Electronic target-language specialised corpora in translator education

Building and searching strategies

Patricia Rodríguez-Inés
Universitat Autònoma de Barcelona

1. Introduction

Today, more information is generated by and readily available to a greater number of people than ever before. Technologies (new or otherwise) are highly accessible, a huge and widely varied volume of documents is produced and all sorts of communication events take place. Against that backdrop, translators operate in an increasingly competitive professional environment where, in addition to providing actual translations, they may well be required to act as terminologists, documentalists, technical writers, localisers, reviewers, post-editors or project managers, to give but a few examples. Even translators who solely translate will almost certainly need to be able to adapt to different fields, genres, registers, etc. Additionally, they may be asked to translate both into and out of their mother tongue (direct and inverse translation).

Given those circumstances, translation teachers responsible for training future translators find problems such as insufficient teaching time in the classroom compounded by their own limitations, in terms, for instance, of the depth of their knowledge and experience of each and every area in which translation trainees should become competent in order to deal with the challenges of professional practice. In other words, translation teachers, as information providers, are unable to cope with all the areas of specialisation and all the register and genre variants to which translation trainees seemingly ought to be exposed. It should also be noted that the impossibility of predicting what sort of texts trainees will encounter as professionals makes specialisation at the training stage inadvisable, despite its apparent desirability or necessity (Gavioli 1996).

> El corpus favorece, pues, el desarrollo de los procesos autónomos de enseñanza-aprendizaje, estableciendo los mecanismos adecuados para la especialización y, al mismo

Babel 59:1 (2013), 57–75.
DOI 10.1075/babel.59.1.04rod ISSN 0521–9744 E-ISSN 1569–9668

> tiempo, la diversificación que el mercado laboral exige al traductor hoy en día (Corpas Pastor 2002: 6).

In the light of the above, I take the view that a translation teacher's ideal role is not that of an information provider but that of an information facilitator, someone who guides students in their learning process, helping them as and when necessary and fostering their acquisition of operative knowledge. That is not to say that I deem the acquisition of declarative knowledge irrelevant for future translators, but rather that I consider developing skills that students will be able to apply when faced with problems in new situations to be much more productive. When translation trainees begin practising professionally, they will have to deal with different topics and a wide range of text types and genres, therefore needing to become competent in any given specialised area in a very short time. In that context, corpora are reusable resources that prove highly valuable in all stages of the translation process and, as I contend, in translation education. Many other authors besides myself not only endorse using corpora in translation training but consider *corpus management in translation* an essential skill for future translators (Maia 1997, 2003; Aston 1999; Friedbichler and Friedbichler 2000; Corpas Pastor 2002; Varantola 2003; Jääskeläinen and Mauranen 2004; and Oster 2007, among others). Or, as Varantola puts it, "[. . .] the knowledge of how to compile corpora and use them is an essential part of modern translational competence" (Varantola 2003: 55).

2. Electronic corpora vs. parallel texts in paper format and dictionaries

The vast majority of researchers who have worked with corpora highlight the practical and pedagogical advantages that such resources and the corresponding methodology entail in comparison to reference sources such as parallel texts in paper format and dictionaries. The advantages in question apply both inside and outside the translation classroom.

Aspects of the publishing process mean that general and specialised dictionaries alike pose problems for users such as translators and translation trainees. Firstly, a substantial amount of time can elapse between a work of the kind in question being researched and actually being made available for consultation, a factor that can have a crucial bearing on its accuracy and coverage, especially in fields in which terminology evolves very quickly (Pearson 1996; Maia 1997). To quote Baroni *et al.* (2006), "regular dictionaries will not cover it [terminology of a specialist area], specialist dictionaries, if they exist, will be hard to find and expensive and are likely to be out of date". Secondly, space restrictions often result in dictionaries (particularly bilingual dictionaries in paper format) presenting terms with little or no context. Where provided, accompanying examples tend to be prefabricated,

offering lexical and/or semantic information but no usage-related insight. Corpas Pastor and Postigo Pinazo (2002) state that a corpus, in contrast, can provide information of many different types, including discursive, cultural, terminological, syntactic, combinatorial, semantic, phraseological and stylistic information, as well as statistical data that helps users to determine the relevance of detected patterns.

As Bowker (2000: 20) suggests, parallel texts can be useful for overcoming certain shortcomings (e.g. a lack of contextual information) of lexicographic resources such as dictionaries. It goes without saying, however, that parallel texts lose their relative advantages if they are in paper format, as the effort involved in collecting and reading through printed documents is disproportionate. It would be impossible, in a short time, for translators to collect and consult a sufficiently varied range of documents to guarantee the presence of all or most concepts, terms and linguistic patterns. Furthermore, as Church *et al.* (1991: 116) observe, humans are prone to making mistakes when analysing large quantities of text and attempting to identify patterns without mechanical assistance. Thus, in comparison to the (now relatively old) practice of using parallel texts in paper format, the use of electronic corpora in translation can be seen as a huge advance with positive implications for translator education.

> [. . .] el uso de corpus para la pedagogía de la traducción constituye, hasta cierto punto, un desarrollo lógico si tenemos en cuenta dos factores clave: primero, la creciente 'informatización' y 'automatización' de la actividad traductora como proceso y como producto; y, segundo, el uso tradicional de textos 'paralelos' como fuente de documentación más fiable y completa que los diccionarios al uso (Corpas Pastor 2002: 5).

