Corpora in Translation Practice

Federico Zanettin

Università per Stranieri di Perugia
Palazzo Gallenga, Piazza Fortebraccio, 4 - Perugia
zanettin@unistrapg.it

Abstract

The aim of this paper is to trace links between work in the corpus linguistics community and the world of practicing translators. The relevance to translation work of corpora in general, and bilingual and parallel corpora in particular, is evaluated by comparing corpora and translation memories and by drawing an analogy between different types of corpora and more traditional reference tools, i.e. dictionaries. Corpus resources available to translators are placed along a cline going from "robust", stable corpora (e.g. large reference corpora such as the BNC) to "virtual", ephemeral corpora (e.g. DIY web corpora). Finally, a few suggestions are put forward in order to encourage a wider diffusion of corpora and concordancing software among professional translators.

1. Introduction

The translator's workplace has changed dramatically over the last ten years or so, and today the computer is undoubtedly the single most important tool of the trade for a translator regardless of whether he or she is a literary translator working for a small publisher, a technical translator working for a translation agency or a legal translator. Today, translators compose their texts on the computer screen, often receive their source texts in electronic format and sometimes their translations will only live as digital information as in the case of web site localization.

The specific hardware and software resources individual translators will resort to will vary depending on the task to be done. While in the case of most literary translators the translated text will probably take shape by means of a general purpose word processor, in the case of technical translators the target text will be produced with the help of the most sophisticated "translator workbench", equipped with all sorts of CAT tools, translation memory and terminology systems, and localization software.

The computer has also flanked, if not substituted, other technological supports in providing access to traditional tools and resources. Translation aids such as monolingual and bilingual dictionaries, terminologies and encyclopedias are now available not only on paper but also in electronic format. Colleagues and expert informants can now be consulted via e-mail and newsgroups besides via telephone, fax and face-to-face encounters. The storage capacity and processing power of personal computers have made access to linguistic and content information easier and quicker than ever before, and the Internet has opened up highways of communication and information retrieval. The problem is now not finding a piece of information, but finding the right and reliable piece of information without wasting too much time.

Corpora and concordancing software can be a way of gaining access to information about language, content, and translation practices which was hardly available to translators before the present stage of ICT development. Corpora and corpus analysis software have been around for quite a long time, but their use is only now beginning to extend beyond a restricted segment of language professionals, such as lexicographers, language engineers, as well as linguists in educational and training institutions.

I would like to suggest that corpora and concordancing software could find a larger place in the translator computerised workstation, and that more corpus resources could and should be made more accessible to professional translators. In order to do so, however, corpus builders and software producers should take into account the specific needs of this group of users. Learning to use corpora as translation resources should also be part of the curriculum of future translators and become part of their professional competence.

2. Corpora and translation

According to the EAGLES text typology elaborated by John Sinclair (1996) we can make a general distinction between Monolingual and Multilingual (including Bilingual) corpora. As regards bilingual (and multilingual) corpora a further distinction can be made between Comparable corpora (corpora compiled using similar design criteria but which are not translations) and Parallel, or Translation Corpora, which are texts in one language aligned with their translation in another. This picture can be further complicated by involving variables such as direction and directness[1] of translation, number of languages, number of translations per text, etc., producing bi-directional, reciprocal, control, star and diamond corpus models (cf. Johansson, forthcoming; Teubert, 1996; Zanettin, 2000; Malmkiaer, forthcoming). Still another type of translation related corpus is the Monolingual Comparable Corpus (Baker, 1993), or a corpus composed of two sub-sections, one of texts originally composed in one language and the other of texts translated into that same language (from a number of other languages). This type of corpus, however, while

[1] (i.e. whether a translation is produced directly from the original text or via an intermediate translation in another language).

undoubtedly an extremely useful tool for translation theorists, researchers and students, is arguably of less immediate relevance for professional translators dealing with actual translation jobs.

