Corpora for translator education and translation practice
Achievements and challenges

Silvia Bernardini

School for Translators and Intepreters
University of Bologna at Forlì, Italy
Corso della Repubblica 136
47100 Forlì, Italy
silvia@sslmit.unibo.it

Abstract

This paper aims to chart some of the ground we have covered in the last decade or so in the area at the interface between corpus linguistics and translator training/translation practice, and to point to some of the challenges (and prospects) lying ahead. Two related issues of central importance for the translation professionals of tomorrow will be focussed upon: the current impact (or lack thereof) of corpus-informed pedagogy on the training of translators, and the increasing availability of tools that facilitate the construction of corpora from the web. The further spread of corpus resources in the translation profession is suggested to crucially depend on two main developments taking place: a greater focus on awareness-raising uses of corpora in translator education, and a greater ease of access to and greater integration of corpus tools with CAT technology.

1. Corpora in the translation classroom

1.1. Achievements

Translation is in many senses an ideal field for corpus applications. The analysis of source texts against specialised and reference corpora can make the identification of stylistic traits, idiosyncrasies and *register*- and *genre*-specific conventions (Trosborg, 1997) easier. The browsing of target language corpora both prior to and during the production of a target text can reduce the amount of unwanted "shining through" (Teich, 2003) of the source language (SL) into the target text (TT), by providing the translator with an inventory of attested "units of meaning", i.e. conventional ways of expressing specific meanings and performing specific functions in the relevant text type/variety within the target language (TL) (Tognini-Bonelli, 2001, p. 131). Table 1 shows a concrete example of the kinds of insights one can gain in this way. Given a turn of phrase typical of the wine tasting domain in Italian (*il vanigliato del legno*), a translator with a specialised corpus for the target language at her disposal can extract and evaluate several likely translation candidates. In this case, the results of a simple search for *vanilla* are presented. These provide supporting evidence for the translation of *legno* (lit. *wood*) as *oak*; they also suggest that the term *vanigliato* can be rendered as, among others, *vanilla notes*, *nuances*, or *hints*.

Original Italian	avere il sopravvento sul *vanigliato del legno*
Original English	...*Vanilla and oak layers*... ...*vanilla and subtle oak undertones*... ...*vanilla characteristics* especially if oak-aged... ...*oak vanilla nuances* in dry wine... ...*subtle vanilla oak hints*... ...*a suggestion of toasty vanilla oak*... ...hint of vanilla oak... ...with *vanilla*, *oak* and apple *notes*... ...oak barrels, it may pick up *vanilla overtones*...

Table 1: Snippets from a search for *vanilla* in a web-derived bilingual comparable corpus on wine tasting

After all, and technological aids apart, these facts are not new to translators, for whom it is standard practice to rely on so-called "parallel texts", i.e., in corpus parlance, on the paper counter-part of *comparable corpora* of texts in the source and target language, matched by genre and subject matter to each other and to the text to be translated.

The last decade has seen a growing interest in the uses of corpora in translator education. Classroom experiences have shown that parallel corpora (of originals and their translations) can raise the students' awareness of professional translator strategies (Pearson, 2003), that comparable corpora can help them to produce more naturally-sounding translations (Zanettin, 2001), and that constructing corpora can itself be a learning activity, whereby students learn to reflect on texts while acquiring the practical skills they need to build their own documentation resources (Varantola, 2003). Several practical and accessible introductions to (aspects of) corpus use aimed at students and professionals have appeared. Bowker and Pearson (2002) is a book-length manual that walks the reader through the steps of building an LSP corpus, annotating it, consulting it, and applying it to different tasks (term and definition extraction, writing, translating).

If corpora are to play a role in the translation professions of tomorrow, it is important that they impact on the education of the students of today. The body of work just mentioned testifies that this is to some extent happening. However, there are also signs that substantial efforts still have to be put into place to convince the majority of translation students and teachers that corpus use can help to reflect on tasks and raise awareness of strategies, and that these are among the central goals translation courses should set themselves. Secondly, as we shall see, professionals still appear to be largely unaware of or unacquainted with corpora. Clearly, a second challenge for translator educators is to reach them as well. Section 1.2. discusses these issues.

1.2. Challenges

1.2.1. Educating educators

It is common practice to speak of the instruction of future translators as "translator training". The term "training" implies that the abilities and competences to be learned are expected to be acquirable through practice with the kinds of tools and tasks one will be faced with during one's future professional carreer, in an environment that reproduces as closely as possible the future work environment. Widdowson (1984) contrasts the training framework, in which learners are prepared to solve problems that can be identified in advance through the application of pre-set or "acquired" procedures, with the education framework, whose aim is to develop the ability to employ available knowledge to solve new problems, and to gain new knowledge as the need arises. According to Widdowson, LSP teaching would be an example of a training setting, while general language teaching would be an example of an educational setting.

