“Professional corpora”: Teaching strategies for work with online documentation, translation memories, and content management

Anthony Pym
Intercultural Studies Group
Rovira i Virgili University
Tarragona, Spain

Paper delivered to the Forum on Translator Training held in Beijing, China, and organized by China Foreign Language Press and the Monterey Institute of International Studies, August 16-17, 2007.

Abstract. The expansion of electronic memory capacity is having fundamental long-term effects on the way texts are produced and used, and thus on the way they are translated. Translators are increasingly working on data bases in non-linear ways, separated from awareness of any active communicative context. This enhances productivity and consistency but challenges more humanistic values like understanding, cooperation, and job satisfaction. It order to address these changes, teaching practices should 1) make students aware of the communicative functions of texts, particularly the ways in which particular parts of texts become high-risk in particular situations, 2) teach students how to use electronic technologies within such a frame, and how to teach themselves about the technologies, and 3) train students for a range of professional communication jobs, incorporating both the technical sides and the various revision and editing techniques now required by the technologies.

Do not worry: there is no universal revolution this week. Translation is still what it has always been, more or less. In some particular fields, however, and indeed in a widening circle of fields, a series of memory-based technologies are fundamentally altering the way translations are produced, and thus the way translators need to be trained. Here we shall focus on those particular changes, in those particular sectors, with the rider that some of the novelties may be coming your way soon.

Where are the texts?

The most basic of these changes concerns the kind of linguistic material that translators work on when dealing with websites, software programs, and product documentation. There are still no doubt texts, with a beginning, a middle and an end, of the kind Aristotle approved of. But much translational work is now carried out on sets of

linguistic data, and is done with the help of sets of linguistic data, and with no visible beginning, middle or end. For example, a translator may be required to locate and translate a series of updates to a website. They will render just those updates, without necessarily seeing the entire website as any kind of text, and quite commonly not as any kind of image either. Or again, translators may work in a team using an online translation-memory system to render hundreds of pages of technical documentation for an impossible deadline: only the project manager, at best, will see the text as a whole (or indeed, as a project); the translators themselves will only see series of small unconnected parts, like foot-soldiers in a battle.

As these examples suggest, the conceptual and cognitive changes are due to technology on two levels. On the one hand, the source material is being generated by piecing together fragments, often for users who are only going to use fragments (think of all the user manuals and software Help files, all produced and read through indexes). On the other, translation memories and content-management systems divide linguistic material into phrases and chunks, at cohesive levels much lower than anything traditionally called a text (usually at the paragraph or section levels). This fundamental change, the absence of initial textuality, underlies all the rest. People no longer use such contents in a linear way, starting at the beginning and reading through to the end. So the documents are not written in a linear way, and they are certainly not translated in a linear way. We are all working on memorized chunks, and updates to memorized chunks. In semiotic terms, the paradigmatic has drawn-and-quartered the syntagmatic.

It took us a few decades to realize that translators work not on sentences but on texts, much to the chagrin of phrase-level linguists in search of equivalence. It then took us the 1990s to see that documents usually come with useful information on deadlines, quality required, readerships and rates of pay, so that what we actually work on are not just texts but all those things bundled into “projects”. And now another decade or so has been necessary for us to see that technology has changed translation activities into something like maintenance operations, of the kind you use when you have your car serviced. When we only translate the updates, when we do not see the beginning or end of the communication act, our work has moved from the “project” to the “program” (yes, like a car-service program). Linearity is no longer in texts, but in the possibility of keeping maintenance contracts over time.

The segments, at whatever level, are universally memorized (memory is what our technologies work on) and paired across languages. Does that mean some kind of return

to phrase-level equivalence? In many respects, yes, except that equivalence is now fixed by company convention, with little disturbance from natural usage. What we work from are increasingly not contextualized fragments of language, with social connotations and the like. What we access and apply are sets of data, lists of linguistic material, or what are elsewhere known as corpora.

Where is the information?

