Chapter XX

Parallel and comparable corpora: What are they up to?

ANTHONY MCENERY & ZHONGHUA XIAO

With ever increasing international exchange and accelerated globalisation, translation and contrastive studies are more popular than ever. As part of this new wave of research on translation and contrastive studies, corpora, and multilingual corpora in particular, have a prominent role. In this chapter, we will illustrate the value of parallel and comparable corpora to translation and contrastive studies.

Since the 1980s, corpus linguistics has developed at an accelerated speed. While the construction and exploitation of English language corpora still dominate the research of corpus linguistics, corpora of other languages, particularly typologically related European languages such as French, German and Portuguese and Asian languages such as Chinese, Korean and Japanese, have also become available and have notably added to the diversity of corpus-based language studies.[1] In addition to monolingual corpora, parallel and comparable corpora have been a key focus of non-English corpus linguistics, largely because corpora of these two types are important resources for translation and contrastive studies. As Aijmer & Altenberg (1996: 12) observe, parallel and comparable corpora 'offer specific uses and possibilities' for contrastive and translation studies:

- they give new insights into the languages compared – insights that are not likely to be noticed in studies of monolingual corpora;
- they can be used for a range of comparative purposes and increase our knowledge of language-specific, typological and cultural differences, as well as of universal features;
- they illuminate differences between source texts and translations, and between native and non-native texts;
- they can be used for a number of practical applications, e.g. in lexicography, language teaching and translation.

In this chapter, we will explore the potential value of such multilingual corpora. Before we explore the value of these corpora, however, is it necessary to clarify some terminological issues.

1. Multilingual Corpora: Terminological Issues

When we refer to a corpus involving more than one language as a multilingual corpus, the term *multilingual* is used in a broad sense. A multilingual corpus, in a narrowed sense, must involve at least three languages while those involving only two languages are conventionally referred to as *bilingual* corpora. In this chapter, we are using *multilingual* and *bilingual* interchangeably. Given that corpora involving more than one language are a relatively new phenomenon, with most research hailing from the early 1990s (e.g. *the English-Norwegian Parallel Corpus (ENPC)*, see

Johansson & Hofland, 1994),[2] it is unsurprising to discover that there is some confusion surrounding the terminology used in relation to these corpora. Generally, there are three types of corpora involving more than one language:

- Type A: Source texts plus translations, e.g. *Canadian Hansard* (cf. Brown, Lai & Mercer, 1991), *CRATER* (cf. McEnery & Oakes, 1995).
- Type B: Monolingual subcorpora designed using the same sampling frame, e.g. *The Aarhus corpus of contract law* (cf. Faber & Lauridsen, 1991).
- Type C: A combination of A and B, e.g. the *ENPC* (cf. Johansson & Hofland, 1994), the *EMIILE*.[3]

Different terms have been used to describe these types of corpora. For Aijmer & Altenberg (1996) and Granger (1996: 38), type A is a translation corpus whereas type B is a parallel corpus; for McEnery & Wilson (1996: 57), Baker (1993: 248, 1995, 1999) and Hunston (2002: 15), type A is a parallel corpus whereas type B is a comparable corpus; and for Johansson & Hofland (1994) and Johansson (1998: 4) the term parallel corpus applies to both types A and B. Barlow (1995, 2000: 110) certainly interpreted a parallel corpus as type A when he developed the *ParaConc* corpus tool. It is clear that some confusion centres around the term *parallel*.

When we define different types of corpora, we can use different criteria, for example, the number of languages involved, and the content or the form of the corpus. But when a criterion is decided upon, the same criterion must be used consistently. For example, we can say a corpus is monolingual, bilingual or multilingual if we take the number of languages involved as the criterion for definition. We can also say a corpus is a translation (L2) or a non-translation (L1) corpus if the criterion of corpus content is used. But if we choose to define corpus types by the criterion of corpus form, we must use it consistently. Then we can say a corpus is parallel if the corpus contains source texts and translations in parallel, or it is a comparable corpus if its subcorpora are comparable by applying the same sampling frame. It is illogical, however, to refer to corpora of type A as translation corpora by the criterion of content while referring to corpora of type B as comparable corpora by the criterion of form. Consequently, in this paper, we will follow McEnery *et al* and Baker's terminology in referring to type A as parallel corpora and type B as comparable corpora. As type C is a mixture of the two, corpora of this type should be referred to as comparable corpora in a strict sense.