Regardless of their specific type, electronic corpora offer a series of advantages that can be summarised as follows:

– They take less time to build than corpora in paper format.
– They are reusable.
– They enable users to search for information quickly and systematically in large volumes of heterogeneous documents.
– They offer many examples and a great deal of context.
– They allow users to concentrate on passages of text containing keywords.
– They make it easy to identify collocational patterns.
– They make it easy to identify frequency data, which helps users determine the relevance of detected patterns.
– They make it easy to identify specific data and, in the case of annotated corpora, speed up its retrieval and analysis.
– They provide users with a solid basis for making strategic decisions or lexical choices.

– They offer many opportunities for incidental learning (serendipity).
– They promote a sense of discovery.

In relation to translation teaching in particular, Aston (1999) suggests that using electronic corpora makes learning environments richer:

> Unlike the dictionary, a concordance leaves it to the user to work out how an expression is used from the data. This typically calls for deeper processing than does consulting a dictionary, thereby increasing the probability of learning (Hultsijn 1992). In more general terms, by drawing attention to the different ways expressions are typically used and with what frequencies, corpora can make learners more sensitive to issues of phraseology, register and frequency, which are poorly documented by other tools (Aston 1999: 295).

Varantola (2007) brings some balance to the topic by reminding us that corpora and dictionaries are very different tools, neither of which should be underrated as they may be used to satisfy different needs.

Despite all the benefits of using electronic corpora and the corresponding methodology in translation and translator education, it must be stressed that doing so requires a degree of previous training and an understanding of the implications entailed. As Bowker (2000: 46) suggests, (future) translators need to be made aware of both the advantages and limitations of corpora and corpus-based resources and tools, just as they need to be made aware of the pros and cons of traditional resources. With that in mind, I feel it is important to be familiar with the types of corpora which are most useful for professional and trainee translators, as well as the kinds of tasks that can be performed with them (Rodríguez-Inés 2008, 2010).

3. Most relevant corpus types in translation teaching

The most common and useful types of corpus in translation teaching are basically as follows:

– The parallel corpus: source texts (STs) in language A aligned with target texts (TTs) in language B.
– The bilingual comparable corpus: STs in language A and STs in language B.
– The monolingual target-language specialised corpus: STs in language B.

This article will focus on target-language specialised corpora. As they are relatively easy to compile, at least where some languages and topics are concerned, and often do not require much processing to be able to offer relevant information, such

corpora are widely used in learning environments (Bowker 2000; Corpas Pastor 2001; Bowker and Pearson 2002; López Rodríguez 2002; Zanettin, Bernardini and Stewart 2003; Wilkinson 2005a, 2005b; Rodríguez-Inés 2008; and Buendía-Castro and López-Rodríguez 2010, among others), as well as in some professional environments (Maher, Waller & Kerans 2008; Scott 2011).

4. Target-language specialised corpus, monolingual comparable corpus, ad hoc corpus, virtual corpus, disposable corpus or DIY corpus

4.1. On corpus-related terminology

Like that of any other specialised area, the terminology specific to the field of corpora can be somewhat confusing. A monolingual corpus composed of parallel texts may be called a monolingual comparable corpus. If restricted to a particular topic and/or genre in the target language (TL), such a corpus may be referred to as a target-language specialised corpus. If built to fulfil a specific purpose related to the translation of a particular text, it can also be considered an *ad hoc* corpus, a type of resource which, according to Varantola (2003), is not intended for long-term use. Such a corpus may be described as *virtual* (Ahmad, Holmes-Higgin and Abidi 1994) and *disposable* (Varantola 2000).

> These [disposable corpora] are small, specialized corpora created ad-hoc to serve the needs of the translator for a specific translation project, and their value lies not only in their analysis but even more so in their creation (Zanettin 2002: 239).

The fact that corpora can now easily be created by gathering documents from the internet has given rise to the term DIY corpus (Zanettin 2002), which refers to a collection of HTML-format web pages compiled for the purpose of translating a specific text. A DIY corpus is open, in that more documents can be added to it at any point, and involves no copyright issues.

4.2. On corpus building

Maia (1997, 2000) is one of the staunchest advocates of students creating their own corpora, an activity she feels is of benefit to their learning process. Likewise, I believe that when students look for parallel texts suitable for inclusion in an ad hoc corpus from a range of sources whose reliability they themselves can assess, they are liable to acquire or improve their documentation and textual analysis skills, develop their critical thinking and expand their knowledge of the corresponding topic and genre.

Table 1. Criteria required for a corpus that is a balanced and representative source for specialized translation (Bowker 1996).

Criteria	Requirements for specialized translation
Corpus size	large selection of texts (to determine if usage is widespread or idiosyncratic)
Text size	complete texts (so examples of usage or explanations of concepts are not cut short)
Text type	mixture of instructional, expert, and popularized texts (to help translators achieve an understanding of the subject field and allow them to see different registers of usage)
Date of publication	mainly recent texts, but also some older texts (older texts are useful because concepts are better explained when they first come out; recent texts are needed to reflect the state of the field at present)
Author	texts by a variety of authors (to determine if usage is widespread or idiosyncratic)
Language	preferably texts by authors writing in their native language (to show idiomatic usage)
Culture	texts written by authors with different cultural backgrounds (e.g. British, American, etc.) (to show appropriate regional usage)

To aid the process of building a target-language specialised corpus, Bowker (1996) sets out a list of criteria and requirements that a corpus must meet if it is to be useful where specialised translation and terminology are concerned (see Table 1).

