Corpus-based translation studies: The challenges that lie ahead

Mona Baker
UMIST & Middlesex University

Introduction

Translated text has always had a very raw deal in corpus linguistics. It has been specifically excluded from monolingual corpora, where it is generally treated as unrepresentative of the language being studied, irrespective of the direction of translation: even text translated into one's own native language does not normally qualify for inclusion in a monolingual corpus. Where translated text has been studied at all, the idea has been to show that "translationese" is common, or that some of the language that the corpus linguist is interested in studying is influenced by another language. In fact, the very notion of using corpora to study translation as such – simply to understand it as a phenomenon – does not seem to have occurred to corpus linguists. In a recent collection on corpus-based studies, Lauridsen expresses the overall position of corpus linguists when she asserts that

> . . . one should refrain from using translation corpora unless the purpose of the linguistic analysis is either to evaluate the translation process or to criticise the translation product on the basis of a given translation theory. (Lauridsen 1996:67)

It is only very recently, in the past twelve months or so, that we have started to consider using the techniques and tools of corpus linguistics to study translation as a variety of language behaviour that merits attention in its own right: not in order to criticise or evaluate individual translations but in order to understand what actually happens in the process of translation. This development reflects an increased awareness within translation studies of the distinctive nature of translation as a communicative event which is shaped by its own goals, pressures and context of production.

Emphasis has shifted very gradually in recent years from the source to the target text. This did not happen overnight: for a long time, target-oriented translation studies was only target-oriented in the sense of not imposing the standards of the source text onto the target text, of allowing for the fact that different contexts and communicative goals may require different translation methods. We started with typologies of source texts and assumed that translators have to use different strategies for different source text types. Then we became a little bit more sophisticated and suggested that the function of the translation may sometimes be different from that of the source text, as when a novel is adapted for the stage or for the screen. But pedagogical goals blinded us to the theoretical implications of this position, and instead of going on to develop a typology of translated text – because the implication here is that a translation, like any kind of text production, develops in response to the pressures of its own immediate context and draws on a distinct repertoire of textual patterns – we settled for suggesting that equivalence is a function of the assignment or commission that accompanies a request for translation. Which means that we were still trying to justify translation methods rather than explain translation itself as a textual phenomenon.

With Sager (1984) we see – for the first time, to my knowledge – an interest in establishing "the relation of translation to other forms of text transformation" (p. 333) and, consequently, an attempt to develop not a typology of methods of translation but, uniquely, a typology of translated text *per se*. This typology was developed in a much fuller form in Sager (1994) and is based on the "status and message type of [the] target language document" (1994:178). Sager's "functional types of translation" are an attempt to identify "the distinct functional characteristics of the translated documents" (*ibid.*:179), independently of any source language and, to some extent, independently of the source text, though Sager does state that "the type of translation that can be produced depends first of all on the status attributed to the original" (*ibid.*:178) and that types of translation are only "secondly . . . defined with respect to their functional role as new products" (*ibid.*:179).

With the availability of corpus techniques, we can now go a step further and look not just at the functional types of translation but at the distinctive features of translated text *per se*. The kind of distinctive, universal features that have been proposed in the literature, but never tested on a large scale, include **simplification** (the idea that translators subconsciously simplify the language or message or both), **explicitation** (the tendency to spell things out in translation, including, in its simplest form, the practice of adding background information) and **normalisation** or **conservatism** (the tendency to conform to patterns and practices which are typical of the target language, even to the

point of exaggerating them). A fourth feature which does not seem to have been considered in the literature is what I will call, for lack of a better term, **levelling out**.[1] This concerns the tendency of translated text to gravitate around the centre of any continuum rather than move towards the fringes. I will give examples of all four features shortly, but the idea of "levelling out" simply means that we can expect to find less variation among individual texts in a translation corpus than among those in a corpus of original texts. In other words, translated texts seem to be less idiosyncratic, or more similar to each other, than original texts.

A computerised corpus of translated text, one hopes, will reveal regularities of this type. Logically, *it has to*. We may be wrong about the specific regularities we are expecting to find, and we may well be wrong about the ways in which we currently propose to test the relevant hypotheses. But given that all language is patterned, and that this patterning is influenced by the purpose for which language is used and the context in which it is used, the patterning of translated text must be different from that of original text production; the nature and pressures of the translation process must leave traces in the language that translators produce. Translation is a language activity which is performed in a unique context, quite distinct from normal text production, including text production by learners of a foreign language. Translated text

(a) is normally constrained by a fully developed and articulated text in another language,

(b) looks both ways: it has to respond to the needs of its prospective readers and the context in which it will ultimately function while, at the same time, taking into account the original readership and context of production, and

(c) has to cope with – and linguistically respond to – its own social and textual status, which changes across time and in different cultural communities. This may explain, for instance, why translators tend to stay close to the typical patterns of the target language and why translations seem to gravitate towards the centre of any continuum: creativity on the part of the translator can easily be confused with literalness and assumed to be a sign of interference from the source language.

