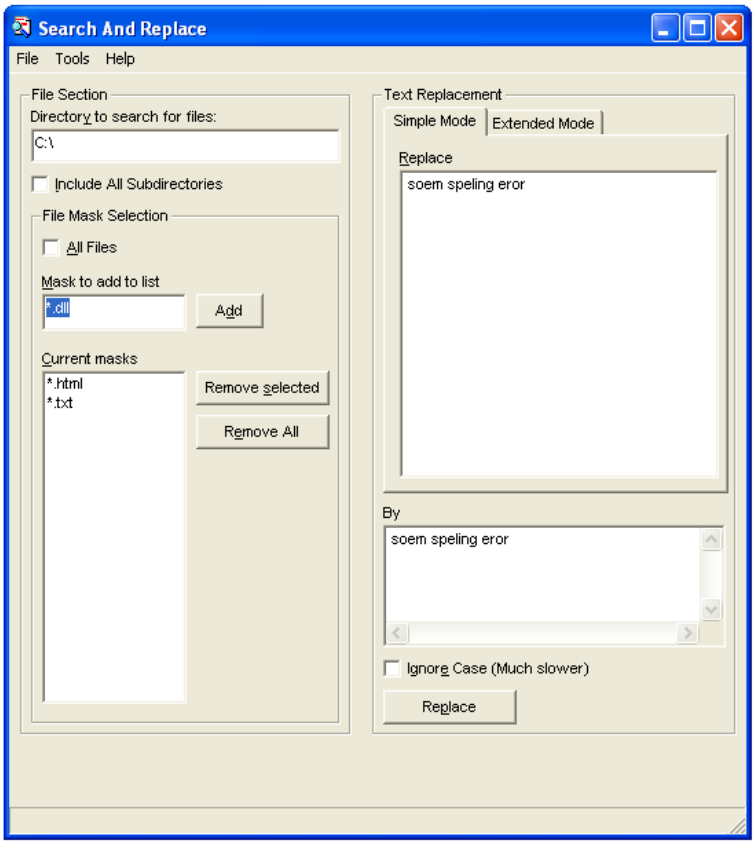

Figure 6. Nodesoft Search and Replace Interface

Concerning file management, besides having insisted on being alert to the directory where files are being downloaded into or converted during the third and fourth stages, we may also focus on renaming parallel texts on the basis of their language, text-type, etc. For this purpose, we normally suggest free *Lupas Rename* (once again we encourage our student to search for different software), whose basic features include operations on the name of the filename and the extension of many files at once.

