Lincoln Fernandes
Faculdades Barddal
lico.fernandes@gmail.com

Corpora in Translation Studies: revisiting Baker's typology

Abstract: This article aims to offer a more flexible way of classifying the different types of corpora in the descriptive and applied branches of Translation Studies. To do so, it goes on to reformulate Baker's (1995) typology of corpora by discussing each of the selection criteria on which corpora are generally designed along with their attributes. The final result is expected to be better typological resources for any translation researcher or *translator educator* interested in carrying out corpus-based work.
Keywords: Baker (1995), typology, corpora, Translation Studies.

Resumo: Este artigo tem como objetivo oferecer uma maneira mais flexível de classificar os diferentes tipos de corpora nos ramos descritivo e aplicado dos Estudos da Tradução. Para tal, o mesmo reformula a typologia de corpora proposta por Baker (1995) através da discussão de cada um dos critérios de seleção sobre os quais corpora são geralmente desenhados juntamente com seus atributos. Espera-se que o resultado final seja melhores recursos tipológicos para qualquer pesquisador em tradução ou *educador de tradutores* interessados em elaborar trabalhos baseados em corpora.
Descritores: Baker (1995), tipologia, corpora, Estudos da Tradução.

Introduction

It has been over a decade now since Baker (1995) proposed a typology for corpora in translation research and pedagogy. In discussing it, the Manchester-based scholar puts forward three main types of corpora "in anticipation of the surge of activity" (p. 230) in this particular area, namely *comparable corpora*, *multilingual corpora* and *parallel corpora* (see Classification Criteria for Corpora in TS below). However, because of the rapid development in Corpus-based Translation Studies (CTS) as well as the "need for greater standardization" (Zanettin, 2000, p. 105), Baker's categorization – as any other attempt to classify something into categories – requires further refinement in order to describe more accurately all types of corpora which have recently come up in the field. In this context, this paper

calls for a rearrangement of Baker's proposal and attempts to offer a more flexible way of classifying the different types of corpora in the descriptive and applied branches of CTS. To begin with, let me first highlight some specific connotations that the term corpus has acquired over the years in studies involving the use of corpora within Translation Studies (TS).

What does the term *corpus* mean in CTS?

The term *corpus* when used in the context of CTS has more specific connotations than traditional definitions such as the one provided, for instance, by the *Oxford Concise English Dictionary* (i.e. "a large collection of written or spoken texts"), which does not carry such connotations. These connotations can be associated with at least four main attributes: *electronic form*, *size*, *representativeness*, and *open-endedness*.

Electronic form – for many years the word 'corpus' was only associated with hard-copy texts, but after the advent of the computer, it nearly always implies a collection of texts held in electronic form which can be read and analyzed automatically or semi-automatically rather than manually (Baker, 1995, p. 226).

Size – from a historical perspective, corpus-based studies have often relied on huge amounts of data in order to increase empirical evidence and knowledge about the world of our experience (see Sinclair, 1991). As a consequence of this fact, the term "corpus" has traditionally been associated with vast quantities of data extracted from large collections of text; nevertheless, in the context of CTS the term has also been used to describe what came to be known as "small-scale corpora" in translator education (see Bowker, 1996 and Pearson, 1998). Therefore the issue of corpus size in CTS becomes a relative one in the sense that qualitative aspects sometimes may be more relevant than quantitative ones. Another important aspect related to size has to do with the use of full texts instead of text fragments. According to Baker (ibid.), corpora which consist of full texts are by and large far more useful than those which consist of text fragments. This is so because full texts allow for the examination of not only microlevel units such as words, phrases and sentences, but also the way texts are structured in their entirety, that is to say, how texts are formed by chapters, sections, paragraphs and so on (p. 225).

Representativeness – in building a corpus covering an area of interest, researchers must know to what extent and in what respects their corpus is representative enough to serve its purpose. Thus, the selec-

tion of texts in a representative corpus is not only related to size, but also to a careful description of what the corpus is intended to represent (see Halverson, 1998 and Kennedy, 1998). Moreover, in the case of parallel corpora one has to establish unequivocally the source texts of the translations as there are times when "a multitude of candidates for a source text may exist" (Toury, 1995, p. 74). By so doing, researchers would not be faced with an injudicious choice of source text, which could certainly lead them astray and consequently produce rather unfortunate results.

Open-endedness – this refers to the flexibility that a corpus in translation studies should have to enable researchers to answer specific research questions. In other words, by means of an opened-ended corpus researchers can select and use the texts of this corpus for different types of comparisons and studies (see Olohan, 2004, p. 48). Therefore, it can be anticipated that the concept of corpus in CTS shall also present as one of its main attributes to allow for a wide range of configurations for data comparison.