Translators still need to know languages and cultures. In highly technical fields, however, good documentation skills, good organization skills, and some basic common sense can often replace developed language competence. For example, I occasionally have to render legalistic documents into Catalan, a language that I do not know well. This mostly concerns university regulations, suggested modifications to university regulations, and explanations of why I do not apply university regulations correctly – usually in my devious attempts to get foreign students enrolled. How do I locate the correct Catalan terms and phrases? Obviously I look at selected parallel texts on websites (texts in Catalan with the same or similar discursive function); I also look closely at the texts that are sent to me (mainly the university regulations themselves, and the official complaints about my non-applications); I might very occasionally check an online bilingual legal glossary; I learn a lot from my Catalan spell-checker; and finally, depending on the quality required, I learn from revision by a native speaker. On a bad day I might write a text in Spanish, a language I do know fairly well, then get a web-based machine translation into Catalan, which gives remarkably good results, and then run all or some of the above checks. Of course, when I do the translation with a translation memory, all future efforts draw on the matching phrases I produce. With all those modes of assistance, I still may not know much about the language or the culture, but I can certainly write and rewrite some effective official letters in Catalan.

My main point here is that all those processes draw on corpora of one kind or another. The parallel texts form a corpus, as do the glossaries and dictionaries and spellcheckers, and the translation memories, and perhaps even the knowledge stored in my reviser's brain. These are certainly not corpora in the sense used by mainstream Corpus Linguistics: the lists of data do not seek to represent a whole national language or generalist patterns of usage; the use of those lists does not require any systematic mining for terminology or phraseology; there is no sophisticated knowledge

management at stake, and I am not about to suggest there should be (professional usage tends toward efficient knowledge management anyway). When we use those corpora, we are simply writing (or rewriting, or translating) in a way that works from lists rather than from texts alone. We are using what I want to term "professional corpora". In order to use those materials, we need skills in addition to those associated with the use of languages and the weaving of texts in cultures. Those skills are what we now have to identify and somehow teach.

What is translation competence?

My second point is no less important, but it is harder to explain. Some years ago I proposed that translation competence properly involves solving problems to which there is more than one correct solution (Pym 1992). When there is just one solution (French "faire un discours" is English "make a speech"), then we are applying terminology, or phraseology, or whatever cultural authority legitimates the equivalence. However, there are many situations in which more than one solution is viable (French "élaborer un discours" might be "develop a speech", "develop a discourse", "elaborate a discourse", and much more, depending on a hundred subtle semantic and rhetorical factors). In those cases, I proposed, properly translational skills were required. The questions of right vs. wrong ("binary" problems) were for language-learning and electronic memories: the problems with more than one solution ("non-binary" problems) were for translation competence. Ours were the skills that paid close attention to textuality, that adapted messages to new purposes, that enacted the interplay of cultures. That is how I sought to define the line between the training of translators and the learning of languages.

That theory enjoyed a certain success in its day. Yet I am now compelled to reconsider it. If we look at the many professional corpora now instantaneously available, can we really pretend that their use is not part of what translators do? Do we really want to say that translators somehow become more "translatory", as it were, the more they work on the more-than-one, in effect the more they use and re-use solution-proposing technologies? That is not a very happy theory. It could even condemn translators to produce multiple alternative renditions, wasting time and in many cases reducing the qualities of outputs: Lorenzo (2002) finds that the longer student translators go over and revise their work, the lower the quality becomes. On that view,

the more translatory the translator, the less efficient the translation process. And that is certainly no longer where the solutions lie.

In the pre-Internet era, which was when I started talking about non-binary problem-solving, the hardest part of translating was the location of data. I remember rendering a Spanish report on the "irregular", "informal" or "grey" economy (the kind that the tax office is not supposed to know about). But which of those terms was correct? Where could I find authoritative English for this particular corner of economics? After consulting a series of Spanish professors (I was in Spain), I finally reached the Information Office of what was then the European Community. Into the archives I dived, and in half an hour or so, with some expert help, I had just what I needed—a Commission report, in English, on the "informal economy of Spain". Perfect! This was my parallel text, the authority I sought. Until, of course, I read the first few pages of the report and began to recognize the style: this was a translation I had produced about one year previously. I had become my own apparent authority. When we work on the front lines, which is most of the time, our decisions are constantly under-determined: we must decide in places where no one has yet decided. Or so it seemed a few decades ago.

Now, of course, the problem is quite the opposite. There is such an abundance of information that the key is to know when and how to discard that which is least authoritative. The processes that, for me, were once so uninteresting as to be non-translatory are now, thanks to technology, essential to efficient and successful translation practice. All the binary problem-solving must now be reconsidered as a substantial part of what translators do.

In this context, a better way of modeling the translator's problem-solving processes is to assess the distribution of effort in terms of communicative risks.