Professional translators working in the technical sector are perhaps more familiar with the parallel concordancing feature of translator memory systems. A translation memory is data bank from which translators automatically retrieve fragments of past translations that match, totally or to a degree, a current segment to be translated, which must match, totally or to a degree - an already translated segment. But it can also be seen as a parallel corpus which translators manually query for parallel concordances of (already translated) specific terms or patterns. Aligned translation units are conveniently displayed on the screen, offering the translator a range of similar contexts from a corpus of past translations. A translation memory is, however, a very specific type of parallel corpus in that:

a) it is "proprietory": TMs are created individually or collectively around specific translation projects. They are highly specialized and very useful when used for the translation or localization of program updates – indeed that is their origin – but are not much help when starting a new translation project on a different topic or text type.
b) TMs tend to closure, to progressively standardize and restrict the range of linguistic options. This may be an advantage from the point of view of terminological consistency and of processing costs for clients or translation agency managers, but is often detrimental for readability (texts translated using a "Workbench" can become very repetitive) and the translators eyesight (translators using a well-known Workbench often testify to a "yellow-and-blue-eye-syndrome).

Translation workbenches and translation memories have indeed become the most successful technological product to be created for professional translators, but – as it often happens with MT products – their use is best limited to specific text types, such as online help files, manuals and all types of reference work which do not require sequential reading and for which the scope of translation can be limited to the sentence of phrase level (and thus left to a machine). When dealing with other types of texts translators are perhaps better off with a different kind of language resource, i.e. the type of corpora which are more familiar to lexicographers and linguists and which are only now beginning to enter the selection of tools available to professional and trainee translators.

3. Corpora as translation aids

The respective potential uses on the part of professional translators of monolingual target corpora, bilingual comparable corpora, and of parallel corpora can be illustrated drawing an analogy with other respected tools of the trade, i.e. dictionaries: Monolingual target corpora can be compared to monolingual target language dictionaries, and comparable source corpora to monolingual source language dictionaries. While dictionaries favor a synthetic approach to lexical meaning (via a definition), corpora offer an analytic approach (via multiple contexts).[2] Translators can use target monolingual corpora alongside target monolingual dictionaries to check the meaning and usage of translation candidates in the target contexts. Like source language dictionaries, source language corpora can be consulted for source text analysis and understanding. Large reference corpora (BNC, CORIS/CODIS, etc.) can function as general dictionaries, while smaller, specialized and bilingual comparable corpora can be seen as analogous to specialized monolingual dictionaries (either or both in the source and in the target language).

Parallel corpora can instead be compared to bilingual dictionaries, with a few important differences: bilingual dictionaries are repertories of lexical equivalents (general dictionaries) or terms (specialized dictionaries and terminologies) established by dictionaries makers which are offered as translation candidates. Parallel corpora are repertoires of strategies deployed by past translators, as well as repertoires of translation equivalents. In selecting a translation equivalent from a general bilingual dictionary a translator has to assess the appropriateness of the candidate to the new context by starting from a definition and a few usage examples. A parallel corpus will offer a repertoire of translation strategies past translators have resorted to when confronted with similar problems to the ones that have prompted a search in a parallel corpus.

Parallel corpora can provide information that bilingual dictionaries do not usually contain. They can not only offer equivalence at the word level, but also non-equivalence, i.e. cases where there is no easy equivalent for words, terms or phrases across languages. A parallel corpus can provide evidence of how actual translators have dealt with this lack of direct equivalence at word level. For example, in the translations by two different Italian translators of a number of novels by Salman Rushdie (Zanettin, 2001b), the word "edges", which usually collocates with a preposition, as in the phrases "around the edges," or "at the edges," was never translated literally, but rather omitted:

1. …biting the skin around the edges of a nail…
   …*mordicchiandosi la pelle attorno all'unghia…*
2. …around the edges of Gibreel Farishta's head…
   …*intorno alla testa di Gibreel Farishta…*
3. …around the edges of the circus-ring…
   …*intorno alla pista da circo…*
4. …and there was a fluidity, an indistinctiness, at the edges of them…
   …*vicinissime a loro c'erano una fluidità e un'indeterminatezza…*
5. …the horses grew fuzzy at the edges…
   …*i cavalli diventavano sempre più sfocati…*

[2] So-called "production dictionaries", which focus on usage information, can be thought of as standing somehow in between the two.