A parallel corpus can be defined as a corpus that contains source texts and their translations. Parallel corpora can be bilingual or multilingual. They can be uni-directional (e.g. from English into Chinese or from Chinese into English alone), bi-directional (e.g. containing both English source texts with their Chinese translations as well as Chinese source texts with their English translations), or multi-directional (e.g. the same piece of writing with English, French and German versions). In this sense, texts which are produced simultaneously in different languages (e.g. EU and UN regulations) also belong to the category of parallel corpora (cf. Hunston, 2002: 15). In contrast, a comparable corpus can be defined as a corpus containing components that are collected using the same sampling frame and similar balance and representativeness (cf. McEnery, 2003: 450), e.g. the *same proportions* of the texts of the *same genres* in the *same domains* in a range of *different languages* in the *same sampling period*. However, the subcorpora of a comparable corpus are not translations of each other. Rather, their comparability lies in their same sampling frame and similar balance.

By our definition, corpora containing components of varieties of the same language (e.g. *International Corpus of English (ICE)*) are not comparable corpora as suggested in the literature (e.g. Hunston, 2002: 15), because all corpora, as a source for linguistic research, have 'always been pre-eminently suited for comparative studies' (Aarts, 1998), either intra-lingual or inter-lingual. *Brown*, *LOB*, *Frown* and *FLOB* are typically designed for comparing language varieties synchronically and diachronically. The *British National Corpus (BNC)*, while designed for representing modern British English, is also a useful basis for various intra-lingual studies (e.g. spoken vs. written, monologue vs. dialogue, and variations caused by socio-economic parameters). Nevertheless, these corpora are generally not referred to as comparable corpora.

While parallel and comparable corpora are supposed to be used for different purposes (i.e. translation and contrastive studies respectively, see section 2), the two are also designed with different focuses. For a comparable corpus, the sampling frame is essential. The components representing the languages involved must match with each other in terms of proportion, genre, domain and sampling period. For a parallel corpus, the sampling frame is irrelevant, because all of the corpus components are exact translations of each other. Once the source texts are selected using a certain sampling frame, it does not apply twice to translations. However, this does not mean that the construction of parallel corpora is easier. For a parallel corpus to be useful, an essential step is to *align* the source texts and their translations, i.e. to produce a link between the two, at the sentence or word level. Yet the automatic alignment of parallel corpora is not a trivial task for some language pairs (cf. Piao, 2000, 2002).

Depending on the specific research question, a *specialised* (i.e. containing texts of a particular type, e.g. computer manuals) or a *general* (i.e. balanced, containing as many text types as possible) corpus should be used. Parallel and comparable corpora can be of either type. For terminology extraction, specialised parallel and comparable corpora are clearly of use while for the contrast of general linguistic features such as tense and aspect, balanced corpora are supposed to be more representative of any given language in general. Existing parallel corpora appear to suggest that corpora of this type tend to be specialised (e.g. contract law and genetic engineering). This is quite natural, considering the availability of translated texts by genre (in machine-readable form) in different languages (cf. Johansson & Hofland, 1994: 27; Mauranen, 2002: 166; Aston, 1999), and indeed, as will be seen later in our discussion, specialised parallel corpora can be especially useful in domain-specific translation research.[4]

While most of the existing comparable corpora are also specialised, it is relatively easier to find comparable text types in different languages. Therefore, in relation to parallel corpora, it is more likely for comparable corpora to be designed as general balanced corpora. For instance, as the *Korean National Corpus* (Park, 2001) and the *Chinese National Corpus* (Zhou & Yu, 1997) have adopted a sampling frame quite similar to that of the BNC, these corpora can form a balanced comparable corpus that makes contrastive studies of these three languages possible.