According to Maher, Waller & Kerans (2008), corpus design is of the utmost importance where work involving a highly specialised area and discourse community is concerned. They observe that "identifying the right content is the key to confident decision making later." They "discourage sampling of the Internet by topic keywords alone if working to a high standard within a specialism", and recommend

> attention to genre and a discourse community's reading preferences, given that our goal is not primarily to find 'a good explanation of subject matter' (López-Rodríguez and Tercedor-Sánchez 2008), a purpose for which other research strategies can be equally effective. Rather, we wish to open a window that allows us to observe a community's language use (Maher, Waller & Kerans 2008:62).

Additionally, they suggest that translators seek advice from an expert if unsure as to what texts can be deemed truly representative of a given field.

Bowker's proposed approach to corpus building and that of Maher, Waller & Kerans are clearly complementary. The former is more inclusive in terms of text types that provide not only "good models for the target text" but also explanations.

Despite such recommendations on corpus compilation, a corpus is unlikely

to be balanced or representative if it has been built in the context of a translation classroom activity (i.e. by translation students) or if relevant documents in the language, topic or genre being dealt with are hard to come by. However, Atkins, Clear & Ostler (1992) do not see this as an obstacle to working with corpora, but rather as a factor that should be taken into account when interpreting the data obtained.

> [. . .] we have found any corpus – however "unbalanced" – to be a source of information and indeed inspiration. Knowing that your corpus is unbalanced is what counts. It would be short-sighted to wait until one can scientifically balance a corpus before starting to use one, and hasty to dismiss the results of corpus analysis as "unreliable" or "irrelevant" simply because the corpus used cannot be proved to be balanced (Atkins, Clear & Ostler 1992: 6).

As for text sources, a user-compiled corpus will, in all likelihood, comprise texts extracted from the internet, an approach that has been termed Web for Corpus (WfC) (De Schryver 2002). With or without the help of web downloaders (e.g. WebCopier, Free Download Manager, Flashgot) or purpose-specific web corpus builders (BootCat, Web Concordancer), online texts can be found through specialised search engines (e.g. Buscopio), collected from text repositories (e.g. Medline, Scirus) or simply saved from relevant web pages. In contrast, the approach termed Web as Corpus (WaC) (ibid.) consists of searching the Web as a corpus in its own right. This involves using search engines, which may be rather general (e.g. Google) or designed to fulfil more specific needs (e.g. online WaC query resources, such as WebCorp and KwicFinder).

With the WfC and WaC approaches alike, users will, sooner or later, have to enter search words (or 'seeds' or 'seed words', as some authors refer to them – Baroni & Bernardini 2004) to either create an ad hoc corpus or retrieve relevant information from one.

Working with electronic texts and IT tools for searching and analysing them allows students and other users to compile large corpora. A note of caution must be sounded, however, as corpus size should not be an absolute criterion for quality in translator education. When building and working with corpora in that context, the emphasis should not be on merely accumulating texts to obtain a very large corpus, but on using corpora in a meaningful, well oriented, rewarding manner, so that students can see the benefits their use entails.

> La compilación de corpus *ad hoc* para la docencia en traducción e interpretación resulta especialmente útil, por cuanto supone una fuente de documentación rápida, económica y fiable para conocer la materia, las unidades terminológicas y fraseológicas de un ámbito especializado dado, así como el índice de variación formal y conceptual de las unidades terminológicas de la especialidad de acuerdo con la situación comunicativa (cf. Cabré Castellví 1999). En este sentido, el corpus *ad hoc* constituye una

> *macrofuente* de documentación, en tanto fuente gramatical y discursiva, lexicográfica, terminológica y cognitiva-especializada (Corpas Pastor 2001: 173).

As shown, regardless of the term used to refer to them, comparable corpora have great potential for providing information. The key is knowing how to obtain that information. Several authors (Zanettin 2001; Bowker & Pearson 2002; López Rodríguez 2002; Sánchez-Gijón 2003; Wilkinson 2005a, Rodríguez-Inés 2008; and Gatto 2009; among others) have suggested ways of extracting different types of information from comparable corpora. Bowker (2000: 22) is of the opinion that the greatest challenge involved in using a target-language specialised corpus lies in designing a method for searching for an unknown TL term in cases in which the corresponding source language (SL) term cannot be used as an access point.

I shall now proceed to set out a number of strategies I have found highly useful when building target-language specialised corpora and subsequently extracting relevant information from them for translation purposes. Where appropriate, illustrative examples corresponding to direct and/or inverse translation will be provided.

5. Strategies for building a target-language specialised corpus

5.1. Aim: Obtaining TL key terms for finding parallel texts

5.1.1. *Strategy 1: Use terms and vocabulary present in the ST*

Procedure:

1. Extract a frequency word list from the ST.
2. Extract concordances or clusters of key terms or terms that appear frequently.
3. Look at the co-text of the ST terms.
4. Identify words whose translation equivalent is in no doubt.
5. Enter the TL equivalents in a specialised search engine (e.g. Buscopio), text repository (e.g. Medline, Scirus) or corpus builder (e.g. BootCat, Web Concordancer).