We may wonder whether translator education is in fact closer to the training or to the education end of the cline. Gouadec (2002, pp. 32ff) explicitly champions the former position:

> [W]e are supposed to train people to perform clearly identified functions in clearly identified environments where they will be using clearly identified tools and "systems". [...] No serious translator training programme can be dreamt of unless the training environment emulates the work station of professional translators. [...] [T]he curriculum should [...] concentrate on emulating the actual work conditions of language services providers.

These views are certainly not unusual, and indeed are rather popular with students and prospective employers, who often lament a limited role of technology in translator education. While I am obviously sympathetic to the general issue of technology in the translation classroom, I think it would be dangerous to carry these views to their extreme consequences, for two main reasons.

First, if translation skills are best taught by simulating actual work conditions, we should abandon the idea of education for translators (something that even Gouadec would probably not want to suggest) and turn to apprenticeship instead: a professional environment should arguably provide a more appropriate setting for the simulation of actual work conditions than an academic one. Second, and more importantly, actual work conditions - and time pressure in particular - require that translator's strategies have become proceduralised, as is the case with mature professionals. Jääskeläinen (1997) finds that semi-professionals (translator trainees) show more extensive processing than both professionals and non-professionals. She suggests that this may be because they are aware of the problems involved but have not yet automatised the necessary problem-solving strategies. Automatic processes are typically very efficient but little flexible, such that there is the danger, pointed out e.g. by Wills (1994, p. 144), "of problems being forced into a certain structure, because it is believed to offer a solution". In an education setting, students are still to develop the strategies that will then become proceduralised. Forcing them to work under realistic time constraints as would happen in a simulation activity could therefore work *against* the development of professionality.

Translation instruction viewed as education, on the other hand, would make time for just the kind of activities and reflections that future professional translators will not have time for. A challenging aspect that is often neglected is how we can teach our students to identify problems in the first place. Going back to Gouadec (2002, p. 33), he claims that professional translators should possess, among others, the following skills:

1. Fully understand material to be translated
2. Detect, interpret and cope with cultural gaps [...]
3. Transfer information, facts, concepts [...]
4. Write and rewrite
5. Proofread
6. Control and assess quality

These skills translate into know-how; translators should know how to:

1. Get the information and knowledge required
2. Find the terminology
3. Find the phraseology
4. Translate
5. Proofread
6. Rewrite
7. Manage their task(s)
8. Manage a project (and other people)

Comparing the two lists, one notices that neither item 1 nor item 2 in the first (the "skills" list) translate into any of the know-hows in the second. In other words, there is a gap between "fully understand the material/detect any gaps etc." and "getting the information and knowledge required".

While illustrating this point with sufficient detail would take more space than is available here, a simple example can be provided. The phrases in the first column of Table 2 are taken from the *Time Out Barcelona Guide* (2002, Penguin). They are all titles of short sections devoted to different events or places, and they all involve wordplay. In these cases, to "fully understand the material to be translated" one needs to understand the relationship between the facts being recounted or places being described and the lexicalised expressions used. While the texts themselves no doubt provide hints for getting at the more "congruent" sense, the less congruent sense is normally not as easily inferrable from the texts, since it is assumed to be available to the reader (this is in fact a precondition for the success of the wordplay). A student who is not aware of these layers

Title	Topic	Senses
Get into the habit	Montserrat Monastery	*in the habit of doing* something: having a habit [...] of so doing. So *to [...] get into the habit* (OED) *the habit*, monastic order or profession (OED)
Getting high	*Castells* (human towers)	*high*: under the influence of drugs or alcohol (OED)
Death on the mountain	Montjuïc (site of executions)	James Still poem Japanese movie
On the tiles	The work of Architect J.M. Jujol	*on the tiles*: on a spree, on a debauch (OED) *Josep Maria Jujol*: Catalan architect, his activity ranged from furniture designs and painting, to architecture (wikipedia)

Table 2: Titles and senses: wordplay in the *Barcelona Time Out Guide*

of meaning may be misled into taking such expressions as *on the tiles* and *getting high* at face value only.