Technology and risk-management

When a problem has more than one possibly right answer, the translator must decide. In making that decision, the translator takes a risk—if I render "élaborer un discours" as "develop a speech", I am betting that this option will not hinder the communication flow concerned, but there is no absolute guarantee that this is not the case. Translators take risks.

Now, what happens when we are working with professional corpora rather than texts? First, when the technology and the teamwork do not allow us to see the whole text (the whole website for example), we have little way of assessing the purpose of the communication. This means there is not much chance of us evaluating any communicative risk—all problems potentially involve the same risk. That is indeed what a lot of the technology is forcing us into: we have to take the same risks as ever, but more blindly than before. We cannot see which parts of the text are high-risk (and worthy of our best non-binary efforts), and which are low-risk (to be rendered as quickly as possible, probably in a binary way).

At the same time, however, when a corpus presents the translator with a possible rendition, it always does so with the backing of a certain authority. The very existence of the corpus means that people beyond the translator have invested effort into that solution. If the translator accepts the solution, their risk is thus partly transferred to the authority behind the corpus, especially when the corpus comes from the client or language-service provider: the solution may be dead wrong, but the translator will not be entirely to blame! This is one of the main ways in which translators can reduce the enormous risk burden imposed by "blinding" technologies (the main other risk-reduction strategies are generalization and omission). The corpora thus at once enhance the initial risk and offer the possibility of risk transfer. Hence the extreme importance of translators being able to assess the relative authorities behind different corpora (in which case we simply generalize the basic lessons of website usage: do not believe everything you are see).

The key to risk analysis lies in the distribution of the translator's effort. As we have said, not everything is equally important in a text; translators should work hard only on those parts that are important, the parts that impinge on the communicative success conditions, the parts that are in themselves high-risk, and merit high effort.

Unfortunately, to see where those parts are, the translator really needs to see the entire communication act. Where the technologies and organizations do not allow this, all the translator can do is distribute effort in accordance with the authorities of the corpora. Translators are thus virtually obliged to accept the renditions that come from the client (the glossaries and translation memories); they will not challenge deceptive "exact matches"; they are not motivated to improve the quality of the memories. The result is that database technologies, particularly web-based translation memories, lead to

a progressive degradation of translation quality. The more the technologies are used, the worse the translations become.

Happily, that is not the end of the story.

The return of textuality?

If various corpora technologies would seem to spell doom for any kind of high-quality translating, let alone any semblance of humanized communication, those who work on large-scale translation projects are aware of this. They do not all accept a passive ride on a down-hill slide. So how are the risks reduced, in spite of the technologies?

A first step is in the use of controlled writing to produce source texts. When textuality is restricted from the outset, there are likely to be fewer problems with it further down the line. If the initial documents are kept as simple and as controlled as possible, machine translation works reasonably well, and revisers can take it from there. In other words, translation quality is maintained by reducing the role of the human translator (and indeed in reducing the role of textuality).

A second step is the accumulation of organizational power in the figure of the project manager, who is ideally able to perceive and control the entire translation process. In best-case scenarios, project managers do indeed see beyond the technologies, and they can intervene accordingly. (In current reality, it seems they are mostly concerned with time management rather than communication, but we live in hope.)

A third step is the investment of serious time and effort in product-testing and text-revision processes. The translator may not be able to see where the text goes and what it has do to, but an expert reviser will have some access to this wider picture, and the project manager will ideally receive feedback on actual text usage.

Those three measures are obviously not equal in nature or extent of actual application. Yet together they potentially allow our technical translation processes to work both with and against corpus technology. As such, they represent a certain return to the textuality to which the translator is granted only restricted access.

All is not lost.

Lessons for training institutions

How should our training institutions respond to these challenges? As we have said, there is no complete revolution, so there is no need to throw out our traditional training methods. At the same time, however, some parts of our institutions will want to address the new demands. Here we attempt to summarize the main tasks to be faced, offering just a few hints as to what can be done under each rubric. As will be seen, this involves working both with and against the new technologies:

1. Cultivate awareness of risk distribution.

Students must be aware that not everything in a document is of the same importance, and that in many places it is both rational and efficient to accept corpus-produced solutions without further ado. That awareness, however, requires quite advanced skills in precisely the aspects that the current technologies hide: purpose-oriented communication and text organization. Here are some ideas on how to cultivate those skills:

1a. Get students to summarize texts, and have them produce only summaries for three months or so, until they seriously appreciate that not all parts of a text are of equal importance. (This is a point repeatedly stressed by the French translator trainer Daniel Gouadec.)

