Unit 5 Multilingual corpora

5.1 Introduction

Having covered some of the important concepts and practices in corpus linguistics, we will consider, in units 5-8, a range of related issues. In this unit we will consider the multilingual dimension of corpus linguistics.

While the construction and exploitation of English language corpora still dominate corpus linguistics, corpora of other languages, particularly typologically related European languages such as French, German and Portuguese as well as Asian languages such as Chinese and Japanese, have also become available (see the website accompanying this book for a survey of well-known and influential corpora) and have added notably to the diversity of corpus-based language studies. 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 and 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 gained from the study 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 unit, we will address issues related to multilingual corpora. This unit consists of four sections. Unit 5.2 is concerned with terminological issues. Unit 5.3 introduces the alignment of parallel corpora. Unit 5.4 concludes the unit.

5.2 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 narrow sense, must involve at least three languages while those involving only two languages are conventionally referred to as *bilingual* corpora. In this book, we are generally using *multilingual* to refer to corpora containing two or more languages. Given that corpora involving more than one language are a relatively new phenomenon, with most related research hailing from the early 1990s (e.g. the English-Norwegian Parallel Corpus (ENPC), see Johansson and Hofland 1994), 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 and Mercer 1991), and Crater (cf. McEnery and Oakes 1995);

- Type B: Monolingual subcorpora designed using the same sampling techniques, e.g. the Aarhus corpus of contract law (cf. Faber and Lauridsen 1991);
- Type C: A combination of A and B, e.g. the ENPC (cf. Johansson and Hofland 1994) and EMILLE (cf. Baker et al 2004).

Different terms have been used to describe these types of corpora. For Aijmer and Altenberg (1996) and Granger (1996: 38), type A is a translation corpus whereas type B is a parallel corpus; for McEnery and 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 and Hofland (1994) and Johansson (1998: 4) the term parallel corpus applies to both types A and B. Barlow (1995, 2000: 110) assumed a parallel corpus was type A when he developed the ParaConc corpus tool (see case study 6 in Section C). It is clear that some confusion centres around the term *parallel*.

In this book a *parallel* corpus is one which is composed of source texts and their translations in one or more different languages while a *comparable* corpus refers to one which is composed of L1 data collected from different languages using the same sampling techniques. 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 techniques. 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 book, we will follow McEnery and Wilson (1996) and Baker's (*ibid*) 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.

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 techniques 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 comparable sampling techniques and similar balance (see unit 2).

By our definition, corpora containing components of varieties of the same language (e.g. the International Corpus of English, see unit 7.6) are not comparable corpora because all corpora, as a resource for linguistic research, have 'always been

pre-eminently suited for comparative studies' (Aarts 1998), either intra-lingual or inter-lingual. The Brown, LOB, Frown and FLOB corpora 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 sociolinguistic variables). Nevertheless, these corpora are generally not referred to as comparable corpora. In this book we label corpora containing components of varieties of the same language *comparative corpora*.

While parallel and comparable corpora are supposed to be useful for different purposes (i.e. translation and contrastive studies respectively, see unit 10.6), 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 (see unit 5.3), 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 specialized (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, specialized 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 specialized (e.g. contract law and genetic engineering). This is quite natural, considering the availability of translated texts (in machine-readable form) by genre in different languages (cf. Johansson and Hofland 1994: 27; Mauranen 2002: 166; Aston 1999), and indeed, as will be seen unit 10.6, specialized parallel corpora can be especially useful in domain-specific translation research, though readers are advised to refer to Halverson (1998) for an argument for the need for representative parallel corpora. While most of the existing comparable corpora are also specialized, 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, the Chinese National Corpus (Zhou and Yu 1997) and the Polish National Corpus have adopted a sampling frame quite similar to that of the BNC (see unit 7.2), these corpora can form a balanced comparable corpus that makes contrastive studies for these four 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 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 and Altenberg 1996: 13), 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 and Wilson 2001: 71-72; McEnery and 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 unit 10.6). 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.