6. …blurred at the edges, my father…
   …*con la mente annebbiata, mio padre…*
7. …looking somewhat ragged at the edges…
   …*con l'aria di un uomo distrutto…*
8. …Mrs Qureishi, too, was beginning to fray at the edges…
   …*anche Mrs Qureishi si stava consumando…*

In all these cases, the two professional translators have consistently chosen to resort to "zero-equivalence", which being a translation strategy rather than a case of comparative linguistic knowledge would be hardly reported in any bilingual dictionary.

4. Corpus resources for translators

Not all dictionaries are the same, nor are all corpora. Apart from translation memories, corpus resources which are of potential use for professional translators could be classified along a scale which goes from "robust" to "virtual." A "corpus" is a collection of electronic texts assembled according to explicit design criteria which usually aim at representing a larger textual population. "Robust" corpora are ready-made corpora created and distributed by the research community and the language industry on CD-ROM or accessible through the Internet. Prototypical examples are large reference national corpora, such as the *British National Corpus* (BNC) for British English, and the *Dynamic Corpus or Written Italian* (CORIS/CODIS) for Italian. This type of resource, which requires a large building effort, is only now becoming available to the wider public outside the (corpus) linguistics community, and will probably require some "customisation" effort in order to become more widespread among language services providers.

Parallel corpora are usually smaller and even less available to the general public than monolingual corpora. Their construction requires more work than that of monolingual corpora. Among other factors, text pairs (rather than single texts) have to be located and before they can be used they need to be aligned, at least at the sentence level (cf. Véronis, 2000).

There are of course varying degrees of robustness, according to the effort and care which has been put in achieving a balanced and representative selection of texts, in providing explicit linguistic and extralinguistic information (corpus annotation) and the means (the software) to query the corpus for that information (McEnery & Wilson, 1996). Corpus design criteria also vary according to the purpose for which a corpus is built, e.g. a comparable monolingual corpus for descriptive translation research. In this sense, the less "robust" (i.e. the more "virtual") corpora are the most truly professional type, with reference to translators, since they are "rough-and-ready" products created for a specific translation project. A distinction is usually made by corpus linguists between "corpora" and "archives" of electronic texts. An "archive" is simply a repository of electronic texts: In this sense the WWW is an immense (multimedia) text archive. Virtual or "disposable" corpora are created by a translator using the WWW as a source "archive". The WWW and HTML documents need not to be the only source for small, specialized DIY corpora, and textual archives of various types and targeted to various users (newspapers, collections of laws, encyclopedias, etc.) are available on cd-rom. The WWW is however certainly the most familiar and user friendly environment for translators: it is always available; it is the most comprehensive source of electronic texts, and corpus creation, management and analysis can be a relatively straightforward operation (Austermühl, 2001; Zanettin, forthcoming). Building a corpus of web pages basically involves an information retrieval operation, conducted by browsing the Internet to locate relevant and reliable documents which can then be saved locally and made into a corpus to then be analysed with the help of concordancing software. The additional time required by creating and consulting a corpus is compensated for by saving in other translation-related tasks, such as dictionary consultation (both on paper and electronic), paper documentation (often in the form of "parallel texts", e.g. Williams, 1996), help from experts, and by the fact that the corpus contains information not available elsewhere. Moreover, the effort is rewarded by improving quality in terms of terminological and phraseological accuracy (Friedbichler & Friedbichler, 2000).

A number of studies have reported on experiments in translation and language teaching classes with DIY corpora, either made of "disposable" web pages (e.g. Varantola, 2000, forthcoming; Maia, 1997, 2000, forthcoming; Zanettin, forthcoming; Pearson, 2000) or of texts taken from other electronic sources such as newspapers (Zanettin, 2001a) or magazines (Bowker, 1998) on CD-ROM. Corpora created from sources other than web pages can require more time and effort to be built, and can be more or less "disposable" depending on the size of the translation project and on the resources available to create and manage them.

Reports on the use of corpora by professional translators are fewer: Friedichler & Friedbichler, drawing on their experience as translators of medical texts and trainers of technical translators, suggest that domain-specific target language corpora may usefully complement dictionaries and the Web as resources in the translation process, filling the gap between the two. Jääskläinen and Mauranen (2000) report on an experimental study involving a team of researchers from the University of Savonlinna and a team of professional translators translating for the timberwood industry. The researchers created a corpus from a variety of sources (web sites, PDF documents, etc.) following suggestions from the translators, and then trained them in using concordancing software (*WS Tools*, Scott, 1996) to analyse the corpus. In exchange, the translation team agreed to answer a questionnaire. One of the results of the study was learning that translators often complained that the user-friendliness of the concordancing software was very low. This complaint was seconded by translator trainees in other studies with "disposable" corpora where students, usually working in groups, collected a corpus of HTML documents and used them to help them translate a specific text.