All in all a corpus in CTS is not simply a large body of written text or spoken material as traditional definitions have often implied. It is defined more accurately as any open-ended body of machine-readable full texts analyzable automatically or semi-automatically, and sampled in a principled way in order to be maximally representative of the translation phenomenon under examination (cf. Baker, 1995).

Classification Criteria for Corpora in TS

The classification criteria presented here do not intend to be innovative as most of the points listed below have already been discussed in the field (see Baker, 1995; Zanettin, 2000; Kenny, 2001; Olohan, 2004). As a matter of fact, the discussion that follows aims to present a more flexible way of classifying the various types of corpora in translation research and pedagogy. To achieve that aim, I will revisit Baker's (1995) typology, attempting to discuss each of the selection criteria on which corpora are generally designed along with their attributes (p. 229) in order to classify the types of corpora being used in the descriptive and applied branches of TS. Figure 1 displays a snapshot of Baker's typology.

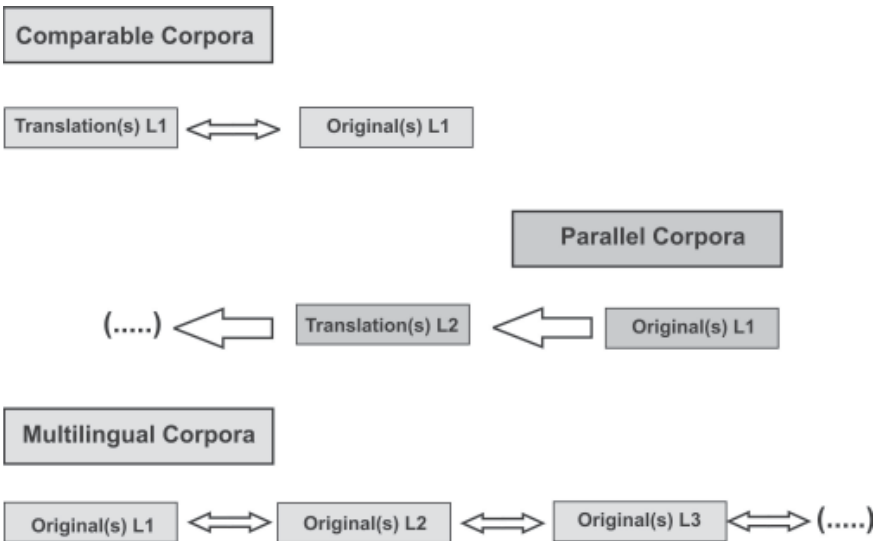

Fig. 1 – Baker's (1995) Typology for Corpora in TS

According to Baker's (1995) terminology, there are basically three main types of corpora for translation research and pedagogy:

(i) Comparable corpora – which "consist of two separate collections of texts in the same language: one corpus consists of original texts in the language in question and the other consists of translations in that language from a given source language or languages" (p. 234);

(ii) Parallel corpora – consist of "original, source language-texts in language A and their translated versions in language B" (p. 230);

(iii) Multilingual corpora – which are "sets of two or more monolingual corpora in different languages, built up either in the same or different institutions on the basis of similar design criteria" (p. 232).

In my view, Baker's tripartite classification can be re-arranged under only two main categories: *comparable* and *parallel*. This is due to the fact that the term *multilingual* does not have any contrastive feature that could make it distinctive from the other two types of corpora (see below). Moreover, such a classification does not seem to have caught on in the field, since the term *multilingual comparable corpora*

has often been used in replacement of *multilingual corpora* (see Teubert, 1996 and Kenny, 2001). Additionally, it is worth noting that in the first introductory book on CTS, namely *Introducing Corpora in Translation Studies* by Maeve Olohan (2004), the author focuses on *comparable* and *parallel* corpora only, which may indicate a change of perspective on the way the types of corpora are classified. Fig. 2 displays a proposal for the classification of corpora in TS.

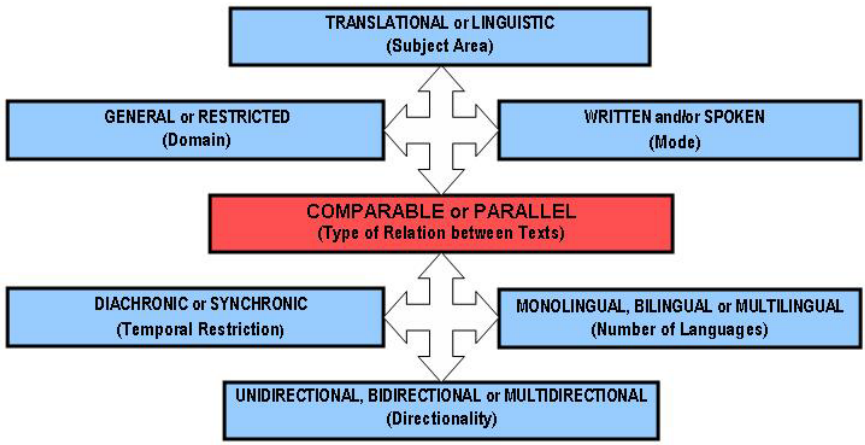

Fig. 2 – Classification Criteria for Corpora in TS

Type of Relation between Texts: Comparable or Parallel?