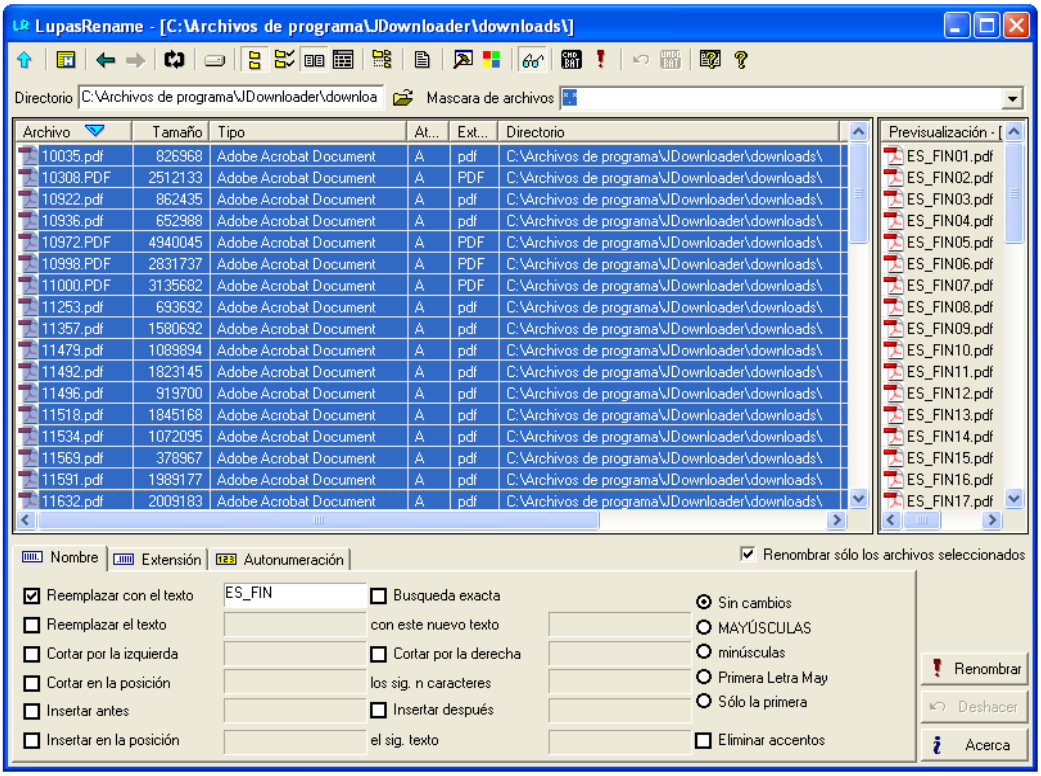

Figure 7. Lupas Rename Interface

4. Conclusion

We have tried to share our experience in teaching how to build DIY corpora in our business translation courses. The model we have presented is based on the use of free software which originally was not specifically developed to help in building corpus but helps in completing the different stages of the process. Teaching specific functions related to corpus building of each program does not really take much time because of their user-friendly interfaces. Furthermore, translator trainees seem to be more and more familiar with computers each year.

Besides having presented different freeware for these stages and focused on source text analysis, we have presented three ways of retrieving parallel texts for the translation of notes to annual accounts as an introduction to DIY corpus building. These strategies draw upon information retrieval through search engines or client-based meta-search engines for both

locating parallel texts' sources and parallel texts themselves. Particularly, we have presented three types of sources from which download managers and off-line browsers may download parallel texts en masse: corporate websites, official register websites and the Web itself. These strategies may take place when dealing with other economic, commercial or financial genres. Therefore, each time translator trainees face a new text, we focus on the retrieval of parallel texts during half an hour or so before starting to build corpora and to translate the text.

It seems to us that these strategies and the model of building DIY corpora prove to be a very interesting exercise to develop not only the extra-linguistic sub-competence but also the instrumental one. Among the different aspects concerning the instrumental sub-competence translator trainers may develop by this kind of exercise we may find: file management (directories, file renaming, searching, file formats, etc.); freeware for different stages (bulk downloading, file conversion, OCR, etc.), evaluation of Internet sources (types of websites, content of websites, URL structure, etc.); word processing (search and replace, regular expressions, table to text conversion, text sorting, macros, etc.); information retrieval (operators, filters, search engines, noise and silence, etc.); Internet (websites vs. web pages, invisible web, HTML code, etc.); Windows (keyboard shortcuts, mouse buttons, context menu, etc.).

Notes

1. In this paper, the concept of 'parallel texts', which is also defined by Vinay & Darbelnet (1958/1977: 272) as texts with situations and stylistic features similar to those of the source text ("situation identique ou comparable […] rédaction semblable quant au style et par conséquent à l'emploi stylistique de la langue par rapport à une situation comparable"), should not be

confused with the concept of ‘parallel corpus’ which normally refers to a collection of original texts and their translations.

2. We will not discuss specific software programs for automatic corpus building in this paper.

References

Aston, G. (1999). “Corpus Use and Learning to Translate”. *Textus,* 12, 289-314. Available at: http://www.sslmit.unibo.it/~guy/textus.htm (accessed: 15 January 2010)

Aston, G. (2000). “I corpora come risorse per la traduzione e per l’apprendimento”. In S. Bernardini and F. Zanettin (eds.) *I corpora nella didattica della traduzione: Corpus use and learning to translate.* Bologna: CLUEB, 21-30.