Parallel and comparable corpora are used primarily for translation and contrastive studies. The two types of corpora have their own advantages and disadvantages, and thus serve for different purposes. While the source and translated texts in a parallel corpus are useful for exploring 'how the same content is expressed in two languages' (Aijmer & Altenberg, 1996: 13),[5] they alone serve as a poor basis for cross-linguistic contrasts, because translations (i.e. L2 texts) cannot avoid the effect of translationese (cf. Hartmann, 1985; Baker, 1993: 243-5; Teubert, 1996: 247; Gellerstam, 1996; Laviosa, 1997: 315; McEnery & Wilson, 2001: 71-2; McEnery & Xiao, 2002). In contrast, while the components of a comparable corpus overcome translationese by populating the same sampling frame with L1 texts from different languages, they are less useful for the study of how a message is conveyed from one language to another. Also the development of application software for machine aided and machine translation, while it may be based on comparable data, has clearly benefited from having access to parallel data, for example to bootstrap example-based machine translation systems (see section 2). Nonetheless, comparable corpora are a useful resource for contrastive studies and translation studies when used in combination with parallel corpora. Note, however, that comparable corpora can be a poor basis for contrastive studies if the sampling frames for the comparable corpora are not fully comparable. In the section that follows, we will illustrate, through examples, the value of corpora, particularly parallel and comparable corpora, to translation and contrastive studies.

2. Corpus-based Translation and Contrastive Studies

As Laviosa (1998a) observes, 'the corpus-based approach is evolving, through theoretical elaboration and empirical realisation, into a coherent, composite and rich paradigm that addresses a variety issues pertaining to theory, description, and the practice of translation.' Corpus-based translation studies come in two broad areas: theoretical and practical (Hunston, 2002: 123). In theoretical terms, corpora are used mainly to study the translation process by exploring how an idea in one language is conveyed in another language and by comparing the linguistic features and their frequencies in translated L2 texts and comparable L1 texts. In the practical approach, corpora provide a workbench for training translators and a basis for developing applications like machine translation (MT) and computer-assisted translation (CAT) systems. In this section, we will discuss how corpora have been used in each of these areas.

Parallel corpora are a good basis for studying how an idea in one language is conveyed in another language.[6] Xiao & McEnery (2002a), for example, use an

English-Chinese parallel corpus containing 100,170 English words and 192,088 Chinese characters to explore how temporal and aspectual meanings in English are expressed in Chinese. In that study, the authors found that while both English and Chinese have a progressive aspect, the progressive has different scopes of meanings in the two languages. In English, while the progressive canonically (93.5%) signals the ongoing nature of a situation (e.g. *John is singing*, Comrie, 1976: 32), it has a number of other specific uses that do not seem to fit under the general definition of progressiveness' (Comrie, 1976: 37). These 'specific uses' include its use to indicate contingent habitual or iterative situations (e.g. *I'm taking dancing lessons this winter*, Leech, 1971: 27), to indicate anticipated happenings in the future (e.g. *We're visiting Aunt Rose tomorrow*, *ibid*: 29) and some idiomatic use to add special emotive effect (e.g. *I'm continually forgetting people's names*, *ibid*) (c.f. Leech, 1971: 27-29). In Chinese, however, the progressive marked by *zai* only corresponds to the first category above, namely, to mark the ongoing nature of dynamic situations. As such, only about 58% of situations referred to by the progressive in the English source data take the progressive or the durative aspect, either marked overtly or covertly, in Chinese translations. The authors also found that the interaction between situation aspect (i.e. the inherent aspectual features of a situation, e.g. whether the situation has a natural final endpoint) and viewpoint aspect (e.g. perfective vs. imperfective) also influences a translator's choice of viewpoint aspect. Situations with a natural final endpoint (around 65%) and situations incompatible with progressiveness (92.5% of individual-level states and 75.9% of achievements) are more likely to undergo viewpoint aspect shift and presented perfectively in Chinese translations.[7] In contrast, situations without a natural final endpoint are normally translated with the progressive marked by *zai* or the durative aspect marked by *-zhe*.

Note, however, that the direction of translation in a parallel corpus is important in studies of this kind. The corpus used in Xiao & McEnery (2002a), for example, is not suitable for studying how aspect markers in Chinese are translated into English. For that purpose, a Chinese-English parallel corpus (i.e. L1 Chinese plus L2 English) is required.

Another problem that arises with the use of a one-to-one parallel corpus (i.e. containing only one version of translation in the target language) is that the translation only represents 'one individual's introspection, albeit contextually and cotextually informed' (Malmkjær, 1998). One possible way to overcome this problem, as suggested in Malmkjær, is to include as many versions of a translation of the same source text as possible. While this solution is certainly of benefit to translation studies, it makes the task of building parallel corpora much more difficult. It also reduces the range of data one may include in a parallel corpus, as many translated texts are translated once only. It is typically texts such as literary works where multiple translations of the same work are available. These works tend to be non-contemporary and the different versions of translation are usually spaced decades apart, thus making the comparison of these versions less meaningful.