As far as I see it, it would be more profitable to focus on the terms "comparable" and "parallel" from the perspective of their contrastive features. These features have to do with the kind of relation that holds between the texts which comprise these corpora (cf. Teubert, 1996). In a comparable corpus, for instance, texts are put together on the basis of textual resemblance (i.e. texts are gathered based on their similarity of topic, text-type, communicative function, etc). In a parallel corpus, on the other hand, texts are grouped together on the basis of translational resemblance (i.e. one text could be taken as the translation of the other and vice versa).

Now if we return to Baker's (1995) tripartite classification of corpora (see above), it is possible to note that the term "multilingual" does not have any contrastive feature that could make it distinctive from the other two types of corpora. The term multilingual seems to acquire a contrastive feature only when compared to other corpora in

terms of language number (see below). In this sense, what Baker (ibid.) refers to as a multilingual corpus could be classified according to this new perspective as a *linguistic multilingual comparable corpus*. Linguistic because corpora of this kind are not primarily concerned with the study of translation (see below), multilingual because of the number of languages involved and comparable due to the fact that the texts comprising this kind of corpus are assembled on the basis of textual resemblance.

Subject Area: Linguistic or Translational?

This second criterion I want to suggest in the classification of corpora is related to the distinction between corpus-based studies designed for the study of languages and those built up with a view to investigating translation products and processes. For obvious reasons, I would like to call the former "linguistic" and the latter "translational". CTS is primarily concerned with the latter, but some scholars interested in translator education (see Schäffner, 1998; Zanettin, 1998; Stewart, 2000; and more recently Zanettin, Bernardini and Stewart, 2003) also make use of linguistic corpora as an aid for improving and developing trainee translators' language competence and translation specific skills. Despite their acknowledged importance, linguistic corpora do not provide realistic models for trainee translators because they only suggest procedures that should or could have been used in specific decision-making situations without actually showing what procedures are being used by professional translators. What is more, they seem less serviceable in providing theoretical discussions of how translated texts function in real communicative situations, since they only contrast two different language systems not at all linked by translation (cf. Shuttleworth and Cowie, 1998, p. 109).

Domain: General or Specialized?

The term domain refers to the area of language enquiry on which a corpus focuses. As far as domain is concerned, there are basically two main types of corpora: general and specialized (Baker, 1995, p. 229). As its own name suggests, a general corpus is broader in scope because it is built to study the language of translated material as a whole. The Translation English Corpus (TEC) hosted by Centre for Translation and Intercultural Studies (CTIS) at the University of Manchester, for instance, is a general translational corpus. It has been compiled to investigate the language of translated English. By contrast, a specialized corpus looks into the language of specific trans-

lated genres or text-types. As an illustration, Kenny's German-English Parallel Corpus of Literary Texts (GEPCOLT) is a specialized corpus which main focus is to investigate the language of translated literary texts from German into English.

Mode: Written and/or Spoken?

Mode has to do with the way the original contents of a text are delivered. For instance, a text transcribed from an audio or video source is considered "spoken" and a text scanned from a book and converted to electronic form is considered "written". According to Atkins et al. (1992), when the mode of delivery is not specified, it will be "written" by default. Moreover, they point out some problematic types of text that are sometimes written to be read (e.g. academic speeches) or spoken to be written (e.g. dialogues in a narrative text). It is important to note, though, that there are cases in which the texts of a corpus can consist of both written and spoken language. This is what happens with general linguistic corpora such as the British National Corpus (BNC), for instance, which currently has a 100 million word collection of samples of written and spoken language from a wide range of sources.

Temporal Restriction: Diachronic or Synchronic?

As to restrictions of time period, a corpus can be categorized as either *synchronic* – when it focuses on an object of study at one particular point in time – or *diachronic* – when it is concerned with the historical development of this object through time (Atkins et al., 1992, p. 6). Munday's (1998) analysis of translation shifts, for instance, is a typical example of a synchronic corpus-based study. His small-scale corpus, comprised of a short-story by Gabriel García Márques published in Spanish, focuses on the publication year of the English translation (i.e. 1993). Now if Munday (ibid.) had decided to include other English translations of the same short story published in different dates – aiming at examining the way these translations changed over time – the study would be of a diachronic kind.

