

Alma Mater Studiorum Università di Bologna  
Archivio istituzionale della ricerca

Corpora

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

Silvia Bernardini, Dorothy Kenny (2020). Corpora. London : Routledge [10.4324/9780203872062].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/713760> since: 2023-02-16

*Published:*

DOI: <http://doi.org/10.4324/9780203872062>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

**Bernardini, S. and Kenny, D. (2020). "Corpora". In Baker, M. and Saldanha, G. (eds). *Routledge Encyclopedia of Translation Studies*. London: Routledge. 110-115.**

The final published version is available online at: <https://doi.org/10.4324/9781315678627>

#### Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

*This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)*

***When citing, please refer to the published version.***

## **Corpora**

### **Silvia Bernardini and Dorothy Kenny**

A corpus is a collection of texts in electronic form that are the object of literary or linguistic study. Today, such collections normally include vast quantities of texts (hundreds of millions, or even billions, of words), that can be searched by means of fast and flexible dedicated software applications known as corpus query tools, corpus analysis tools, or concordancers. While most definitions stress the need for corpora to be assembled according to explicit design criteria and for specific purposes (Sinclair 2004), several corpora collected by crawling the Web in more or less opportunistic ways also exist. These are used both for finding out about general language use, and as a baseline against which to highlight features that are typical of specialized corpora.

Corpus linguists insist on the primacy of authentic data, as attested in texts, that is, instances of spoken, written or signed behaviour that have occurred “naturally, without the intervention of the linguist” (Stubbs 1996:4). Corpus linguists thus take an approach to the study of language that is consistent with the empiricism advocated in descriptive translation studies in the 1970s. At that time, scholars became particularly critical of the use of introspection in translation theory (Holmes 1988:101) and of approaches that viewed translations as idealized entities rather than observable facts (Toury 1980:79–81). While Toury conceded that isolated attempts had been made to describe and explain actual translations, he called for a whole new methodological apparatus that would make individual studies transparent and repeatable. It was Baker (1993) who saw the potential for corpus linguistics to provide such an apparatus, and her early work in the area (Baker 1993, 1995, 1996) launched what became known as corpus-based translation studies, or CTS. Researchers in CTS now pursue a range of agendas, drawing on a variety of corpus types and processing techniques.

### **Corpus creation and processing**

Best practice in corpus creation requires designers to make informed decisions on the types of language they wish to include in their corpora, and in which proportions. Design criteria crucially depend on the envisaged use of the corpus but have, in the past, centred on the idea that corpora should somehow be ‘representative’ of a particular type of language production and/or reception. The statistical notion of representativeness is however extremely difficult to apply to textual data, and many commentators now prefer to aim for a ‘balanced’ sample of the language in which they are interested (Leech 2006).

A general-purpose monolingual corpus might thus include both written language and transcribed spoken language, and, within each, samples of a variety of text types, dating from specific time periods. There may also be a trade-off between including fewer but more useful, full-length texts on the one hand, and more numerous but textually ‘compromised’ partial texts on the other (Baker 1995:229–30). The decision must be made on the basis of one’s research purpose: if, for example, the research is meant to shed light on a linguistic feature that is evenly distributed throughout most texts, then text samples might be adequate, whereas if the object of study is a rare feature, or one whose distribution is not stable, then complete texts might be preferred (Kenny 2014: 110–11). Once a suitable breakdown of text types, author profiles and other parameters has been decided upon, the

actual texts chosen for inclusion in a corpus can be selected randomly, or through more deliberate handpicking. Unless the texts are released with a copyleft licence, permission to include them in the corpus may have to be sought from copyright holders, particularly if the corpus is meant for sharing with other researchers or for sale (McEnery and Hardie 2012).

Depending on the intended uses of the corpus, various levels of contextual, structural and linguistic annotation are desirable. Metadata may be added that describe the genre of texts, name their authors, and specify a range of other information. Structural mark-up may be used to indicate the main divisions in a text, such as headings, paragraphs and sentences. Part-of-speech tagging may be employed to assign each word to a part-of-speech category. More complex forms of linguistic annotation include parsing, where words and phrases are assigned a syntactic function, and semantic tagging, which consists in assigning words to concepts such as 'people' or 'geographical names'. The level of mark-up that a corpus is subjected to will have implications for the kind of searches one is able to perform. In a raw text corpus, searches rely exclusively on sequences of characters and only basic statistics can be obtained, such as type-token ratio, average sentence length or lexical density. If a corpus is annotated with contextual and structural information, this can be used to restrict queries to specific subcorpora or to specific text portions. If linguistic annotation is available, queries can be made for parts of speech, syntactic functions or concepts. These different search parameters can be combined to deliver very sophisticated queries (Jones and Waller 2015), and employed as translation aids (Mikhailov and Cooper 2016). Zanettin (2012) provides an accessible introduction to the creation and annotation of corpora for translation research and practice.