Example: Finding the English translation equivalent of the Spanish medical term *habón* may cause difficulties. However, a ST in which the term appears may offer relevant information, such as the fact that an *habón* is a type of raised skin lesion associated with other terms such as *edema*, *urticaria* and *hipersensibilidad*. The translations of the three terms in question are straightforward (*oedema*, *urticaria*, and *hypersensitivity*) and could be entered as search terms for finding parallel texts containing the English equivalent of *habón*.

5.1.2. *Strategy 2: Use basic knowledge of ST topic and genre*
Procedure:

1. Perform an initial search for parallel texts (same topic and/or genre as the ST) in the TL.
2. Extract keywords from the parallel texts identified to subsequently identify a new set of keywords.
3. Use the new keywords to perform a second search for parallel texts.

5.1.3. *Strategy 3: Collect TL texts listed in the ST bibliography section (if included)*
Procedure: Collect the TL texts on which the ST is based, as they will contain key terminology and phraseology and even full quotes in the TL.

5.1.4. *Strategy 4: Examine ST bibliographic references (if included)*
Procedure: Use TL words or phrases that appear in the bibliography section (e.g. in titles) to search for parallel texts. Such texts may solve initial conceptual or terminological problems.

Additionally, if using a search engine such as Google to mine the Web for documents, it is highly advisable to:

- Specify a language in the search criteria (e.g. English, to retrieve documents in the language in question).
- Restrict the domain in the search query (e.g. geographical domain: .uk or .es; other domains: .ac or .edu, denoting academic websites). For instance, entering the search string *"on-site courses" site:ac.uk* and then *"face-to-face courses" site:ac.uk* in Google would indicate the relative frequency of use of the two terms in the UK, and the comparison would have a solid basis since all the results would come from British academic websites. Another example, in this case to observe the use of the academic term *MOMD* (Modules Outside the Main Discipline), would consist of searching for *"MOMD" site:ac.uk*. The results would clearly show that the term is exclusively used by the University of Birmingham.

The domain restriction strategy can also be used to search for documents within a specific website. For instance, entering the search string *electives site:http://www.manchester.ac.uk* would make it possible to observe the use made of the term in question by the University of Manchester only. Similarly, searching for *"free-choice credits" site:http://www.uab.cat* would show how this term has been used in the English version of the Universitat Autònoma de Barcelona's website.

6. Strategies for searching for translation equivalents in a target-language specialised corpus

6.1. Aim: Identifying important terms

6.1.1. *Strategy 1: Focus on the terms that occur most frequently in a field*
Procedure: Extract a frequency wordlist to gain an insight into a specialised field's most common or important terms.

6.1.2. *Strategy 2: Focus on the terms that are specific to a field*
Procedure: Extract a keyword list by comparing wordlists from a specialised corpus and a general language corpus to identify words that are specific to the relevant specialised field.

6.2. Aim: Investigating a linguistic or conceptual hypothesis

6.2.1. *Strategy 1: Use cognates*
Procedure: Extract concordances of TL words that are morphologically similar to SL words.
Example: The English verb *stipulate* and the Spanish verb *estipular* are cognates. It is necessary, however, to be aware of the potential pitfalls of false friends when using this strategy.

6.2.2. *Strategy 2: Combine the word that is the subject of a hypothesis with others that appear in its context*
Procedure: Extract concordances of a SL term whose translation equivalent is clear in combination with one or more TL words to obtain the translation equivalent of a complex term.
Example: In the case of the SL unit *trim kit: dimensions*, a term used in relation to microwave ovens, it could be hypothesised that if we were working from English into Spanish, the word *kit* might be present in a TL corpus containing texts on home appliances. On that basis, *kit* could be entered as a search word in combination with *dimensiones* as a context word.

6.3. Aim: Selecting an equivalent from various options

6.3.1. *Strategy 1: Try spelling variants*
Procedure: Extract concordances of a term with and without a dash, upper case letters, an accent, etc.
Example: Entering the string *real*time ultrasonography* in a TL medical corpus

will show that *real-time* tends to be written with a dash when used as an adjective. The same applies to *first-trimester, second-trimester* and *third-trimester* when those terms are used adjectivally.

6.3.2. *Strategy 2: Use context words in searches*
Procedure: Extract concordances of two words or expressions that seem to be synonymous by entering one of them as a context word to see if both appear in the same context and how each is used.
Example: Enter *restriction* and *retardation* in a corpus of medical texts on fetal anomalies. Although the terms *fetal growth restriction* and *fetal growth retardation* have the same meaning, political correctness is causing the use of the latter to decrease.

6.3.3. *Strategy 3: Use wildcards (e.g. asterisks) to truncate searches*
Procedure: Enter a truncated search and extract concordances to obtain various possible solutions and analyse any differences there may be between them.
Example: A search for *chromosom* aberration* in a corpus of medical texts will show cases of *chromosome/-al/-ic aberration*.

6.3.4. *Strategy 4: Check the usage of synonyms and near-synonyms*
Procedure: Extract concordances of words that are synonymous or appear to be so to see if they really convey the same meaning, are used in the same way (i.e. are interchangeable) or have a hierarchical relationship.
Example: *Stock exchange* and *bourse* have the same meaning but are not 100% interchangeable, as the latter term is used in a more restricted range of contexts.

6.3.5. *Strategy 5: Check word order*
Procedure: Combine search words by entering them in different orders, entering one of them as a context word, or entering one of them and then resorting the context.
Example: Search for *merger** in an English corpus of financial texts. Doing so will show that the most common phrase and word order among the results is *mergers and X* (e.g. *mergers and acquisitions*, *mergers and takeovers*).