Number of Languages: Monolingual, Bilingual or Multilingual?

In terms of language number, a corpus can be classified as *bilingual* or *multilingual* when more than three languages are involved. Another aspect related to the number of languages being represented in the corpus has to do with language varieties. In other words, if a corpus is described as bilingual, for instance, and the languages in-

volved are Portuguese and English, it seems important to specify the language variety of these two languages (e.g. European Portuguese vs. Brazilian Portuguese and British English vs. American English).

Directionality: Unidirectional, Bidirectional or Multidirectional?

Zanettin (2000) sees directionality as the translation direction of the texts which comprise the corpus. For instance, in a corpus comprised of texts originally written in L1 and their respective translations in L2 the direction of the translations functions in just one direction, so in such cases they are called unidirectional. Now if a corpus is made up of texts originally written in L1 and their translations in L2 plus originals in L2 and their translations in L1, it is called bidirectional. Multidirectional corpora are also possible, especially when more than two languages are involved and their translation direction is not centered on L1, but on the interaction among all the languages constituting the corpus (p. 106).

One last point worth making has to do with the combination of corpora. Depending on one's research purposes, a corpus can be combined with other corpora in order to achieve those particular purposes. Users of TEC, for instance, have to rely on the BNC in order to have their comparable corpora, which points out to the fact that greater standardization in terms of text encoding is necessary so that more and more corpora can be combined and their use spread all over the world.

Final Remarks

I hope that the classification criteria presented here may offer more flexible means of classifying the different types of corpora in the descriptive and applied branches of Translation Studies. Moreover, it is worth noting that the classification above cannot, in any real sense, exhaust completely the discussion of types of corpora. CTS is still in its infancy and owing to its technological symbiosis, it is rapidly and constantly developing. Yet as those changes are realized, the effect of this reformulation of Baker's (1995) terminology will hopefully be better resources for any translation researcher or translator educator interested in carrying out corpus-based work.

References

Atkins, Sue; Clear, Jeremy & Ostler, Nicholas. "Corpus Design Criteria". *Literary and Linguistic Computing, 7*(1). 1992. pp. 01-16.

Baker, Mona. "Corpora in Translation Studies. An Overview and Suggestions for Future Research". *Target, 7*(2). 1995. pp. 223-243.

Baker, Mona. "Corpus-based Translation Studies. The Challenges that Lie Ahead". In: Somers, Harold (Ed.). *Terminology, LSP and Translation*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1996. pp. 175-186.

Biber, Douglas. "Representativeness in Corpus Design". *Literary and Linguistic Computing, 8*(4). 1993. pp. 243-257.

Bowker, Lynne. "Towards a Corpus-based Approach to Terminography". *Terminology, 3*(1). 1996. pp. 27-52.

Concise Oxford Dictionary. [CD-ROM]. Oxford: Oxford University Press, 1996.

Kennedy, Graeme. *An Introduction to Corpus Linguistics*. London/New York: Longman, 1998.

Kenny, Dorothy. *Lexis and Creativity in Translation. A Corpus-based Study*. Manchester, UK: St Jerome, 2001.

Matthews, Peter. *The Concise Oxford Dictionary of Linguistics*. Oxford: Oxford University Press, 1997.

McEnery, Tony & Wilson, Andrew. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996.

Munday, Jeremy. "A Computer Assisted Approach to the Analysis of Shifts". *Meta, 43*(4). 1998. pp. 543-556.

Olohan, Maeve. *Introducing Corpora in Translation Studies*. London/New York: Routledge, 2004.

Pearson, Jennifer. *Terms in Context*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1998.

Schäffner, Christina. "Parallel Texts in Translation". In: Bowker, Lynne; Cronin, Michael; Kenny, Dorothy & Pearson, Jennifer (Eds.). *Unity in Diversity? Current Trends in Translation Studies*. Manchester, UK: St. Jerome, 1998.

Shuttleworth, Mark & Cowie, Moira. *Dictionary of Translation Studies*. Manchester, UK: St Jerome, 1997.

Sinclair, John. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

Vinay, Jean-Paul & Darbelnet, Jean. *Comparative Stylistics of French and English: A Methodology for Translation* (Juan C. Sager & Marie-Josée Hamel, Trans.). Amsterdam/Philadelphia: John Benjamins Publishing Company, 1995.

Zanettin, Federico. "Parallel Corpora in Translation Studies: Issues in Corpus Design and Analysis". In: Olohan, Maeve (Ed.), *Intercultural Faultlines. Research Models in Translation Studies I Textual and Cognitive Aspects*. Manchester, UK: St Jerome, 2000.