The distinctive features of translated language can be identified by comparing the translations with comparable L1 texts, thus throwing new light on the translation process and helping to identify translation norms. Laviosa (1998b), for example, in her study of L1 and L2 English narrative prose, finds that translated L2 language has four core patterns of lexical use: a relative lower proportion of lexical words over function words, a relatively higher proportion of high-frequency words over low-frequency words, relatively greater repetition of the most frequent words, and less variety in the words that are most frequently used. Other studies show that translated language is characterised, beyond the lexical level, by nominalization, simplification (Baker, 1993, 1998), explication (i.e. increased cohesion, Øverås, 1998) and sanitisation (i.e. reduced connotational meanings, Kenny, 1998). As these features are regular and typical of translated English, further research based upon these findings may not only uncover the translation norms or what Frawley (1984) calls the ‘third code’ of translation, it will also help translators and trainee translators to become aware of these problems.

McEnery & Xiao (2002), on the basis of a specialised English-Chinese parallel corpus of healthcare, found that the ratio of overt/covert marking of aspectual meanings was exceptionally low in Chinese translations. As Chinese is recognised as an aspect language (cf. Xiao & McEnery, forthcoming), the authors hypothesised that the low frequency of aspect markers was atypical of the target L1 language and was attributable to the translated nature of the data in this case. To test this hypothesis, they constructed a comparable L1 Chinese corpus using the same sampling frame and compared the frequencies of two well-established perfective aspect markers in the two datasets, namely, the translated Chinese and L1 Chinese. The experiment showed that in the translated Chinese, the two aspect markers occurred 27.32 times per 10,000 words whereas they occurred 62.33 times per 10,000 words in the comparable L1 Chinese data. A cross-tabulation between the word numbers and actual frequency counts showed a log-likelihood ratio of 49.1 for 2 degrees of freedom, which is statistically significant at the level $p<0.001$. Therefore, the authors’ null hypothesis that the difference in frequencies of aspect markers in the two datasets existed by chance was rejected and they were able to claim that translated Chinese is indeed different from L1 Chinese in terms of aspect marking.

The above studies show that translated language is translationese. The effect of source language on the translations is strong enough to make the L2 data perceptibly different from the target L1 Chinese. As such, a uni-directional parallel corpus is a poor basis for cross-linguistic contrast. This problem, however, can be alleviated by a bi-directional parallel corpus (e.g. Maia, 1998; Ebeling, 1998), because the effect of translationese is averaged out to some extent. In this sense, a well-matched bi-directional parallel corpus can become the bridge that brings translation and contrastive studies together. To achieve this aim, however, the same sampling frame must apply to the selection of source data in both languages. Any mismatch of proportion, genre, or domain, for example, may invalidate the findings derived from such a corpus.

While we know that translated language is distinct from the target L1 language, it has been claimed recently that parallel corpora represent a sound basis for contrastive studies. James (1980: 178), for example, argues that ‘translation equivalence is the best available basis of comparison’ while Santos (1996: i) claims that ‘studies based on real translations are the only sound method for contrastive analysis.’ Mauranen (2002: 166) also argues, though not as strongly as James and Santos, that translated language, in spite of its special features, ‘is part of natural language in use, and should be treated accordingly’, because languages ‘influence each other in many ways other than through translation’ (*ibid*: 165). While we agree with Mauranen that ‘translations deserve to be investigated in their own right’, as is done in Laviosa (1998b) and

McEnery & Xiao (2002), we hold a different view of the value of parallel corpora for contrastive studies. It is true that languages in contact can influence each other, but this influence is different from the influence of a source language on translations in respect to immediacy and scope. Basically, the influence of languages in contact is generally gradual (or evolutionary) and less systematic than the influence of a source language on the translated language. As such, translated language is at best an unrepresentative special variant of the target language. If this special variant is confused with the target L1 language and serves alone as the basis for contrastive studies, the results are clearly misleading to teachers and students of second languages, because contrastive studies are 'typically geared towards second language teaching and learning' (Teich, 2002: 188). Using parallel corpora alone, for example, McEnery & Xiao (2002) would have come to the misleading conclusion that aspect markers occurred only infrequently in Chinese. As Chinese as an aspect language relies heavily on aspect to encode temporal information, which is different from English which encodes both tense and aspect, this false conclusion would inevitably have an adverse effect on Chinese learners of English. Parallel corpora can serve as a useful starting point for cross-linguistic contrasts because findings based on parallel corpora invite 'further research with monolingual corpora in both languages' (Mauranen, 2002: 182). In this sense, parallel corpora are 'indispensable' to contrastive studies (*ibid*). Based on the preliminary findings in McEnery & Xiao (2002) and Xiao & McEnery (2002a), we have initiated an ESRC-funded project on contrasting tense and aspect in English and Chinese on the basis of two one-million-word L1 corpora of the two languages.