6.4. Aim: Identifying standard phraseology for a topic or genre.

6.4.1. *Strategy 1: Look at word clusters*
Procedure: Extract word clusters of different sizes.
Example: Extracting clusters containing the word *información* from a corpus of Spanish instruction manuals will provide phrases such as *Para más/mayor*

información. . .; Si desea/s / necesita/s más información, contacte/contacta con . . .

6.4.2. *Strategy 2: Observe collocations*
Procedure:
1. Extract a search word's collocations.
2. Observe the collocations, bearing in mind that collocates might appear several words before or after a search word rather than immediately preceding or following it.

Example: Observing the collocates of the noun *tax* in a corpus of websites and leaflets on mortgages will show that one of the verbs that collocate with it is *levy* (as in *Plusvalía is a tax that is levied by the local council. . .*).

6.5. Aim: Identifying a text's register

6.5.1. *Strategy 1: Look at the words that appear most frequently in the text*
Procedure:
1. Extract a frequency wordlist from the TL corpus.
2. Extract concordances of words that might be indicative of register.

Example: In a Spanish corpus (of instruction manuals, for instance), look at subject and personal pronouns (e.g. *tú*/*usted*, *tu*/*su*), verb endings (e.g. *enciende*/*encienda*) or other elements that might provide clues (e.g. conditional *si* as part of expressions such as *si se introduce* or *si introduce/s*) as to the most common way of addressing the reader in a particular genre.

Example: In an English corpus of medical texts, one option would be to see whether the terms used are of Greek-Latin or Anglo-Saxon origin (e.g. *stratum* or *layer*, *gestation* or *pregnancy*). Terms of the former type usually denote a higher register (e.g. genres, such as research articles, in which expert-to-expert communication takes place), whereas the latter kind tends to appear in genres intended for a broader readership (e.g. patient information leaflets).

6.6. Aim: Investigating ideology and semantic prosody

6.6.1. *Strategy 1: Look at the context and co-text of a word or phrase*
Procedure:
1. Extract concordances of a given word or phrase.
2. See if the words surrounding the search word or phrase have positive or negative connotations.

Example: Despite meaning *modernisation*, the Spanish word *reconversión* usually entails job losses.

6.7. Aim: Retrieving possible equivalents for 'unknown' terms

6.7.1. *Strategy 1: Examine the ST bibliography section (if included)*
Procedure:
1. Extract concordances of TL words that appear in the ST bibliography section.
2. See if there are other related TL words that are equivalent to other ST key terms or could at least be useful when performing new searches.

6.7.2. *Strategy 2: Perform truncated searches (e.g. using an asterisk)*
Procedure: Search for fragments of relevant words or expressions using truncation.
Example: Searching for *pre** (based on the formula *prefix/root of word + asterisk*) in an English corpus of medical texts will show how the uses of *prebirth* and *prenatal* differ.
Example: Searching for **00* (based on the formula *asterisk + last part of a number*) in a Spanish corpus of technical descriptions of household appliances will show the most conventional way of expressing power ratings (*X Watt* / *X Watts* / *X vatios* / *X W.*), although a certain degree of noise should be expected in a case like this.

6.7.3. *Strategy 3: Use known equivalents of related terms or of parts of complex terms*
Procedure: Extract concordances, collocates or clusters of known equivalents of SL terms related to the one being searched for, or of known equivalents of parts of SL complex terms.
Example: An approach to finding an equivalent of the Spanish medical term *longitud cráneo-nalga* in an English corpus of medical texts might consist of entering the word *length* (which seems almost certain to be part of the TL equivalent) and looking at combinations present in the TL corpus (a corpus of medical abstracts related to fetal monitoring techniques will not contain many possible patterns: e.g. *ear length*, *humerus length*, *telomere length*, *crown-rump length*, etc.). Observation, common sense and verification in a different resource will confirm that the last of the aforementioned terms is the equivalent sought.

6.7.4. *Strategy 4: Use knowledge of syntactic structures*
Procedure: Search for likely TL syntactic structures rather than possible equivalent words.
Example: Search for * *a* * in a Spanish corpus if translating an English expression with the structure * *by* * or * *to* * (e.g. *step by step* → *paso a paso*; *face to face* → *cara a cara*).

6.7.5. *Strategy 5: Search for abbreviations or acronyms or their extended forms*
Procedure: Enter an acronym, an abbreviation or a word that is part of the extended form of an acronym or abbreviation. In some cases, the SL item will be retained in the TL.
Another option consists of searching for brackets ((*)) in the corpus, although such a search might generate a considerable amount of noise.

6.7.6. *Strategy 6: Search for proper nouns or product names*
Procedure: If looking for equivalents of words positioned close to proper nouns, search the TL corpus for names (of companies, people, international organisations, etc.) as they appear in the ST, as their SL form is likely to remain unchanged in TL texts (the degree of likelihood can vary from genre to genre).

Conclusion

This paper has set out to show that exploiting corpora in translation and translator education can be beneficial. Furthermore, both building and using electronic corpora can enhance the student's learning experience. As stand-alone resources or in conjunction with others, such as dictionaries, corpora can provide a wealth of information on various aspects of a language (e.g. phraseology, terminology, concepts, usage, etc.).