1. I borrow the term from Shlesinger (1989), who uses it to refer to the shifts that occur in simultaneous interpreting along the oral–literate continuum.

2. Investigating translation as a phenomenon: Some initial challenges

In early 1995, work began on designing and physically putting together a translation corpus which consists of translated English text in a number of domains and from a variety of source languages.[2] This corpus is currently housed at UMIST and, to my knowledge, is the first of its kind, in the sense that it is specifically designed for the study of translation: it is not designed to provide a resource for bilingual lexicography nor to isolate instances of "translationese", but to identify types of linguistic behaviour which are specific to translated text, patterns of linguistic behaviour, in other words, which are generated by the process of mediation during translation.

This type of venture faces many challenges. First, there is the question of setting up the corpus in the first place; here there are thorny theoretical issues associated with corpus design, including the elaboration of criteria for including or excluding individual texts. There are even thornier practical issues to do with the physical creation of the corpus and with acquiring copyright permission for each text. Challenges of this type are not unique to a translation corpus of course; monolingual corpus creation has had to cope with such issues for many years, with the difference that it is harder and less straightforward to obtain permissions for translated text.

Challenges which are specific to a translation corpus designed to research universal features of translation include the difficulty of isolating the influence of particular source languages. Researchers have to develop techniques for assessing the significance of a pattern in terms of whether it only emerges in translations from a particular source language, or a particular family of languages (such as Romance languages), or whether it is typical of translation *per se*. This means that the researcher has to think carefully about the parameters that have to be recorded and how they may be recorded and displayed in such a way as to make the process of assessment reliable and convenient. At UMIST, we record a variety of parameters in relation to each text in the corpus, including source language, sex (and, where possible, sexual orientation) of the translator and of the author, direction of translation, place of publication, and various other things which we feel may have a bearing on the features identified.

Another theoretical issue which poses a genuine challenge and which is well worth noting early on is the need to maintain a reasonable balance between

2. The basic structure of the corpus was first conceived in Baker (1995), where examples of the kind of research that can be carried out on such a corpus were also discussed. The elaboration of design principles and the physical creation of the corpus, however, were undertaken by Sara Laviosa-Braithwaite as part of her PhD research at UMIST (see Laviosa-Braithwaite, in prep.).

the general and the specific, between the norm and the exception, in corpus-based analyses of translation. This point has already been made by Malmkjær (forthcoming, a) who warns against putting too much emphasis on norms and typical patterns at the expense of the one-off or atypical example.

Researchers who set out to study the patterns of translated text may be tempted to treat these patterns. The problem of course is that once we have very large amounts of text on the computer, and the ability to generate all kinds of statistics and frequencies at the press of a button, there is a strong temptation to emphasise the norm, what is typical, at the expense of the one-off, the more creative use of language, even to the point of labelling it as "wrong" because it does not fit in with the norm. But we must not lose sight of the fact that one of the main reasons we want to study the patterning of any kind of language or text production, including translation, is that patterns are the backdrop against which creativity can take shape: norms enable the creative use of language and identifying them allows us not only to capture universal features of translation, and hence understand translation as a phenomenon in its own right, but it also allows us to make sense of the individual example which, like Malmkjær, I think is very important indeed. So, as well as devising techniques for capturing universal features of translation, we also have to be prepared to deal with and explain less typical features, even the one-off, the atypical example.

These, then, are some of the challenges that await us and that we need to address systematically. The real challenge, however, and the one I wish to address in more detail, concerns the way one proceeds to investigate a translation corpus in the context of trying to identify universal features of translation. What kinds of things do we look for and what sort of search techniques do we use to locate them? In other words, how should we go about capturing our evidence? We need to develop a sense of what we should be looking for and the skill and relevant procedures for locating it. And we need to think about what may or may not be feasible to investigate using corpus techniques. Corpora do have their limitations, and some questions are better addressed using other research methodologies.

3. Locating universal features in a translation corpus

We know by now that we are interested in features such as simplification, explicitation, conservatism and so on. But these are very vague and abstract notions, and corpus linguistics does not lend itself to abstraction: we cannot search for something like simplification in a corpus. We have to ask ourselves first what kind of concrete manifestations each feature might have in text, and

then we have to consider how we might locate these concrete manifestations, by no means an easy matter as I hope to show shortly.