With reference to practical translation studies, as corpora can be used to raise linguistic and cultural awareness in general (cf. Hunston, 2002: 123; Bernardini, 1997), they provide a useful and effective reference tool and a workbench for translators and trainees. In this respect even a monolingual corpus is helpful. Bowker (1998), for example, found that corpus-aided translations were of a higher quality with respect to subject field understanding, correct term choice and idiomatic expressions than those undertaken using conventional resources. Bernardini (1997) also suggests that traditional translation teaching should be complemented with LCC (large corpora concordancing) so that trainees develop 'awareness', 'reflectiveness' and 'resourcefulness', the skills that 'distinguish a translator from those unskilled amateurs.'

In comparison to monolingual corpora, comparable corpora are more useful for translation studies. Zanettin (1998) demonstrates that small comparable corpora can be used to devise a 'translator training workshop' designed to improve students' understanding of the source texts and their ability to produce translations in the target more fluently. In this respect, specialised comparable corpora are particularly helpful for highly domain-specific translation tasks, because when translating texts of this type, as Friedbichler & Friedbichler (1997) observe, 'the translator is dealing with a language which is often just as disparate from his/her native language as any foreign tongue.' Several studies show that translators with access to a comparable corpus with which to check translation problems 'are able to enhance their productivity and tend to make fewer mistakes' (*ibid*) when translating into their native language. When translation is from a mother tongue into a foreign language, 'the need for corpus tools grows exponentially and goes far beyond checking grey spots in L1 language competence against the evidence of a large corpus' (*ibid*). For example, Gavioli & Zanettin (1997) demonstrate how a very specialised corpus of texts on the subject of hepatitis helps to confirm translation hypotheses and suggest possible solutions to problems related to domain-specific translation.

While monolingual and comparable corpora are of use to translation, it is difficult to generate 'possible hypotheses as to translations' with such data (Aston, 1999). Furthermore, verifying concordances is both time-consuming and error-prone, which entails a loss of productivity. Parallel corpora, in contrast, provide '[g]reater certainty as to the equivalence of particular expressions', and in combination of suitable tools (e.g. *ParaConc*), they enable users to 'locate all the occurrences of any expression along with the corresponding sentences in the other language' (*ibid*). As such, parallel corpora can help translators and trainees to achieve improved precision with respect to terminology and phraseology and have been strongly recommended for these reasons (e.g. Williams, 1996). A special use of a parallel corpus with one source text and many translations is that it can offer a systematic translation strategy for linguistic structures which have no direct equivalents in the target language. Buyse (1997), for example, presents a case study of the Spanish translation of the French clitics *en* and *y*, where the author illustrates how a solution is offered by a quantitative analysis of the phonetic, prosodic, morphological, semantic and discursive features of these structures in a representative parallel corpus, combined with the quantitative analysis of these structures in a comparable corpus of L1 target language. Another issue related to translator training is translation evaluation. Bowker (2001) shows that an evaluation corpus, which is composed of a parallel corpus and comparable corpora of source and target languages, can help translator trainers to evaluate student translations and provide more objective feedback.

Finally, in addition to providing assistance to human translators, parallel corpora constitute a unique resource for the development of machine translation (MT) systems. Starting in the 1990s, the established methodologies, notably, the linguistic rule-based approach to machine translation, have been challenged and enriched by an approach based on parallel corpora (cf. Hutchins, 2003: 511; Somers, 2003: 513). The new approaches, such as example-based MT (EBMT) and statistical MT, are based on parallel corpora. With EBMT, for example, a new input is matched against the database of already translated texts to extract suitable examples which are then combined to generate the correct translation (see Somers: *ibid*). As well as automatic MT systems, parallel corpora have also been used to develop computer-assisted translation (CAT) tools for human translators, such as translation memories (TM), bilingual concordances and translator-oriented word processor (cf. Somer, 2003; Wu, 2002).