5.3 Corpus alignment

We have so far assumed that parallel corpora means *aligned parallel corpora*. It is clear that simply having a corpus containing parallel texts presents problems as well as promises. Without alignment, we cannot easily determine which sentences in the target language are translations of which in the source language. An aligned parallel corpus solves this problem and makes available to researchers, language learners, etc. information regarding the translation in the parallel corpus that the users of that corpus may not be able to provide for themselves. For example, in the aligned English-Chinese CEPC-health parallel corpus which we will use in case study 6, you will see *What is organ donation* and *Shenme shi qiguan juanzeng* align, though it is unlikely that you are capable of identifying this translation without the aid of the annotation. As well as sentence alignment, sub-sentential level alignment is also undertaken, notably phrase (multi-word unit) or word level alignment. In the above sentence, one might align *what* with *shenme*, *is* with *shi*, *organ* with *qiguan* and *donation* with *juanzeng*. Currently most multilingual corpus tools (e.g. ParaConc) only take pre-aligned parallel texts as input, though Multiconcord (cf. Woolls 2000) is able to align non-aligned parallel texts presented to the system by the user in 10 European languages. In either case, alignment is an essential step in the construction and exploitation of parallel corpora.

The aim of corpus alignment is to find translation equivalents of sentences, phrases or words between the source and translated texts in a parallel corpus. Sentence alignment is generally the first step to phrase and word alignment. The source and translated texts in an aligned parallel corpus may appear in a single file, with translation equivalents aligned together. They may also appear in separate files, with the source and target text of each translation equivalent being linked together with a unique identifier or pointer (i.e. stand-alone annotation, as recommended by CES, see unit 3.3). ParaConc only works on parallel corpora of the latter type.

In this unit we will present a non-technical description of sentence alignment only because this form of alignment can be achieved automatically with a relatively high degree of accuracy. Readers who are interested in the technical aspects of sentence alignment (e.g. the precise alignment algorithms), and word alignment can refer to Oakes and McEnery (2000), Piao (2000, 2002) and Simard, Foster, Hannan, Macklovitch and Plamondon (2000) for more information.

There are basically three approaches to sentence alignment: statistical (probabilistic), linguistic (knowledge-based) and hybrid. The statistical approach to sentence alignment is generally based on sentence length in terms of words (e.g.

Brown, Lai and Mercer 1991) or characters (e.g. Gale and Church 1993) per sentence while the lexical approach (e.g. Haruno, Ikehara and Yamazaki 1996; Kupiek 1993) uses morpho-syntactic information to explore similarities between languages. The lexical approach may achieve more accurate alignment than the statistical approach, but it is 'necessarily slow' and not suitable for aligning large corpora (cf. Brown et al 1991: 169). The most widely used approach to sentence alignment is the hybrid approach, which integrates linguistic knowledge into a probabilistic algorithm to achieve improved accuracy (e.g. Hofland 1996; McEnery and Oakes 1996; Simard et al 2000: 49-50). As the research of alignment has focused on closely related European language pairs, sentence alignment among these language pairs has achieved a very high precision rate, e.g. 100% for Polish-English alignment (McEnery and Oakes 2000: 7), 98% for English-French (McEnery and Oakes 1996) and English-Norwegian alignment (Johansson and Hofland 1994). Recently, however, great success has also been achieved for typologically different languages such as English and Chinese. Piao (2000: 153), for example, reported a quite stable performance for his alignment system, with a success rate ranging from 92.93% to 100% on texts of various sizes and domains.

5.4 Unit summary and looking ahead

In this unit, 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 multilingual 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. In this unit we also introduced corpus alignment, an important process applied to parallel corpora.

In unit 10.6, we will return to discuss the use of comparable and parallel corpora in contrastive and translation studies. Readers interested in exploring the further use of multilingual corpora in language studies are advised to read Aijmer, Altenberg and Johansson (1996), Johansson and Oksefjell (1998), and Botley, McEnery and Wilson (2000). In the next unit, we will discuss how to make statistical claims in corpus-based language studies.