These studies have underlined, nonetheless, the value of corpus building as a way of getting acquainted with the content and terminology of the translation. They have stressed the importance of type and topic of the text to be translated as well as of the target language (some text types, topics, and target languages are better helped with corpora than others) and also of adopting sound criteria in choosing suitable texts for inclusion in the corpus. Most of the corpora in these experiments were target monolingual corpora, though some use of bilingual comparable and even parallel corpora was reported.

The main benefits and shortcoming of DIY corpora may be summed up as follows:

Benefits:

- They are easy to make.
- They are a great resource for content information.
- They are a great resource for terminology and phraseology in restricted domains and topics.

Shortcomings:

- Not all topics, not all text types, not all languages are equally suitable or available.
- The relevance and reliability of documents to be included in the corpus needs to be carefully assessed.
- Existing concordancing software is not well equipped to handle HTML or XML files, i.e. web pages. There are no or few parallel corpora, since while some parallel texts (i.e. source texts + translations) can be found on the Internet, hardly all of them could be included in a parallel corpus designed to provide instances of professional standards (Maia, forthcoming).

DIY web corpora stand midway the WWW itself, which can be used as if it were a corpus and robust, "proper" corpora. As for the Web, a "quasi-concordance" view of documents indexed and retrieved is provided by such as search engines Google (http://www.google.com) or Copernic (http://www.copernic.com). Corpus linguistics-oriented software currently being constructed for browsing the WWW as a corpus, such as *KwicFinder* (Fletcher, 2001) and *WebConc* (Kilgarriff, 2001), will certainly prove a useful tool for translators among other language professionals. However, while this "web as corpus" approach has certainly advantages in terms of time over DIY web corpora (the "corpus" is always already there), it necessarily looses in precision and reliability.

The advantages of "robust" corpora over "virtual" corpora can instead be summed up as follows:

- They are usually more reliable.
- They are usually larger.
- They may be enriched with linguistic and contextual information.
- If parallel, they are already aligned.
- They come with user-friendly, customised software (though, again, not necessarily targeted to the needs of professional translators).

5. Conclusions

Translators can tolerate the learning curve necessary to adopt corpora and concordancing software among their everyday working tools only if they derive benefits. These benefits are the fact that corpora provide information not available elsewhere at an affordable cost.

As a way of concluding, I would like to point out possible improvements for existing corpora and concordancing software:

a) "Robust " reference corpora need to become more accessible: for instance, a BNC license is still relatively expensive and the interrogation software might do with some customization; the CORIS/CODIS corpora and others have limited access.

b) In order for "virtual" corpora to become more widespread among translators, concordancing software for work with small monolingual corpora has to become capable of dealing with HTML and, increasingly, XML texts. For example, it may be useful to interface the concordancing software with the Internet browser to provide facilities for file downloading and management, and for allowing the user to switch between concordance lines and full text view, in order to take advantage of multimedia features of electronic texts.

c) Bilingual and parallel corpora are scarcely available and usually of limited size. Bilingual concordancers require bilingual corpora, and given what it takes to locate and align text pairs, it is not very likely that individual translators will resort to consulting parallel concordances unless parallel (aligned) corpora are already available. The creation of more corpora of this kind is a matter of computational resources (especially parallel concordancers and efficient aligning utilities) as well as of more awareness of the usefulness of this resource among translators and language resources providers.

6. References

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

Baker, M. (1993). "Corpus linguistics and translation studies. Implications and applications". In M. Baker, G. Francis & E. Tognini-Bonelli (eds.) *Text and technology*. Philadelphia/Amsterdam: John Benjamins, 233-252.

BNC web site, http://info.ox.ac.uk/bnc

Bowker, L. (1998). "Using specialized monolingual native-language corpora as a translation resource: a pilot study", in *META* 43:4, 631-651.