Austermühl, F. (2001). *Electronic Tools for Translators.* Manchester/Northampton: St. Jerome.

Baroni, M. and S. Bernardini (2003). “The BootCaT Toolkit. Simple Utilities for Bootstrapping Corpora and Terms from the Web”. Available at: http://sslmit.unibo.it/~baroni/Readme.BootCaT-0.1.2 (accessed: 15 January 2010)

Bernardini, S., M. Baroni and S. Evert (2006). “A WaCky Introduction”. In M. Baroni and S. Bernardini (eds.) *WaCky! Working Papers on the Web as Corpus.* Bologna: Gedit Edizioni, 9-40. Available at: http://wackybook.sslmit.unibo.it/ (accessed: 15 January 2010)

Corpas Pastor, G. (2001). “Compilación de un corpus *ad hoc* para la enseñanza de la traducción inversa especializada”. *Trans,* 5, 155-184. Available at: http://www.trans.uma.es/indice.html (accessed: 27 October 2008)

Corpas Pastor, G. (2004a). “La traducción de textos médicos especializados a través de recursos electrónicos y corpus virtuales”. *El español, lengua de traducción,* 2, 137-164. Available at: http://www.turicor.com/pdf/corpas2004c.pdf (accessed: 27 October 2008)

Corpas Pastor, G. (2004b). “Localización de recursos y compilación de corpus vía Internet: Aplicaciones para la didáctica de la traducción médica especializada”. In C. Gonzalo García and V. García Yebra (eds.) *Manual de documentación y terminología para la traducción especializada.* Madrid: Arco, 223-258.

Corpas Pastor, G. and M. Seghiri (2009). “Virtual corpora as documentation resources: Translating travel insurance documents (English-Spanish)”. In A. Beeby, P. Rodríguez Inés and P. Sánchez Gijón (eds.) *Corpus Use and Translating.* Amsterdam/Philadelphia: John Benjamins, 75-107.

Gallego Hernández, D. (2010). “La caza del tesoro en el aula de traducción económica, comercial y financiera: metodología para el análisis de textos y fuentes de documentación”. *VIII Jornadas de Redes de Investigación en Docencia Universitaria: nuevas titulaciones y cambio universitario.* Available at: http://www.eduonline.ua.es/jornadas2010/comunicaciones/283.pdf (accessed: 8 July 2010)

Gerzymish-Arbogast, H. (1989). “The Role of Sense Relations in Translating Vague Business and Economic Texts”. In M. Snell-Hornby and E. Pöhl (eds.) *Translation and Lexicography: Papers read at the Euralex Colloquium, July 1987.* Amsterdam: John Benjamins, 187-196.

Kilgarriff, A., P. Smrz, P. Rychly and D. Tugwell (2004). “The Sketch Engine”. Proc Euralex, Lorient. Available at: http://trac.sketchengine.co.uk/raw-attachment/wiki/SkE/DocsIndex/sketch-engine-elx04.pdf (accessed: 15 January 2010).

Maia, B. (2000). “Making corpora: A learning process”. In S. Bernardini and F. Zanettin (eds.) *I corpora nella didattica della traduzione: Corpus use and learning to translate.* Bologna: CLUEB, 47-60.

Nazar, R, J. Vivaldi Palatresi and M. T. Cabré Castellví (2008). “A Suite to Compile and Analyze an LSP Corpus”. *Proceedings LREC 2008 (The 6th edition of the Language Resources and Evaluation Conference). Marrakech (Morocco).* Available at: http://melot.upf.edu/lrec2008/Nazar_Vivaldi_Cabre_LREC2008.pdf (accessed: 15 January 2010)

PACTE (2007). “Une recherche empirique expérimentale sur la compétence de traduction”. In D. Gouadec (ed.) *Quelle qualification pour les traducteurs ?.* Paris: La maison du dictionnaire, 95-116. Available at: http://webs2002.uab.es/pacte/publicacions/recherche_empirique.pdf (accessed: 9 September 2009)

Pizarro Sánchez, I. (1998). “La traducción de informes financieros: Problemas fundamentales”. In L. Félix Fernández and E. Ortega Arjonilla (eds.) *II Estudios sobre traducción e interpretación.* Málaga: Universidad, 1009-1014.