There are two main problems here: one is that the same feature may be expressed in different ways on the surface, and the same surface expression may point to different features or tendencies. The second problem is that we do not have any clear definitions of the features in question. What do we mean by simplification? Or explicitation? Is there a difference between the two, or is it the same feature under different names? Take something like finite as opposed to non-finite structures for instance (cf. (1) and (2)).

(1) I do not feel shocked, having expected him to take the easiest way out.

(2) I do not feel shocked, because I expect/expected him to take the easiest way out.

If we find that finite structures, as has been suggested in some studies, are more common in translation than non-finite structures, or than finite structures in original text, do we take that as proof of simplification or explicitation or both? Finite structures may be easier and simpler to process, but they also allow the speaker to make things like tense and causal relations more explicit. It is clear, then, that we need to refine our definitions of the individual features we propose to investigate. My own feeling is that the process of refining the definition will go hand in hand with that of verifying the feature: definition and verification are interdependent in the sense that it is only by investigating the various concrete manifestations of these abstract notions that we will be able to refine the concepts themselves.

With this proviso in mind, I would now like to look in some detail at the individual features and see how we might investigate them, how they might be expressed on the surface.

3.1 Explicitation

I take "explication" to mean that there is an overall tendency to spell things out rather than leave them implicit in translation. The evidence for this tendency may be found in a range of textual phenomena. First, we might look at text length. For a long time, many people have suggested, without carrying out any empirical research, that translations are usually longer than their originals, irrespective of the languages concerned. More recently, Stig Johansson, who works with Norwegian and English corpora at the University of Oslo, has reported (1995:23) that he found an average increase of about 10% in the number of words in English translations vs. Norwegian originals, as well as a slight average increase in the other direction (Norwegian translations compared to English originals).[3] This type of search strategy obviously involves comparing a corpus of source texts with a corpus of target texts, on a text-by-text basis.

But explicitation may also be expressed syntactically and lexically, and this we might investigate using, in addition to corpora of source and target texts, corpora of original and translated texts in the same language (and the same domain of course); this is the kind of corpus that we are working with at UMIST. If we compare original and translated English text, for example, we might find that the optional *that* in reported speech is spelled out more often in translation than in original English text. Lexically, the tendency to make things explicit in translation may be expressed through the use or overuse of explanatory vocabulary and conjunctions, and can be seen in examples such as the following:

(3) a. English original (William Faulkner's *Sanctuary*):
Nobody wanted her out there. Lee has told them and told them they must not bring women out there, and I told her before it got dark they were not her kind of people and to get away from there. It was that fellow that brought her.

b. Hebrew translation:
But it was that fellow that brought her *who caused* [the incident].
(Ben-Shahar 1994:211)

So one thing we could do is to compare the frequencies of words such as *cause*, *reason*, *due to*, *lead to* and conjunctions or adverbs such as *because*, *therefore*, *consequently* in a corpus of original and translated text in the same language. This would enable us to confirm whether translations tend to draw more heavily on the explanatory vocabulary of the language and hence make more explicit the relations between propositions in a text.

3.2 Simplification

We can tentatively define "simplification" as the tendency to simplify the language used in translation. Translators, for instance, may be inclined to break up long sentences in translation, so we might look at average sentence length in both source vs. target texts, and in original vs. translated language. Based on an investigation of a subsection of the corpus held at UMIST, Laviosa-Braithwaite (1996) provides evidence that average sentence length in the translated section of *The Guardian* is significantly lower than the average sentence length in a comparable non-translated section of the same newspaper.

3. Text length, in terms of number of words, will inevitably be influenced in part by the morphological structure of each language. What is interesting about this finding is that there is an increase in the direction of Norwegian as well.

Simplification involves making things easier for the reader (but not necessarily more explicit), but it does tend to involve also selecting an interpretation and blocking other interpretations, and in this sense it raises the level of explicitness by resolving ambiguity. I have already mentioned the use of finite as opposed to non-finite structures, where there is a clear overlap between simplification and explicitation. Another similarly problematic but, it would seem, very rich area of investigation, is what happens to *punctuation* in translated text.

Several studies have appeared recently which suggest that punctuation tends to be changed in translation in order to simplify and clarify. For example, in a study of English translations of Hans Christian Andersen's stories, Malmkjær (forthcoming, b) explains that Andersen uses unusual punctuation by Danish standards and suggests that his translators consistently simplify this use in translation. Assuming that we can rank punctuation in terms of strength (comma – semicolon – period), she finds that "translators, whenever they alter ST's [the source text's] punctuation, alter it from a weaker to a stronger mark": commas become semicolons or periods, semicolons become periods. Strengthening punctuation, I imagine, is part of a subconscious strategy to make things easier, simpler, by making them more clear-cut.