CORIS/CODIS web site, http://www.cilta.unibo.it

Fletcher, W. (2001). "Concordancing the web with KWiCFinder", presentation given at the *Third North American Symposium on Corpus Linguistics and Language Teaching*, Boston, MA, 23-25 March 2001. Available at http://miniappolis.com/KWiCFinder/Corpus2001.htm.

Friedbichler, I. & Friedbichler, M. (2000). in S. Bernardini & F. Zanettin (eds.) *I corpora nella didattica della traduzione. Corpus Use and Learning to Translate*, Bologna: CLUEB, 107-116.

Jääskeläinen, R. & Maurannen, A. (2000) *Work Package 5: Development of a Corpus on the Timber Industry - Final Report, Project SPIRIT MLIS-programme:*

MLIS-3008 SPIRIT 24637, University of Joensuu, Savonlinna School of Translation Studies.

Johansson, S. (forthcoming). "Reflections on corpora and their uses in cross-linguistic research", in F. Zanettin, S. Bernardini, & D. Stewart (eds.) *Corpora in translator education*.

Kilgarriff, A. (2001). "Web as corpus". In P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds.) *Proceedings of the Corpus Linguistics 2001 conference, UCREL Technical Papers: 13.* Lancaster University, 342-344.

Maia, B. (1997). "Do-it-yourself corpora ... with a little bit of help from your friends!" in B. Lewandowska-Tomaszczyk & P. J. Melia (eds.) *PALC '97 Practical Applications in Language Corpora*. Lodz: Lodz University Press, 403-410.

Maia, B. (2000) "Making corpora: A learning process", in S. Bernardini & F. Zanettin (eds.) *I corpora nella didattica della traduzione*. *Corpus Use and Learning to Translate*, Bologna: CLUEB, 47-60.

Maia, B. (forthcoming) "Training translators in terminology and information retrieval using comparable and parallel corpora", in F. Zanettin, S. Bernardini & D. Stewart (eds.) *Corpora in translator education*.

Malmkiaer, K. (forthcoming). "On a pseudo-subversive use of corpora in translator training", in F. Zanettin, S. Bernardini & D. Stewart (eds.) *Corpora in translator education*.

McEnery, T. & Wilson, A. (1996) *Corpus linguistics*. Edimburgh: Edimburgh University Press.

Pearson, J. (2000). "Surfing the Internet: teaching students to choose their texts wisely". In Lou Burnard and Tony McEnery (eds.) *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt am Maim et al: Peter Lang, 235-239.

Scott, M. (1996). *Wordsmith Tools*. Oxford: Oxford University Press.

Sinclair, J. McH. (1996) *EAGLES Preliminary recommendations on Corpus Typology, EAG--TCWG--CTYP/P*. Online: http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Teuberg, W. (1996) "Comparable or parallel corpora?" *International journal of lexicography*, 9:3, 238-264.

Varantola, K. (2000). "Translators, dictionaries and text corpora" in in S. Bernardini & F. Zanettin (eds.) *I corpora nella didattica della traduzione*. *Corpus Use and Learning to Translate*, Bologna: CLUEB, 117-136.

Varantola, K. (forthcoming). "Translators and disposable corpora" in F. Zanettin, S. Bernardini & D. Stewart (eds.) *Corpora in translator education.*

Véronis, J. (2000) *Parallel text processing. Alignment and use of parallel corpora*. Dordrecht: Kluwer.

Williams, I. A. (1996) "A translator's reference needs: Dictionaries or parallel texts". *Target* 8, 277:299.

Zanettin, F. (2000). "Parallel corpora in translation studies: issues in corpus design and analysis", in Olohan, M. (ed.) *Intercultural Faultlines. Research Models in Translation Studies I. Textual and cognitive aspects*, Manchester: St Jerome. 93-118.

Zanettin, F. (2001a). "Swimming in words: Corpora, translation, and language learning", in G. Aston (ed.) *Learning with corpora*, Bologna/Houston,TX: CLUEB/Athelstan, 177-197.

Zanettin, F. (2001b). *IperGrimus*. In *inTRAlinea* (online) http://www.intralinea.it

Zanettin, F. (forthcoming). "DIY corpora. The WWW and the translator", *Proceedings of the "Training the language services provider for the new millennium" International Conference, Porto, Portugal, 25-26 May 2001*.