As Maher, Waller & Kerans (2008) observe in relation to specialised translation,

> Using corpora to guide translation or editing work is a way to compensate for any or all of the following: a) uneven field knowledge; b) non-contact with language genres and registers outside our normal range of use; and c) source language interference from lack of contact with our native language. In general terms, using corpora can help us mature as specialist language users (Maher, Waller & Kerans 2008: 71).

It has been stated that the greatest challenge that the use of target-language specialised corpora entails is retrieving the translation equivalents such resources contain. As Varantola (2002: 180) puts it, "even the search strategies must sometimes be elaborate. If no adequate search string or term springs to mind, the corpus compilers need to think of indirect ways of finding what they are looking for."

In this paper, I have attempted to establish strategies to systematise such indirect approaches to obtaining information from a target-language specialised corpus. While most of the examples in the paper correspond to the English<>Spanish language combination, the strategies described can be applied to other languages and language pairs.

Intuition-driven searches may not always give straightforward, satisfactory results. For example, when trying to come up with an equivalent or a partial equivalent for the medical term *longitud cráneo-nalga*, as featured in a previous example, a user's general English knowledge might suggest equivalents such as *length*, *extension* or *measure* for the Spanish term *longitud*; *cranium*, *skull*, *head* or *top* for *cráneo*; or *buttock* or *bottom* for *nalga*. However, only one of them (*length*) would provide meaningful concordances from which the complete equivalent could be obtained. Corpora do not provide results by magic, and a great deal of trial and error (not to mention observation) is involved. Hopefully, the strategies set out here might help users to exploit corpora more successfully.

Bibliography

Ahmad, Khurshid, Paul Holmes-Higgin, and Syed Sibte Raza Abidi. 1994. "A Description of Texts in a Corpus: 'Virtual' and 'Real' Corpora." In *EURALEX-94: Proceedings*, ed. by Willy Martin *et al.*, 390–402. Amsterdam: Vrije Universiteit.

Aston, Guy. 1999. "Corpus Use and Learning to Translate." *Textus: English Studies in Italy: rivista dell'Associazione italiana di anglistica* 12 (2): 289–314.

Atkins, Sue, Jeremy Clear, and Nicholas Ostler. 1992. "Corpus Design Criteria." *Literary and Linguistic Computing* 7 (1): 1–16.

Baroni, Marco, and Silvia Bernardini. 2004. "BootCaT: Bootstrapping Corpora and Terms from the Web." In *Proceedings of LREC 2004*, 1313–1316. Lisbon: ELDA.

Baroni, Marco *et al.* 2006. "WebBootCaT: Instant Domain-specific Corpora to Support Human Translators." In *EAMT 2006 - 11th Annual Conference of the European Association for Machine Translation. Oslo: The Norwegian National LOGON Consortium and the Departments of Computer Science and Linguistics and Nordic Studies at Oslo University (Norway)*, 247–252. On line at: http://trac.sketchengine.co.uk/wiki/WikiStart (Access date: 12 June 2011).

Bowker, Lynne. 1996. "Towards a Corpus-based Approach to Terminography." *Terminology* 3 (1): 27–52.

Bowker, Lynne. 2000. "Towards a Methodology for Exploiting Specialized Target Language Corpora as Translation Resources." *International Journal of Corpus Linguistics* 5 (1): 17–52.

Bowker, Lynne, and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge. 242 pp.

Buendía-Castro, Miriam, and Clara Inés López-Rodríguez. 2010. "The Web for Corpus and the Web as Corpus in Translator Training." In *Proceedings of the International Symposium on 'Using Corpora in Contrastive and Translation Studies*, ed. by Richard Xiao. On line at: http://www.lancs.ac.uk/fass/projects/corpus/UCCTS2010Proceedings/papers/buendialopez.pdf (Access date: 12 June 2011).

Cabré Castellví, María Teresa. 1999. "Fuentes de información terminológica para el traductor." In *Técnicas documentales aplicadas a la traducción*, ed. by María Pinto, and José Antonio Cordón, 19–40. Madrid: Síntesis.

Church, Kenneth, William Gale, Patrick Hanks, and Donald Hindle. 1991. "Using Statistics in Lexical Analysis." In *Lexical Acquisition: Exploiting Online Resources to Build a Lexicon*, ed. by Uri Zernik, 115–164. Hillsdale: Lawrence Erlbaum Associates.

Corpas Pastor, Gloria. 2001. "Compilación de un corpus *ad hoc* para la enseñanza de la traducción inversa especializada." *Trans* 5: 155–184. On line at: http://www.trans.uma.es/Trans_5/t5_155–184_GCorpas.pdf (Access date: 3 July 2011).

Corpas Pastor, Gloria. 2002. "Traducir con corpus: de la teoría a la práctica." In *Texto, terminología y traducción*, ed. by Joaquín García Palacios, and María Teresa Fuentes, 189–226. Salamanca: Almar.

Corpas Pastor, Gloria, and Encarnación Postigo Pinazo. 2002. "Aplicaciones del corpus para la redacción en inglés de textos científicos originales o traducidos: a propósito de la sigla *ACTH*." In *Estudio del léxico: análisis y docencia*, ed. by María Dolores, and Fernández de la Torre Madueño, 39–71. Málaga: Servicio de Publicaciones de la Universidad de Málaga.

De Schryver, Gilles-Maurice. 2002. "Web for/as Corpus: A Perspective from the African Languages." *Nordic Journal of African Studies* 11 (2): 266–282. On line at: http://www.njas.helsinki.fi/pdf-files/vol11num2/schryver.pdf (Access date: 3 July 2011).