1b. Get students to do oral translations, in face-to-face situations with human users, before they start to write down their translations. The oral should precede the written; the spontaneous explanation, with immediate feedback, should precede the use of fixed databases. (This point has been picked up by the German theorist Hans Vermeer, and has become common enough in theories, although rare in practice.)

1c. Get students to use each others' outputs, so that they develop supplementary awareness of communicative risks by effectively becoming product testers. In bilingual classes, you can get students to translate each other's most intimate texts (their first love, for example), so that the authors have the uneasy experience of being translated. (I have stolen this idea from Andrew Chesterman.)

2. Teach the effective use of professional technologies.

This is the most obvious and perhaps the most deceptive of responses. Hours of class-time need not be set aside for teaching every new piece of software on the market. There are too many programs out there; they virtually all have web-based demos and tutorials available for free; a lot of on-site training happens when students do work

placements. The role of training institutions should probably be limited to saying what is out there (particularly the market leaders), conveying the more general coping strategies, and letting the students discover technology for themselves.

2a. Train students how to teach themselves technology.

Get them to learn different programs by themselves or in pairs, quickly, looking for advantages and disadvantages, and then reporting back to the group. Whatever software we train them in now will be obsolete in five years' time; students must learn how to learn (and the younger they are, the more competent they are in this particular aspect).

2b. Have students work together with web-based materials.

Even in class, the commercial demos and tutorials can be used, along with the web-based learning materials that each course should develop. The role of the teacher should be to provide over-the-shoulder support when really needed.

2c. Use work placements.

The advantages and disadvantages of technology can only really be seen in the context of professional practice. On-site work placements can provide this, just as they must provide much of the real-world learning experience.

3. Train for diversified professional roles.

Given the changes in translation competence, pure translating is perhaps not the most exciting or lucrative of the many task-sets available. All students should be trained in neighboring skills, not just to enhance employability but also so they can understand their place within large translation activities.

3a. Teach controlled writing.

Controlled input is also referred to as "writing for translation". It can work as a reverse pedagogy, moving from error analysis to the removal of the structures that lead to such errors. Or it can be carried out as an experiment: have the controlled and uncontrolled texts translated by the group, and see whether the problems really are removed.

3b. Teach professional revision strategies.

Revision strategies should really be part of technical writing courses, along with writing for translation. In the context of translation projects, much has to be done on how to correct the kinds of errors associated with non-linear text production.

3c. Teach corpus management.

Since professional corpora are ubiquitous, students should know what they are, how they are constituted, and how they can be managed. Terminology and phraseology are

by no means to be equated with translation, but their basics are increasingly part of our professional context.

3d. Teach project management.

Project management requires business and organization skills that translators often do not have. Those skills can be learned as theoretical precepts, just as project-management software is fairly easily explained. However, project management is only developed in the context of actual projects, which is precisely what class groups should be involved in.

3d. Have students engage in teamwork projects.

Once you have a group with functional skills in controlled writing, terminology, project management, revision, and of course translation, they should be made to work together on sizeable projects, with clear and rotating divisions of labor, both within the class and extending projects over weeks without fixed class hours. Teamwork is one of those things that can be learned in practice but rarely in theory. The more use we make of it, the better prepared our students will be to work on large-scale translation projects.

3e. Have students use and re-use memories.

At advanced levels, a series of projects in the same field can bring in all the problems of exchanging and re-using translation memories. Here, especially, only actual practice will convince students that they need to update memories and keep them clean.

Our three rubrics are thus risk distribution, use of technologies, and diversification of roles. Work on those three fronts can hopefully prepare our students for efficient—and often quite lucrative—interaction with professional corpora.

Some will object, of course, that we lack the teaching staff able to do all of this. On that score, I believe the answer is simple. Just as we must allow the professional corpora to enter our classes, so we should be inviting translation professionals to come into the classroom to explain what they do and to convey their skills.

Gone are the days when the linguists and their particular corpora could hope to explain the world of professional translation.

References

Lorenzo, María Pilar. 2002. “Competencia revisora y traducción inversa.” *Cadernos de Tradução* 10: 133-166.

Pym, Anthony. 1992. “Translation Error Analysis and the Interface with Language Teaching”, *The Teaching of Translation,* ed. Cay Dollerup & Anne Loddegaard, Amsterdam: John Benjamins. 279-288.