3. Conclusion

In this chapter, we first clarified the confusion surrounding the terminology related to multilingual corpora. It was argued that consistent criteria should be applied in defining types of corpora. For us this means that parallel corpora refer to those that contain collections of L1 texts and their translations while comparable corpora refer to those that contain matched L1 samples from different languages.

The main concern of this chapter was the potential value of parallel and comparable corpora to translation and contrastive studies.[8] We maintain that while parallel corpora are well-suited to research and teaching in translation studies, they provide a poor basis for cross-linguistic contrasts if used as the sole source of data. They should most often be used in conjunction with L1 target and source corpora. These L1 target and source corpora may or may not be comparable. Parallel corpora are undoubtedly a useful starting point for contrastive research, however, and may lead to further research in contrastive studies based upon comparable corpora. In contrast, comparable corpora used alone are less useful for translation studies. Nonetheless, they certainly serve as a reliable basis for contrastive studies. It appears then that a carefully matched bi-directional parallel corpus provides a sound basis for both translation and contrastive studies. Yet the ideal bi-directional parallel corpus will often not be easy, or even possible, to build because of the heterogeneous pattern of translation between languages and genres. So we must accept that, for practical reasons alone, we will often be working with corpora that, while they are useful, are not ideal for either translation or contrastive studies.

In this chapter, we also discussed the pros and cons of the use of different types of corpora in translation and contrastive studies and evaluated proposals for possible solutions to related problems. It is our belief that as the number of parallel and comparable corpora grows, the corpus-based paradigm will soon enter the mainstream of translation and contrastive studies.

Acknowledgements

This work is supported in part by a grant (Reference No. RES-000-220135) from the Economic and social Research Council (ESRC).

Notes

1. Lists of available corpus resources invloving different languages, , both monolingual and multilingual, can be found at the websites of *Evaluations and Language Resource Distribution Agency* (*ELDA*, http://www.eida.fr/cata/tabtxt1.html), *TELTRI Research Archive of Computational Tools and Resources* (*TRACTOR*, http://tractor.bham.ac.uk/tractor/catalogue.html), *Oxford Text Archive* (*OTA*, http://ota.ahds.ac.uk) and *Linguistic Data Consortium* (*LDC*, http://www.ldc.upenn.edu/Catelog/byType.jsp).
2. It is interesting to note, however, an earlier corpus-based contrastive study, namely Hilipovic 1969, dates back as early as the 1960s.
3. An introduction to the *EMILLE* project can be found at the following URL http://www.emille.lancs.ac.uk.
4. Readers are advised to refer to Halverson (1998) for an argument for the need for representative parallel corpora.
5. This view has been challenged recently, however, notably by Mauranen (2002: 167), who argues that interpreting translation as ‘the decoding and re-encoding of fixed contents, which presumably, exist outside languages’ is ‘hardly an adequate view of either language or translation.’ However, if we

interpret the relationship between *contents* and *languages* as that between *meanings* (the *carried*) and *forms* (the *carrier*), this view is quite natural.

6. However, the quality of translation is an important factor which should be taken into serious consideration during corpus construction.
7. *Situaions*, *telic*, *individual-level states* and *achievements* are commonly used terms in aspect theory. Readers can refer to Xiao & McEnery (2002b) for a more elaborate account of situation aspect.
8. Apart from translation and contrastive studies, Botley, McEnery & Wilson (2000) give a fine account of other potential uses of parallel and comparable corpora.

References

Aarts, J. (1998) Introduction. In S. Johansson & S. Oksefjell (eds.) *Corpora and Cross-linguistic Research* (pp. ix-xiv). Amsterdam: Rodopi.

Aston, G. (1999) Corpus use and learning to translate. *Textus* 12, 289-314. Available online at http://www.sslmit.uniho.it/guy/textus.htm.

Baker, M. (1993) Corpus linguistics and translation studies: implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.) *Text and technology: in honour of John Sinclair* (pp. 233-52). Amsterdam: Benjamins.

Baker, M. (1995) Corpora in translation studies: an overview and some suggestions for future research. *Target* 7, 223-243.

Baker, M. (1999) The role of corpora in investigating the linguistic behaviour of professional translators. *International Journal of Corpus Linguistics* 4, 281-98.

Barlow, M. (1995) *A guide to ParaConc*. Huston: Athelstan.

