Corpora in translation studies

DOROTHY KENNY

**Corpus linguistics** is the branch of linguistics that studies language on the basis of corpora, i.e., 'bodies of texts assembled in a principled way' (Johansson 1995:19). A text, in turn, can be defined as 'an instance of language in use, either spoken or written: a piece of language behaviour which has occurred naturally, without the intervention of the linguist' (Stubbs 1996:4). Corpus linguists thus take an empirical approach to the description of language: they insist upon the primacy of authentic, attested instances of use, an approach which has been mirrored in recent years by developments in descriptive translation studies. For example, scholars such as Holmes (1988:101) have expressed dissatisfaction with the use of introspection by translation theorists, and Toury (1980a:79-81) has decried approaches that view translations as idealized, speculative entities, rather than observable facts. Toury (1980a: 81) concedes that isolated attempts have been made to describe and explain actual translations but calls for a whole methodological apparatus that would make individual studies transparent and repeatable. In this regard he shares the same concerns as corpus linguists such as Atkins et al. (1992), Engwall (1994), Sinclair (1991) and Stubbs (1993, 1995, 1996), who have variously addressed issues like corpus composition and bias, the complementary roles of intuition and observation in linguistic research, and the limitations of the computational and statistical tools currently in use in the processing of corpora. And although Toury (1980a:61) bemoaned the lack of 'strict statistical methods for dealing with translational norms, or even to supply sampling rules for actual research' in the mid-1970s, much has been achieved in corpus linguistics since then, and theorists such as Baker (1993, 1995, 1997) have been instrumental not only in incorporating the methods and tools of corpus linguistics into descriptive translation studies, but also in highlighting the particular challenges that translation poses for corpus studies. Before moving on to the specific details of corpus-based translation studies, however, it is worth mentioning some issues of interest to both translation-oriented and general corpus studies.

**Corpus design and basic processing**

Within corpus linguistics generally, the term **corpus** is usually used to mean 'any collection of running texts ... held in electronic form and analysable automatically or semi-automatically (rather than manually), (Baker 1995:226). The fact that corpora are held
electronically, that is in a form accessible to computers, means that huge quantities of text can be stored: contemporary monolingual corpora such as the British National Corpus (BNC) and the Cobuild Bank of English, for instance, run to 100 million and 200 million words respectively (British National Corpus 1995; Stubbs 1996: xviii). Corpora differ from other large collections of machine-readable text (for example, archives or electronic text libraries), in that they are built "according to explicit design criteria for a specific purpose" (Atkins et al. 1992:1). Design criteria crucially depend on the envisaged use of the corpus and centre on the idea that corpora should be somehow 'representative' of a particular type of language production and/or reception. Design criteria typically involve decisions such as whether spoken and written language are to be included, what text types should be accounted for, what period of text production is to be covered, and whether text samples or full texts are to be included (Atkins et al. 1992; Baker 1995:229-30; Sinclair 1991). Once a suitable breakdown of text types, author profiles, etc., has been decided upon, the actual texts chosen for inclusion in a corpus can be selected randomly. Alternatively, a corpus--builder can intervene more deliberately in the selection of actual texts, choosing texts according to increasingly specific criteria. The two approaches are exemplified by the British National Corpus (1995) and Engwall (1994). Finally, corpus-builders must secure permission from copyright-holders in order to store texts in electronic form and to use them in subsequent research. Baker (1995:234) and Atkins et al. (1992:4) discuss the often thorny issue of copyright. Once texts have been selected for inclusion in a corpus, a decision has to be made as to how they should be represented in electronic form. Basic mark-up may involve indicating the main divisions in a text, for instance, or the addition of descriptive headers to individual texts. Johansson and Hofland (1994) and Johansson et al. (1996) describe how a bilingual corpus of Norwegian and English texts is marked up using a scheme compatible with that of the Text Encoding Initiative (see Sperberg-McQueen and Burnard 1994). Higher level mark-up may include part-of-speech tagging for every word in the corpus, or even syntactic or semantic annotation (Leech 1991). The level of mark-up that a corpus is subjected to will have implications for the kind of electronic processing the corpus can undergo. Raw corpora, i.e., untagged, unparsed corpora, can be treated as sequences of characters in running text that are delimited by spaces, in other words, sequences of orthographic words. Such running words or tokens can be counted and the number of times a single form or type appears can be calculated. The frequency of occurrence of single types can thus be worked out for the whole corpus. The number of different types compared to the total number of tokens in the corpus yields the type-token ratio for that corpus. Type-token ratios tell us something about the variety of the vocabulary used in a corpus (see Baker 1995:236). Another measure, that of lexical density, indicates what percentage of running words is made up of lexical (vocabulary) words, and hence what percentage is occupied by grammatical words. In simplistic terms, a low lexical density would mean high redundancy and hence predictability in a text (see Stubbs 1996:73). A further type of processing outputs a KWIC (key word in context) concordance for an input word, revealing the contexts in which the input word actually occurs in the
Corpora in translation studies