Similarly, the Russian and French translators of Virginia Woolf apparently adjust her unusual punctuation in order to make the texts easier to read. This example, from Virginia Woolf's *To the Lighthouse*, is discussed in May (forthcoming):

(4) a. English original:

> But no, he wanted nothing. His hands clasped themselves over his capacious paunch, his eyes blinked, as if he would have liked to reply kindly to these blandishments *(she was seductive but a little nervous) but could not,* sunk as he was in a gray-green somnolence which embraced them all, without need of words, in a vast lethargy of well-wishing; all the house; all the world; all the people in it, for he had slipped into his glass at lunch a few drops of something, which accounted, the children thought, for the vivid streak of canary yellow in moustache and beard that were otherwise milk white. No, nothing, he murmured.

b. Russian translation:

> But no, it seemed, he needed nothing . . . he blinked, as if he would have been glad to reply kindly to her obliging offers *(she spoke seductively, though she was a bit nervous), but he could not* penetrate the gray-green somnolence . . .

c. French translation:

> No, he needed nothing . . . his eyes blinked as though he would have liked to reply kindly to these sweet attentions *(she appeared seductive if a bit embarrassed), but wasn't able to* since somnolence overcame him . . .

In this example, the parenthetical aside is set apart from the main thought by moving the comma forward to just before the conjunction, giving the reader a chance to digest one bit of information before moving on to the next. May, in the same paper, gives several other examples which show that, as she puts it, "in published translations the clarifying uses of punctuation outweigh its interpretive or creative ones".

Shifts of this type, which involve subtle changes in the placement of a punctuation mark, may well prove difficult to investigate using the current techniques of corpus analysis. But there are other expressions of simplification which lend themselves particularly well to corpus analysis. These include **lexical density** and **type–token ratio** (Baker 1995). Lexical density relates to the proportion of lexical as opposed to grammatical words in a corpus: using more grammatical and fewer lexical words is a way of building in more redundancy and making a text easier to process. There is some evidence that lexical density is generally lower in translated vs. original English text (Laviosa-Braithwaite, in prep.). Type–token ratio is a measure of the range of vocabulary that is used in a text or corpus, i.e. whether translated text uses less or more varied vocabulary than original text in the same language. Using less varied vocabulary (a narrower range) is a feature of text addressed to non-native speakers of a language, and means that these texts are easier to process. If we find that translated text tends to draw on a narrower range of vocabulary than original text in the same language, this could be interpreted as a subconscious strategy of simplification on the part of translators.

3.3 Normalisation/conservatism

"Normalisation" (or "conservatism") is a tendency to exaggerate features of the target language and to conform to its typical patterns. This tendency is quite possibly influenced by the status of the source text and the source language, so that the higher the status of the source text and language, the less the tendency to normalise. Normalisation is most evident in the use of typical grammatical structures, punctuation and collocational patterns or clichés.

In a study of simultaneous and consecutive interpreting in multilingual trials, Shlesinger (1991:150) found many instances of interpreters rounding off unfinished sentences and grammaticising ungrammatical utterances, not to mention getting rid of hesitations and false starts, even when they were

intentional. Marked, rather than ungrammatical structures, are also often "normalised" in translation, as in the following example from Faulkner's *Sanctuary* (Ben-Shahar 1994:212):

(5) a. English original:
He couldn't hardly walk, even.

b. Hebrew translation (standard Hebrew phrase):
He could hardly stand on his feet.

There also seems to be a very noticeable tendency to avoid carrying over experimental uses of punctuation in translated text. Studies by Malmkjær (forthcoming, b), May (forthcoming) and Vanderauwera (1985) all offer extensive evidence of shifts towards normalising punctuation, even when the source writers are known for their experimental use of punctuation.

3.4 Levelling out

"Levelling out" is not a feature that has received much attention in the literature. It concerns the tendency of translated text to gravitate towards the centre of a continuum. Unlike normalisation, which is target-language dependent (in the sense of exaggerating features of the target language), the process of levelling out is neither target-language nor source-language dependent. It involves steering a middle course between any two extremes, converging towards the centre, with the notions of centre and periphery being defined from within the translation corpus itself. For example, there is some evidence that the individual texts in an English translation corpus are more like each other in terms of features such as lexical density, type–token ratio and mean sentence length than the individual texts in a comparable corpus of original English. Laviosa-Braithwaite (1996, in prep.) found that variance (a statistical measure of heterogeneity) is consistently lower on lexical density and type–token ratio – and highly significantly so on mean sentence length – for the translated section of *The Guardian* than for a comparable original section of the same newspaper.