Barlow, M. (2000) Parallel texts and language teaching. In S. Botley, A. McEnery & A. Wilson (eds.) *Multilingual corpora in teaching and researching* (pp. 106-15). Amsterdam: Rodopi.

Bernardini, S. (1997) A ‘trainee’ translator’s perspective on corpora. Paper presented at *Corpus use and learning to translate* held at Bertinoro, Nov. 1997. Available online at URL http://www.sslmit.unibo.it/introduz.htm.

Botley, S., McEnery, A. and Wilson, A. (2000) *Multilingual corpora in teaching and research*. Amsterdam : Rodopi.

Bowker, L. (1998) Using specialised native-language corpora as a translation resource: a pilot study. *Meta* 43(4). Available online at the following URL : http://www.erudit.org/meta/1998/v43/n4/index.html.

Bowker, L. (2001) Towards a methodology for a corpus-based approach to translation evaluation. *Meta* 46(2), 345-64.

Brown, P., Lai, J. and Mercer, R. (1991) Aligning sentences in parallel corpora. In *29th Annual Meeting of the Association for Computational Linguistics* (pp. 169-176). Berkeley, CA.

Buyse, K. (1997) The study of multi- and unilingual corpora as a tool for the development of translation studies: a case study. Paper presented at *Corpus use and learning to translate* held at Bertinoro, Nov. 1997.

Comrie, B. (1976) *Aspect*. Cambridge: Cambridge University Press.

Ebeling, J. (1998) Contrastive linguistics, translation, and parallel corpora. *Meta* 43(4).

Faber, D. and Lauridsen, K. (1991) The compilation of a Danish-English-French corpus in contract law. In S. Johansson & A-B. Stenström (eds.) *English computer corpora. Selected papers and research guide* (pp. 235-43). Berlin: Mouton de Gruyter.

Filipovic, R. (1969) The choice of the corpus for the contrastive analysis of Serbo-Croatian and English. In *The Yugoslav Serbo-Crotian-English contrastive Project B Studies 1* (pp. 37-46). Institute of Linguistics, University of Zagreb.

Frawley, W. (1984) Prolegomenon to a theory of translation. In W. Frawley (ed.) *Translation: Literary, linguistic and philosophical perspectives* (pp. 159-75). London & Toronto: Associated University Press.

Friedbichler, I. and Friedbichler, M. (1997) The potential of domain-specific target-language corpora for the translator's workbench. Paper presented at *Corpus use and learning to translate* held at Bertinoro, Nov. 1997.

Gavioli, L. and Zanettin, F. (1997) Comparable corpora and translation: a pedagogic perspective. Paper presented at *Corpus use and learning to translate* held at Bertinoro, Nov. 1997.

Gellerstam, M. (1996) Translations as a source fro cross-linguistic studies. In K. Aijmer, B. Altenberg and M. Johansson (eds.) *Language in contrast: papers from a symposium on text-based cross-linguistic studies, Lund, March 1994* (pp. 53-62). Lund: Lund University Press.

Grange, S. (1996) From CA to CIA and back: an integrated approach to computerised bilingual and learner corpora. In K. Aijmer, B Altenberg and M. Johansson (eds.) *Language in contrast: papers from a symposium on text-based cross-linguistic studies, Lund, March 1994* (pp. 38-51). Lund: Lund University Press.

Halverson, S. (1998) Translation studies and representative corpora: establishing links between translation corpora, theoretical/descriptive categories and a conception of the object of study. *Meta* 43(4).

Hartmann, R. (1995) Contrastive textology. *Language and Communication* 5, 25-37.

Hutchins, J. (2003) Machine translation: general overview. In R. Mitkov (ed.) *Oxford handbook of computational linguistics* (pp. 501-11). Oxford: Oxford University Press.

Hunston, S. (2002) *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

James, C. (1980) *Contrastive analysis*. London: Longman.

Johansson, S. (1998) On the role of corpora in cross-linguistic research. In S. Johansson & S. Oksefjell (eds.) *Corpora and cross-linguistic research: Theory, method, and case studies* (pp. 3-25). Amsterdam: Rodopi.

Johansson, S., Ebeling, G. and Hofland, K. (1996) Coding and aligning the English-Norwegian parallel corpus. In K. Aijmer, B Altenberg and M. Johansson (eds.) *Language in contrast: papers from a symposium on text-based cross-linguistic studies, Lund, March 1994* (pp. 87-112). Lund: Lund University Press.