While it is easy to find out about these expressions, i.e. "get the information and knowledge required" with the resources currently available to any translator, I am arguing that the real and often underestimated challenge lies in teaching students to identify wordplay or other types of "layered" meaning in the first place. By drawing their attention to regularities in language performance as displayed in corpora, and making them reflect on the implications of (un)conventional usages, corpus-based activities such as those described in Sinclair (2003), Stubbs (2001) and Hoey (2005), especially if set within a translation-relevant framework, could help to fill this gap in translation pedagogy.

1.2.2. Informing professionals

While sensitising students and instructors is of great importance for reaching the professionals of tomorrow, we should not forget the professionals of today. Reading about translation aids, one seldom finds references to corpora and concordancing tools. This impression is confirmed by surveys attempting to find out whether professional translators are aware of the existence of corpora, and to what extent they use them in their work.

Surveying the Canadian market, Bowker (2004) finds that professional associations are aware of the existence of corpora, but are generally more interested in translation memory (TM) technology, and that job advertisements never mention corpora.

A more thorough investigation of the perception professional translators have of corpora is being conducted in the framework of the LEONARDO-funded MeLLANGE project, as part of an attempt to define user needs for learning materials on translation technology.[1] In the first round of submissions 623 questionnaires were returned, 90.8% of which completed by professional translators from the UK (the majority), France, Germany and Italy, and 9.2 by students of translation in the same countries. Out of the total respondents, 40.5% reported collecting reference materials, and more than half of them specified that they collect texts in eletronic format (69.4% of those who reported collecting materials). 46.9% read these collections of texts (rather than *searching through* them), and, of those who do search through them, a majority use search facilities in word processors (65.9%), with only a minority using a concordancer (19%, recall that data are for professionals *and* students).

While many translators are not acquainted with corpora, there seems to be widespread interest in learning more about them: 78.6% of respondents would be interested in a service which provides domain specific corpora, 77.9% in a tool for extracting terms from corpora, and 82.4% in learning more about their potential (MeLLANGE, 2005) (results are summarised in Table 3). Thus, there is clearly a need for tailor-made learning materials addressed to translation professionals, which highlight the value added of corpora with respect to other tools and resources, and which adopt a practical (but not uncritical) perspective.

2. Building corpora

2.1. Achievements

Bowker (2004) mentions different possible reasons why corpora and corpus analysis have not as yet received an enthusiastic welcome in the professional world. One of these is the fact that the design, compilation and exploitation of corpora can be very time-consuming while not providing a tangible immediate increase in productivity. The success of translation memories is instead partly explainable because both their creation and their consultation require minimal effort. Similarly, the fact that a large majority of the questionnaire respondents (above) reported consulting the Web through *Google* (93.4%), despite several drawbacks (that most of them are aware of), suggests that, for corpora to be successful with translation professionals, their construction and use has to be made substantially easier and faster.

One of the achievements of the past decade has certainly been the creation of tools that facilitate the extraction of textual information from the World Wide Web in ways that are more amenable to linguistic analysis. While search engines such as *Google* provide fast and effective retrieval of information from the Web, they are less than ideal when it gets to basic linguistic procedures such as highlighting patterns (i.e. sorting results) or selecting subsets of solutions, not to mention conducting searches for linguistically-annotated sequences (e.g. all verb lemmas preceding a certain noun lemma) (Thelwall, 2005).

A solution to some of these problems has been provided by tools like Fletcher's *KWiCFinder* (Fletcher, 2004), an online concordancer that supports regular expressions, implements concordance-like displays and functionalities (e.g. sorting), and allows off-line perusal of the retrieved texts. Along similar lines, another freely available tool, Matthias Hüning's *TextStat* concordancer[2], allows one to specify a URL and retrieve a file or set of files from a single website directly from within the concordancer, thus conflating and speeding up the processes of retrieving and searching texts.

While *KWiCFinder* is designed mainly with language learning applications in mind (searching for a given word

[1] `http://mellange.upf.es/`

[2] `http://www.niederlandistik.fu-berlin.de/textstat/software-en.html`

Do you collect domain specific texts?	59.5% No 40.5% Yes
How do you collect them?	69.4% In electronic form 30.6% On paper
How do you use them?	53.1% Search through with software 46.9% Read them
Do you use corpora in your translation practice?	60.2% No 39.8% Yes
If yes, do you use?	26.1% Corpora of the target language 23.1% Corpora of the source language 19.7% Parallel corpora 15.3% Domain specific corpora 13.6% Comparable corpora 2.3% General language corpora
What do you use to search them?	65.9% Search facility in word processor 19.0% Concordancer 14.4% Other search tools (specify: Trados, Concordance in translation memory) 0.7% UNIX utilities
If you do not use corpora, why?	41.9% Never heard about them 19.9% I don't have time to build them 17.8% I don't know how to use a concordancer 8.7% I can't see any advantage over *Google* 6.8% I can't see any advantage over translation memories 5.0% Other (1 specified - Not sure if it will work with Macintosh)
Would you be interested in a service which quickly provides domain- and language-specific corpora tailored to your needs?	78.6% Yes 21.4% No
Would you be interested in a tool for extracting terms from a domain-specific corpus?	77.9% Yes 22.1% No
Would you be interested in learning more about the potential that corpora offer?	82.4% Yes 17.6% No