Friedbichler, Ingrid, and Michael Friedbichler. 2000. "The Potential of Domain-specific Target-language Corpora for the Translator's Workbench." In *I corpora nella didattica della traduzione: Corpus Use and Learning to Translate*, ed. by Silvia Bernardini, and Federico Zanettin, 107–116. Bologna: CLUEB.

Gatto, Maristella. 2009. *From Body to Web. An Introduction to the Web as Corpus*. Bari: Università degli Studi di Bari University Press on-line. 180 pp.

Gavioli, Laura. 1996. "Corpus di testi e concordanze: Un nuovo strumento nella didattica delle lingue straniere" [Text corpora and concordances: A new tool for foreign language teaching]. *Rassegna Italiana di Linguistica Applicata* 2: 121–146.

Hultsijn, Jan H. 1992. "Retention of Inferred and Given Word Meanings: Experiments in Incidental Vocabulary Learning." In *Vocabulary and Applied Linguistics*, ed. by Pierre J. L. Arnaud, and Henri Bejoint, 113–125. London: Macmillan.

Jääskeläinen, Riitta, and Anna Mauranen. 2004. "Translators at Work: A Case Study of Electronic Tools Used by Translators in Industry." In *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, ed. by Geoff Barnbrook, Pernilla Danielsson, and Michaela Mahlberg, 49–53. London: Continuum.

López Rodríguez, Clara Inés. 2002. "Training Translators to Learn from News Report Corpora: The Case of Anglo-American Cultural References." In *Training the Language Services Provider for the New Millennium*, ed. by Belinda Maia, Johann Haller, and Margherita Ulrych, 213–222. Porto: AstraFlup.

López-Rodríguez, Clara Inés, and María Isabel Tercedor-Sánchez. 2008. "Corpora and Students' Autonomy in Scientific and Technical Translation Training." *Journal of Specialised Translation* 9: 2–19. Online at: http://www.jostrans.org/issue09/art_lopez_tercedor.php (Access date: 12 June 2011).

Maher, Ailish, Stephen Waller, and Mary Ellen Kerans. 2008. "Acquiring or Enhancing a Translation Specialism: The Monolingual Corpus-guided Approach." *The Journal of Specialised Translation* 10. On line at: http://www.jostrans.org/issue10/art_maher.pdf (Access date: 12 June 2011).

Maia, Belinda. 1997. "Do-it-yourself Corpora. . . with a Little Bit of Help from Your Friends." In *PALC'97: Practical Applications in Language Corpora*, ed. by Barbara Lewandowska-Tomaszczyk, and Patrick James Melia, 403–410. Lodz: Lodz University Press.

Maia, Belinda. 2000. "Making Corpora – A Learning Process." In *I corpora nella didattica della traduzione: Corpus Use and Learning to Translate*, ed. by Silvia Bernardini, and Federico

Zanettin, 47–59. Bologna: CLUEB.
Maia, Belinda. 2003. "Some Languages are More Equal than Others: Training Translators in Terminology and Information Retrieval Using Comparable and Parallel Corpora." Zanettin, Bernardini and Stewart 2003. 43–70.
Oster, Ulrike. 2007. "Working towards Autonomy: Corpora in the Translation Classroom." In *Quo vadis Translatologie? Ein halbes Jahrhundert universitäre Ausbildung von Dolmetschern und Übersetzern in Leipzig. Rückschau, Zwischenbilanz und Perspektiven aus der Außensicht*, ed. by Gerd Wotjak, 311–326. Berlin: Timme.
Pearson, Jennifer. 1996. "Electronic Texts and Concordances in the Translation Classroom." *Teanga* 16: 85–95.
Rodríguez-Inés, Patricia. 2008. *Uso de corpus electrónicos en la formación de traductores (inglés-español-inglés)*. Ph.D thesis, Universitat Autònoma de Barcelona. 727 pp.
Rodríguez-Inés, Patricia. 2010. "Electronic Corpora and Other ICT (Information and Communication Technologies) Tools: An Integrated Approach to Translation Teaching." *The Interpreter and Translator Trainer* 4 (2): 251–282.
Sánchez-Gijón, Pilar. 2003. *Els documents digitals especialtizats: utilització de la lingüística de corpus com a font de recursos per a la traducció especialitzada*. Ph.D thesis, Universitat Autònoma de Barcelona.
Scott, Juliette R. 2011. *DIY Corpora: a pearl in the legal translator's sea of tools. Initial testing of a methodology to enhance the correspondence of legal translations with receiver expectations.* MA dissertation, University of Portsmouth. Abstract on line at: http://eprints.port.ac.uk/2161/ (Access date: 12 June 2011).
Varantola, Krista. 2000. "Translators, Dictionaries and Text Corpora." In *I corpora nella didattica della traduzione: Corpus Use and Learning to Translate*, ed. by Silvia Bernardini, and Federico Zanettin, 117–133. Bologna: CLUEB.
Varantola, Krista. 2002. "Disposable Corpora as Intelligent Tools in Translation." *Cadernos de Tradução: Corpora e Tradução* 1 (9): 171–189. On line at: http://www.cadernos.ufsc.br/online/9/krista.htm (Access date: 12 June 2011).
Varantola, Krista. 2003. "Translators and Disposable Corpora." Zanettin, Bernardini and Stewart 2003. 55–70.
Varantola, Krista. 2007. "The Contextual Turn in Learning to Translate." In *Lexicography, Terminology, and Translation. Text-based Studies in Honour of Ingrid Meyer*, ed. by Lynne Bowker, 215–226. Ottawa: University of Ottawa Press.
Wilkinson, Michael. 2005a. "Using a Specialized Corpus to Improve Translation Quality." *Translation Journal* 9 (3). On line at: http://accurapid.com/journal/33corpus.htm (Access date: 12 June 2011).
Wilkinson, Michael. 2005b. "Discovering Translation Equivalents in a Tourism Corpus." *Translation Journal* 9 (4). On line at: http://accurapid.com/journal/34corpus.htm (Access date: 12 June 2011).
Zanettin, Federico. 2001. "Swimming in Words: Corpora, Translation, and Language Learning." In *Learning with Corpora*, ed. by Guy Aston, 177–197. Bologna: CLUEB.
Zanettin, Federico. 2002. "DIY Corpora: The WWW and the Translator." In *Training the Language Services Provider for the New Millennium*, ed. by Belinda Maia, Johann Haller, and Margherita Ulrych, 239–248. Porto: AstraFlup. On line at: http://1086820439574408617-a-1802744773732722657-s-sites.googlegroups.com/site/federicozanettinnet/dbpublications/DIYcorpora.TheWWWandthetranslator.pdf (Access date: 12 June 2011).