Johansson, S. and Hofland, K. (1994) Towards an English-Norwegian parallel corpus. In U. Fries, G. Tottie and P. Schneider (eds.) *Creating and using English language corpora* (pp. 25-37). Amsterdam: Rodopi.

Johansson, S. and Oksefjell, S. (1998) *Corpora and cross-linguistic research: Theory, method, and case studies*. Amsterdam: Rodopi.

Kenny, D. (1998) Creatures of habit? What translators usually do with words? *Meta* 43(4).

Laviosa, S. (1997) How comparable can ‘comparable corpora’ be? *Target* 9, 289-319.

Laviosa, S. (1998a) The corpus-based approach: a new paradigm in translation studies. *Meta* 43(4).

Laviosa, S. (1998b) Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta* 43(4).

Leech, G. (1971) *Meaning and the English verb*. London: Longman.

Maia, B. (1998) Word order and the first person singular in Portuguese and English. *Meta* 43(4).

Malmkjær, K. (1998) Love thy neighbour: will parallel corpora endear linguists to translators? *Meta* 43(4).

Mauranen, A. (2002) Will ‘translationese’ ruin a contrastive study? *Languages in Contrast* 2(2), 161-86.

McEnery, A. (2003) Corpus linguistics. In R. Mitkov (ed.) *Oxford handbook of computational linguistics* (pp. 448-63). Oxford: Oxford University Press.

McEnery, A. and Oakes, M. (1995) Sentence and word alignment in the CRATER project: methods and assessment. In S. Warwick-Armstrong (ed.) *Proceedings of the Association for Computational Linguistics Workshop SIG-DAT Workshop*. Dublin.

McEnery, A and Wilson, A. (1996) *Corpus linguistics* (1st edition). Edinburgh: Edinburgh University Press.

McEnery, A and Wilson, A. (2001) *Corpus linguistics* (2nd edition). Edinburgh: Edinburgh University Press.

McEnery, A and Xiao, Z. (2002) Domains, text types, aspect marking and English-Chinese translation. *Journal of Languages in Contrast* 2(2), 211-31.

Øverås, S. (1998) In search of the third code: An investigation of norms in literary translation. *Meta* 43(4).

Park, B. (2001). Introducing Korean National Corpus. Talk presented at Corpus Research Group, Lancaster University, Nov. 19th, 2001. See also http://www.sejong.or.kr/english/index.html

Pearson, J. (1998) *Terms in context*. Amsterdam: Benjamins.

Piao, S. (2000) *Sentence and word alignment between Chinese and English*. PhD thesis. Lancaster university.

Piao, S. (2002) Word alignment in English-Chinese parallel corpora. *Literary and Linguistic Computing* 17(2), 207-30.

Santos, D. (1996). *Tense and aspect in English and Portuguese: a contrastive semantical study*. PhD thesis. Universidade Tecnica de Lisboa.

Somers, H. (2003) Machine translation: latest developments. In R. Mitkov (ed.) *Oxford handbook of computational linguistics* (pp. 512-28). Oxford: Oxford University Press.

Teich, E. (2002) System-oriented and text-oriented comparative linguistic research: cross-linguistic variation in translation. *Languages in Contrast* 2(2), 187-210.

Teubert, W. (1996) Comparable or parallel corpora? *International Journal of Lexicography* 9(3), 238-64.

Williams, A. (1996) A translator’s reference needs: dictionaries or parallel texts. *Target* 8(2), 277-99.

Wu, D. (2002) Conception and application of computer-assisted translation. Paper presented at *International Symposium on Contrastive and Translation Studies between Chinese and English*. Shanghai. August 2002.

Xiao, Z. and McEnery, A. (2002a) A corpus-based approach to tense and aspect in English-Chinese translation. Paper presented at *International Symposium on Contrastive and Translation Studies between Chinese and English*. Shanghai 2002.

Xiao, Z. and McEnery, A. (2002b) Situation aspect as a universal aspect: implications for artificial languages. *Journal of Universal Language* 3(2), 139-77.

Xiao, Z. and McEnery, A. (Forthcoming) *Aspect in Chinese*. Amsterdam: John Benjamins.

Zanettin, F. (1998) Bilingual comparable corpora and the training of translators. *Meta* 43(4).

Zhou, Q. and Yu, S. (1997) Annotating the Contemporary Chinese Corpus. *International Journal of Corpus Linguistics* 2(2), 239-58.