Pizarro Sánchez, I. (2009). “La comunicación escrita en la empresa: criterios para una taxonomía”. In C. Pérez-Llantada and M. Watson (eds.) *Language for business: a global approach. Seminar Proceedings.* Ávila, 194-160. Available at: http://www.fiu.edu/~ciber/files/spainseminar09_proceedings.pdf (accessed: 6 February 2010)

Rynne, J. (2001). “Approaching the Translation of Spanish Financial Statements”. *The ATA Chronicle,* 30, 6, 33-36.

Sánchez Gijón, P. (2003). “Aplicaciones de la Lingüística de corpus a la práctica de la traducción. Complemento de la Traducción Asistida por Ordenador”. *Puntoycoma,* 79. Available at: http://ec.europa.eu/translation/bulletins/puntoycoma/79/pyc7910_es.htm (accessed: 15 January 2010)

Sánchez Gijón, P. (2004). “La extracción de conocimiento y terminología a partir de corpus ad hoc: el uso de documentos digitales de la web pública”. *Linguistica Antverpiensia,* 3, 179-202.

Sánchez Gijón, P. (2009). “Developing documentation skills to build do-it-yourself corpora in the specialised translation course”. In A. Beeby, P. Rodríguez Inés and P. Sánchez Gijón (eds.) *Corpus Use and Translating.* Amsterdam/Philadelphia: John Benjamins, 109-127.

Varantola, K. (2000). “Translators, dictionaries and text corpora”. In. S. Bernardini and F. Zanettin (eds.) *I corpora nella didattica della traduzione: Corpus use and learning to translate.* Bologna: CLUEB, 117-133.

Varantola, K. (2003). “Translators and Disposable Corpora”. In F. Zanettin, S. Bernardini and D. Stewart (eds.) *Corpora in Translator Education.* Manchester/Northampton: St. Jerome, 55-70.

Vinay, J. P. and J. Darbelnet (1977). *Stylistique comparée du français et de l’anglais. Méthode de traduction. Nouvelle edition revue et corrigée.* Paris: Didier. First edition in 1958.

Zanettin, F. (2002). “DIY Corpora: the WWW and the Translator”. In B. Maia, J. Haller and M. Ulrych (eds.) *Training the Language Services Provider for the New Millennium.* Porto: Faculdade de Letras da Universidade do Porto, 239-248.

Appendix: Some free software for DIY corpus building

PROGRAM	FUNTION
Google <http://www.google.com>	text retrieval (use of field and basic operators)
Copernic Agent Basic <http://www.copernic.com/en/products/agent/index.html>	text retrieval (client-based meta-search program which does not support Google) and generation of a clean list of URLs
Jaguar <http://rc16.upf.es/cgi-bin/jaguar/jaguar.pl>	automatic corpus building (based on Yahoo!) and concordancing (among other statistical functions)
BootCaT <http://sslmit.unibo.it/~baroni/bootcat.html>	automatic corpus building (based on Yahoo!)

Getleft <http://personal.iddeo.es/andresgarci/getleft/english/index.html>	website downloading
HTTrack Website Copier <http://www.httrack.com/>	website downloading and bulk downloading from a list of URLs
JDownloader <http://jdownloader.org/>	bulk downloading (no need of generating a clean list of URLs)
Ultra Document to Text Converter <http://www.ultrashareware.com/Ultra-PDF-To-Text-Converter.htm>	conversion to TXT format
Omniformat <http://www.omniformat.com/>	conversion to TXT format and OCR
Nodesoft Search and Replace <http://nodesoft.com/SearchAndReplace/Default.aspx>	simple cleaning procedures (search and replace functions)
Lupas Rename <http://rename.lupasfreeware.org/>	file management (operations on filenames)
Antconc <http://www.antlab.sci.waseda.ac.jp/software.html>	concordancing and other applications