Zanettin, Federico, Silvia Bernardini, and Dominic Stewart (eds). 2003. *Corpora in Translator Education*. Manchester: St. Jerome. 156 pp.

Search engines, text repositories and software

AntConc < http://www.antlab.sci.waseda.ac.jp/antconc_index.html> (accessed 28 June 2011).
BootCat <http://bootcat.sslmit.unibo.it/> (accessed 28 June 2011).
Buscopio <http://buscopio.net/> (accessed 28 June 2011).
Flashgot <https://addons.mozilla.org/es-es/firefox/addon/flashgot/> (accessed 28 June 2011).
Free download manager <http://www.freedownloadmanager.org/> (accessed 28 June 2011).
Google <http://www.google.com/> (accessed 28 June 2011).
KwicFinder <http://www.kwicfinder.com/KWiCFinder.html> (accessed 28 June 2011).
Medline <http://www.ncbi.nlm.nih.gov/pubmed/> (accessed 28 June 2011).
Scirus <http://scirus.com/> (accessed 28 June 2011).
Web Concordancer <http://webascorpus.org/searchwac.html> (accessed 28 June 2011).
WebCopier <http://webcopier.softonic.com/> (accessed 28 June 2011).
WebCorp <http://www.webcorp.org.uk/> (accessed 28 June 2011).
WordSmith Tools < http://www.lexically.net/wordsmith/index.html (accessed 28 June 2011).

Abstract

The use of monolingual and bilingual electronic corpora in translation training has steadily increased over the last decade due to the advantages they entail in relation to other resources. Whether as a source of teaching materials or as a reference source in their own right, corpora have proved beneficial for the teaching of general and specialised direct and inverse translation. This article will focus on monolingual target-language specialised corpora, which are one of the types of corpora most widely used in translation training, along with parallel corpora and bilingual comparable corpora. Target-language specialised corpora are *technically easy* to compile and may be used with minimal processing, but pose a challenge in terms of the extraction of translation equivalents. A number of strategies for building and exploiting such corpora will be presented in the article.

Keywords: translator education, target-language specialised corpora, advantages, search strategies, unknown equivalents

Résumé

Ces dix dernières années, l'utilisation de corpus électroniques à la fois monolingues et bilingues est de plus en plus fréquente dans la formation de traducteurs, notamment dû au fait que ceux-ci présentent de nombreux avantages par rapport aux autres ressources. Que ce soit comme matériel d'enseignement ou bien comme outil de référence, les corpus se sont révélés très utiles pour l'enseignement de la traduction et de la traduction spécialisée (thème et version). Cet article

présente une étude sur les corpus monolingues spécialisés en langue cible qui forment l'un des types de corpus les plus fréquemment employés dans la formation des traducteurs, avec le corpus parallèle et le corpus comparable bilingue. Bien que la constitution et l'exploitation d'un corpus spécialisé en langue cible soit relativement simple *à un niveau technique*, celui-ci peut s'avérer plus compliqué vis-à-vis de l'extraction d'équivalents de traduction. Dans cet article, quelques stratégies seront présentées pour la bonne constitution et exploitation de ce type de corpus.

Mots clés: formation de traducteurs, corpus monolingue spécialisés en langue cible, stratégies de recherche, équivalents inconnus

About the author

Patricia Rodríguez-Inés is a lecturer in the Faculty of Translation and Interpreting of the Universitat Autònoma de Barcelona in Spain. Her research interests include corpus linguistics applied to translation, translation teaching and translation competence acquisition. Her PhD thesis, entitled 'Using electronic corpora in translation teaching', won a national prize in 2009 and will be published on line shortly. She teaches general and specialised translation between English and Spanish as well as corpus methodology at her faculty. She has published extensively with PACTE, the research group she belongs to (http://grupsderecerca.uab.cat/pacte/en) and individually (e.g. featured article in *The Interpreter and Translator Trainer* 4: 2).

Address: Despatx K-1014, Facultat de Traducció i d'Interpretació, Campus Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain
E-mail: patricia.rodriguez@uab.es