Another example comes from a study of shifts along the oral–literate continuum (Shlesinger 1989). Shlesinger set out to establish whether simultaneous interpreting effects a change in the level of orality or literacy of a text. Interestingly, her overall finding was that oral texts take on more literate features in simultaneous interpreting and literate texts become more oral. In other words, the process of translation tends to move texts towards the centre of the oral–literate continuum, to locate them away from either extreme. In Shlesinger's own words,

> Simultaneous interpretation exerts an equalizing effect on the position of a text on the oral–literate continuum, i.e. it diminishes the orality of markedly oral texts and the literateness of markedly literate ones. . . . The range of the oral–literate continuum is reduced in simultaneous interpretation. (Shlesinger 1989:96)

Again, this suggests strongly that translation tends to pull various textual features towards the centre, to move away from extremes.

4. Conclusion

Translation studies is, in a sense, turning corpus linguistics on its head. Corpus linguistics has traditionally started from the concrete, from such things as frequency lists and concordances, because the interest has always been in individual languages and largely applied in nature. Teachers wanted to know what words they should give priority to in their foreign language classes, lexicographers wanted to know which words were worth including in the dictionary and wanted to establish the collocational and syntactic environment of each word; always starting from the language itself, from concrete forms and patterns, and it is these forms and patterns themselves that have been and remain of immediate interest to the corpus linguist. The situation is quite different in corpus-based translation studies, with the exception of some applied extensions, such as using concordances to establish equivalents. Translation scholars are ultimately not interested in the words or syntactic structures themselves. What they are interested in are abstract, global notions such as explicitation and simplification, which are independent of specific languages and have various manifestations on the surface.

What I believe is important to clarify at this stage is the definitions of the individual features we are interested in investigating, the range of expressions they might have on the surface and, depending on that, the kind of techniques we might use for analysing them and indeed whether they are analysable at all using the current techniques of corpus analysis.

Bibliography

Baker, M. 1995. "Corpora in Translation Studies: an overview and some suggestions for future research". *Target* 7.223-43.

Ben-Shahar, R. 1994. "Translating literary dialogue: a problem and its implications for translation into Hebrew". *Target* 6.195-221.

Johansson, S. 1995. "Mens sana in corpore sano: on the role of corpora in linguistic research". *The European English Messenger* 4.19-25.

Lauridsen, K. 1996. "Text corpora and contrastive linguistics: which type of corpus for which type of analysis?". *Languages in Contrast: Papers from a Symposium on Text-based Cross-Linguistic Studies* ed. by K. Aijmer, B. Altenberg & M. Johansson, 63-71. Lund: Lund University Press.

Laviosa-Braithwaite, S. 1996. "The English comparable corpus: a resource and a methodology". Paper presented at "Translation Studies: Unity in Diversity", a conference held at Dublin City University, 9–11 May 1996.

Laviosa-Braithwaite, S. in prep. The English Comparable Corpus (ECC): A resource and a methodology for the empirical study of translation. PhD Thesis, UMIST, Manchester.

Malmkjær, K. forthcoming, a. "Love thy neighbour: will parallel corpora endear linguists to translators?".

Malmkjær, K. forthcoming, b. "Punctuation in Hans Christian Andersen's stories and in their translations into English". To appear in *Non-verbal Communication in Translation* ed. by F. Poyatos, Amsterdam: John Benjamins.

May, R. forthcoming. "Sensible elocution: how translation works in and upon punctuation". To appear in *The Translator*.

Sager, J. C. 1984. "Reflections on the didactic implications of an extended theory of translation". *Die Theorie des Übersetzens und ihr Aufschlußwert für die Übersetzungs- und Dolmetschdidaktik / Translation theory and its implementation in the teaching of translating and interpreting* ed. by W. Wilss & G. Thome, 333–43. Tübingen: Gunter Narr .

Sager, J. C. 1994. *Language engineering and translation: consequences of automation.* Amsterdam: John Benjamins.

Shlesinger, M. 1989. Simultaneous interpretation as a factor in effecting shifts in the position of texts on the oral-literate continuum. MA Thesis, Tel Aviv University.

Shlesinger, M. 1991. "Interpreter latitude vs. due process: simultaneous and consecutive interpretation in multilingual trials". *Empirical research in translation and intercultural studies* ed. by S. Tirkkonen-Condit, 147–55. Tübingen: Gunter Narr.

Vanderauwera, R. 1985. *Dutch novels translated into English.* Amsterdam: Rodopi.

IV. Language Engineering