Most of the corpora mentioned so far are monolingual and serve the needs of linguists in general. Translation scholars, however, may have different needs, for example for corpora that contain data from more than one language, although this is not necessarily the case: familiar monolingual corpora, especially those containing specialized texts, can be used in translation pedagogy, to reinforce students' knowledge of normal target language patterns, or in terminology extraction (Pearson, forthcoming; Sager 1990:130). Baker (1995) describes various types of electronic corpora that are of specific interest to translation scholars. In Baker's terminology, a parallel corpus consists of texts originally written in a language A alongside their translations into a language B. Parallel corpora have already been compiled for several language pairs, including English-French (Salkie 1995; and see also Church and Gale 1991, who use the Canadian Hansard corpus of parliamentary proceedings), English-Italian (Marinai et al. 1992), English-Norwegian (Johansson and Hofland 1994; Johansson et al. 1996), and English-German (Schmied and Schäffler 1994; 1996). Alignment techniques may be used to provide explicit links between source sentences and target sentences, or source words and target words. Johansson and Hofland (1994) provide a useful summary of alignment procedures based on comparisons of source and target-sentence length and the use of predefined lexical correspondences between source and target languages. Parallel corpora can be used to provide information on language-pair specific translational behaviour, to posit certain equivalence relationships between lexical items or structures in source and target languages (Kenny, forthcoming; Marinai et al. 1992), or to study the phenomenon of translationese (Schmied and Schäffler 1996). Typical applications of parallel corpora include translator training, bilingual lexicography and machine translation. Malmkjler (1993) has also suggested that parallel corpora that include appropriate information about translators' backgrounds could supply useful data for psycho linguists investigating the differences between first and second language acquisition.