Table 3: Corpus section of MeLLANGE questionnaire (first round, closed questions)

or expression as one would search the Internet), and *TextStat* only offers basic web-search facilities (i.e. it does not interact with a search engine, but simply spiders a specified URL), the *BootCaT* toolkit[3] (Baroni and Bernardini, 2004) was created specifically for translation students and professionals, i.e. for users who need relatively large and varied corpora (typically of about 1-2 million words), and who are likely to search the corpus repeatedly for both form- and content-oriented information within a single extended task. Starting from a series of "seeds" (search words), this set of Perl scripts provide facilities for combining the seeds into sequences, submitting queries to *Google*, retrieving URLs (for manual inspection if necessary) and eliminating duplicates. Then for each URL the text is retrieved, cleaned, and printed to a text file. This procedure can be iterated if larger corpora are required, e.g. selecting seeds for a second round of searches from the initial corpus and repeating the various steps. These tools have been used for several projects, including the construction of Internet corpora for several languages (see Sharoff's website[4] and Ueyama (forthcoming)).

The results in Table 1 were derived from a comparable corpus of English and Italian texts on wine tasting collected with *BootCaT* and used in an English to Italian translation course at the School for Translators and Interpreters of the University of Bologna, Italy. The conventions of this genre both in English and in Italian are unknown to virtually all the students in this course. A specialised comparable corpus is indespensable to (learn to) search for genre-restricted phraseology and terminology, two of the central know-hows identified by Gouadec (above). Given the time constraints under which translators normally operate, mastering techniques for the quick-and-dirty construction of corpus resources could be an additional asset.

[3] `http://sslmit.unibo.it/~baroni`

[4] `http://www.comp.leeds.ac.uk/ssharoff/`

2.2. Challenges

While the new tools at our disposal make the construction of corpora from the Web easier for translators, certain obstacles still have to be overcome. First, the *BootCaT* toolkit at the moment requires basic Unix skills and access to a Unix server. A Web interface is a crucial next step if these tools are to reach the average translator.

In the longer term, widespread use of corpora and corpus construction and search facilities is likely to depend on their integration with Computer-Aided Translation (CAT) technology. We could envisage a tool that interacted with a Web search engine to search, retrieve and POS annotate corpora based on user specifications. It would support regular expressions and handle subcorpora, and would provide facilities for monolingual and parallel concordancing (including alignment). Such a tool would extend the productivity of CAT systems by allowing a double search mode: fully automatic matching for golden-standard TMs, and manual concordancing of comparable and parallel texts for hypothesis development and testing where the TM has nothing to contribute:

> [...] translators working with texts that contain a large number of repeated segments, such as revisions, will be well served by the segment processing approach. On the other hand, translators who hope to leverage or recycle information from previous translations that are from the same subject field, but that are not revisions, may find that the bilingual concordancing approach is more productive. (Bowker, 2002, p. 124)

Such a system would also arguably limit some of the drawbacks associated with the use of TM. It has been observed (e.g. by Kenny (1999) and Bowker (2002)) that translators using CAT software may develop a tendency to make their texts more easily recyclable within a TM, regardless of the translation brief, and that they may be led to lose sight of the notion of "text" as a consequence of a rigid subdivision into units. The possibility to search whole texts (rather than translation units) using a concordancer could positively impact on these strategies and attitudes.

While no tool currently combines all these functionalities, some form of integration seems to be underway, thanks to tools such as *MultiTrans*,[5] a commercial CAT package which allows one to search for strings of any length (i.e. not limited to the size of a translation unit), and, if required, displays them in full-text context. Interestingly, while the company producing this software is called *Multicorpora*, no further mention of corpora can be found on the *Multitrans* page: yet another proof that corpora are currently not a buzzword in the translation market?

3. Summing up: prospects for the future

Despite achievements and enthusiasm within academic settings, corpora are still to make an impact on the translation profession. A number of reasons why this might be the case have been suggested, and several challenges have been identified.