Baker (1995:232) uses the term multilingual corpus to refer to 'sets of two or more monolingual corpora in different languages, built up in either the same or different institutions on the basis of similar design criteria'. A multilingual corpus therefore consists of texts that have not been translated; rather they are all originals in their respective languages. As an example of a multilingual corpus Baker gives the corpus compiled by the
Council of Europe Multilingual Lexicography Project. A multilingual corpus can be used for contrastive linguistic work (Aijmer and Altenberg 1996), including bilingual lexicography. Baker (1995:233), however, has expressed reservations about the usefulness of multilingual corpora in theoretical translation studies, claiming that work using multilingual corpora is based on the erroneous assumption that ‘there is a natural way of saying anything in any language, and that all we need to do is to find out how to say something naturally in language A and language B’. Even in contrastive linguistics, multilingual corpora can present practical problems if the texts in the different languages are not well matched in terms of genre and situational factors (Aijmer and Altenberg 1996:14). Baker’s comparable corpus consists of a collection of texts originally written in a language, say English, alongside a collection of texts translated (from one or more languages) into English. Of the three types of corpus she identifies, it is the comparable corpus that Baker suggests will reveal most about the specific features of translated text, i.e., those features that occur exclusively, or with unusually low or high frequency, in translated text as opposed to other types of text production, and that cannot be traced back to the influence of anyone particular source text or language. If such features were to be discovered in a comparable corpus of, say, English, and were subsequently confirmed by studies involving comparable corpora of other languages, they could then be considered candidates for translation UNIVERSALS. Drawing on research conducted by Shlesinger (1991), Toury (1991a) and Vanderauwera (1985), Baker (1993:243-5) outlines the following hypotheses whose universal status could be investigated using comparable corpora: translated texts tend to be more explicit, unambiguous, and grammatically conventional than their source texts or other texts produced in the target language; they also tend to avoid repetitions that occur in the source text and to exaggerate features of the target language. Baker (1997) also suggests that translation may involve some kind of 'levelling out' by gravitating towards the centre of continua, such as that between orality and literacy (Shlesinger 1989b). Some of these hypotheses have already been investigated on a limited scale by researchers such as Shama’a (in Baker 1995), Toury (1980a: 129ff.) and Puurtinen (1995), although these analyses were done manually. Corpus linguistic techniques allow much more powerful generalizations to be made about translation, however. If, for example, translations were to display a high type-token ratio, low lexical density and low sentence length (which can all be calculated automatically using software like Scott’s 1996 WordSmith Tools) vis-a-vis other texts in the same language, then this would support the ‘simplification hypothesis’, for example. This is the line pursued by Laviosa-Braithwaite (1996) in her research using an English comparable corpus. Likewise, increased text length and a disproportionate number of explanatory vocabulary words and conjunctions could support the ‘EXPLICITATION hypothesis’ (Baker 1997:180-1). The attested frequency of certain vocabulary items may point towards greater standardization in translated texts than in original texts, a finding suggested by Gellerstam’s (1986) research using a Swedish comparable corpus. The idea that translations are more conventional than their source texts or other target language texts (cf. Toury 1980a:136) can also be tested by investigating
collocational patterns. Corpus linguistics provides interesting techniques for spotting statistically significant and even unconventional collocational patterns in vast quantities of text (Clear 1993; Louw 1993) and such techniques are being extended to bilingual corpora (Peters and Picchi 1996). Corpora and the software for processing them undoubtedly provide translation scholars with powerful tools for studying the very nature of translation. Some translation theorists have, however, sounded a note of caution: Malmkjler (forthcoming) warns that the bulk of statistical evidence provided by corpora may lead scholars to 'treat as marginal, if not exactly ignore, problematic cases'. Malmkjler also argues that the selection of translated texts for inclusion in a parallel corpus can affect what the observer notices to an undesirable degree, and that 'a parallel corpus still only provides, for each instance, the result of one individual's introspection, albeit contextually and cotextually informed' (ibid.), thus making a case for corpora that contain several translations of a single source text. Finally, Malmkjler stresses that 'in order to be able to provide any kinds of explanation of the data provided by the corpus, rather than mere statistics, analysts really need substantially more context than computers tend to search and display' (ibid.). Comparable corpora too have their problems: it is in the very nature of translation that new genres are introduced from one literature to another, and there may be nothing 'comparable' in the host literature to a text introduced to it through translation from another textual tradition. This difficulty is similar to one faced by scholars working with lesser-used languages: the only exemplars of many (non-literary) text types in Irish Gaelic, for instance, are translations, mainly from English; there are no 'native' texts with which to compare translations. The effects of the economy of translation have also been felt by Johansson and Hofland (1994:26), whose choice of texts for selection in their English-Norwegian bidirectional, parallel corpus has been limited by the fact that 'a large number and a wide range of texts have been translated into Norwegian, but far less in the other direction'. Perhaps the greatest challenge that faces corpus-based research into translation stems from the fact that corpus linguistics has always been data driven: it has proceeded from the bottom up, using concrete facts to make generalizations about particular languages (Baker 1997:185). Much current translation scholarship, however, proceeds top down: theorists are interested in finding evidence to support abstract hypotheses. Translation studies thus makes very particular demands on corpora, and ongoing research in translation studies may lead to new ways of looking at corpora, just as corpora are already leading to new ways of looking at translation.

**Further reading**