[5] `http://www.multicorpora.ca/`

There seem to be three main areas where efforts should be concentrated. First, the role of corpus work for awareness-raising purposes should be emphasised over the more obvious documentation role, and the importance of basic "translation" skills be restored to its central place in translator education:

> [...] the general abilities to be taught at school [...] are the abilities which take a long time to learn: text interpretation, composition of a coherent, readable and audience-tailored draft translation, research and checking, correcting. [...] If you cannot translate with pencil and paper, then you can't translate with the latest information technology. (Mossop, 1999)

Second, translator-oriented (e-)learning materials have to be provided, so as to reach those professionals who are eager to learn about corpora. These materials should ideally be contrastive in focus (i.e., why/when use corpora instead of the Web/TMs/Dictionaries?). They should also include substantial practice primarily with those tools and facilities that translators (rather than linguists or language learners) are likely to find of immediate relevance (e.g., concordancing should arguably be given priority over frequency wordlisting). Finally, corpus construction and corpus searching tools should be made more user-friendly, and ideally integrated with CAT tools.

4. References

M. Baroni and S. Bernardini 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. *Proceedings of LREC 2004*, pp. 1313-1316.

L. Bowker. 2002. Corpus Resources for Translators. In S. Tagnin, editor *Corpora and Translation, TradTerm 10 Special Issue*.

L. Bowker. 2004. *Computer-aided Translation Technology.* Ottawa: University of Ottawa Press.

L. Bowker and J. Pearson. 2002. *Working with Specialized Language. A Guide to Using Corpora.* London: Routledge.

W. Fletcher. 2004. Facilitating the Compilation and Dissemination of Ad-hoc Web Corpora. In G. Aston, S. Bernardini and D. Stewart, editors, *Corpora and language learners*, Amsterdam: Benjamins.

D. Gouadec. 2002. Training Translators: Certainties, Uncertainties, Dilemmas. In B. Maia, J. Haller and M. Ulrych, editors, *Training the Language Services Provider for the New Millennium*, Oporto: Universidade do Porto, pp. 31-41.

M. Hoey. 2005. *Lexical priming.* London: Routledge.

R. Jääskeläinen. 1997. *Tapping the Process: An Explorative Study of the Cognitive and Affective Factors Involved in Translating.* Doctoral dissertation. Joensuu: University of Joensuu.

D. Kenny. 1999. CAT Tools in an Academic Environment. *Target*, 11(1), pp. 65-82.

MeLLANGE. 2005. *Corpora and E-learning Questionnaire: Results Summary.* Internal document, 20.06.05.

B. Mossop. 1999. What Should be Taught at Translation School? In A. Pym, editor, *Innovation in Translator and Interpreter Training - An Online Symposium*. online: http://www.fut.es/ apym/symp/mossop.html [visited: 23.02.06]

K. Pearson. 2003. Using Parallel Texts in the Translator Training Environment. In F. Zanettin, S. Bernardini and D. Stewart, editors, *Corpora in translator education*, Manchester: StJerome, pp. 15-24.

J. McH. Sinclair. 2003. *Reading Concordances.* London: Longman.

M. Stubbs. 2001. *Words and Phrases.* London: Blackwell.

E. Teich. 2003. *Cross-linguistic Variation in System and Text.* Berlin: Mouton.

M. Thelwall. 2005. Creating and Using Web Corpora. *International Journal of Corpus Linguistics*, 10(4), pp. 517-541.

E. Tognini-Bonelli. 2001. *Corpus Linguistics at Work.* Amsterdam: Benjamins.

A. Trosborg. 1997. Text Typology: Register, Genre and Text Type. In A. Trosborg, editor, *Text Typology and Translation*, Amsterdam: Benjamins, pp. 3-23.

M. Ueyama. forthcoming. Evaluation of Japanese Web-based Reference Corpora. In M. Baroni and S. Bernardini, editors, *Wacky! Working Papers on the Web as Corpus*, Bologna: Gedit.

K. Varantola. 2003. Translators and Disposable Corpora. in F. Zanettin, S. Bernardini and D. Stewart, editors, *Corpora in Translator Education*, Manchester: StJerome, pp. 55-70.

H.G. Widdowson. 1984. English in Training and Education. *Explorations in Applied Linguistics II*, Oxford: Oxford University Press, pp. 201-212.

W. Wills. 1994. A Framework for Decision-making in Translation. *Target*, 6(2), pp. 131-150.

F. Zanettin. 2001. Swimming in Words. In G. Aston, editor, *Learning with Corpora*, Houston, TX: Athelstan, pp. 177-197.