### **Translation-oriented corpus typology**

Several scholars have proposed corpus typologies that are of particular relevance to translation studies (Zanettin 2012). At a high level of abstraction, corpora can be divided into those that contain texts in a single language – monolingual corpora – and those that contain texts in two or more languages – bilingual or multilingual corpora. Well-known monolingual corpora outside of translation studies are the BYU corpora (Davies 2011), which include the Corpus of Contemporary American English (COCA), a 560 million word corpus covering the years 1990-2017, and many others. In translation studies, the Translational English Corpus (Olohan 2004:59-60) is perhaps the best-known monolingual translational corpus. Corpora may also be characterized by the relationship that holds between their subcorpora, where these exist. Thus, a monolingual corpus for translation research may consist of two subcorpora; one translational, the other non-translational. If the two sets of texts cover the same genre(s) in roughly the same proportions, were published during the same time period and cover the same domains, then we can speak of a monolingual comparable corpus. Monolingual comparable corpora allow systematic investigations of how translated text differs from non-translated text in the same language, and thus are a vital resource in research that seeks to isolate characteristic features of translation (Laviosa 1998a, 1998b; Mauranen 2004; Jiménez-Crespo 2011, 2013; Delaere, De Sutter and Plevoets 2012; Redelinghuys and Kruger 2015). Likewise, the subcorpora in a bilingual corpus may be related through shared values for attributes such as genre, date and place of publication and domain, among other variables, and thus combine to form a bilingual comparable corpus. These are especially useful in the education of translators (Zanettin 1998; Looock and Lefebvre-Scodeller 2014) and for contrastive analyses (Altenberg and Granger 2002; Aijmer and Altenberg 2013; Cappelle and Looock 2013), though they are not without problems: as

with monolingual comparable corpora, it can be difficult to ensure comparability between the subcorpora (Bernardini and Zanettin 2004), and searching for “cross-linguistic equivalents” (Altenberg and Granger 2002:9) is not straightforward. Baker (1995:233) has also expressed reservations about their usefulness in theoretical translation studies, claiming that their use is based upon 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”.

The subcorpora in a bilingual (or multilingual) corpus may, on the other hand, be related through translation, that is, the corpus may contain texts in one language, alongside their translations into another language (or other languages). Such corpora are commonly known as parallel corpora. Parallel corpora are usually aligned (Ahrenberg 2015); that is, explicit links are provided between units of the source and target texts, usually at the sentence level. This enables bilingual concordancing, where a search for a word in one language returns all sentences containing that word, along with their aligned equivalent sentences in the other language. Parallel corpora exist for several language pairs and for groups of languages. The best known one is probably the multilingual Europarl corpus (Koehn 2005), which contains the proceedings of the European Parliament. Another is InterCorp, which contains manually aligned fiction texts translated from and into several languages as well as other automatically aligned subcorpora in various genres, including subtitles, journalistic and legal texts (Čermák and Rosen 2012).

A number of variations on the basic design are possible: a bilingual parallel corpus can be uni-directional or bi-directional, for instance. Given that bi-directional corpora such as the English-Norwegian Parallel Corpus (Johansson 1998) contain source texts in both languages, they can also be used as bilingual comparable corpora, provided that conditions of comparability obtain. Other parallel corpora may contain, on their target sides, two or more translations into the same language of the same source text, as is often the case in learner translation corpora (Castagnoli et al. 2001), or progressive drafts of the emerging target text (Utku 2004). Parallel corpora have been used widely, for instance in bilingual terminography and lexicography, for the extraction of translation equivalents (Bowker and Pearson 2002; Teubert 2002), as well as for research into translation shifts (Munday 1998; Cyrus 2009), lexical creativity (Kenny 2014) and translator style. Data-driven machine translation also depends on the availability and exploitation of vast parallel corpora.

### **Corpus-based translation studies**

Early work in CTS pursued the research agenda put forward in Baker’s seminal (1993) article, and investigated those recurrent features that make translation different from (or similar to) non-translated language production. Hypothesized typical features, also called universals of translation, include the reported tendency of translated texts to be more explicit, use more conventional grammar and lexis, and be somehow simpler than either their source texts or other texts in the target language. The universal status of the features in question has been questioned by several scholars (Chesterman 2004; Becher 2010), and many, including Olohan (2004), prefer the term ‘typical features of translation’ to ‘universals of translation’. Studies that investigate typical features of translation include Calzada Pérez’s (2017) comparisons of translated European Parliament speeches with original speeches from the same setting and with speeches from the British House of Commons, and Hareide’s (2017a,

2017b) particularly thorough attempts to verify the unique items hypothesis (Tirkkonen-Condit 2004) and the gravitational pull hypothesis (Halverson 2003).

The search for generalizations has evolved in two diverging directions. On the one hand, the suggestion has been put forward that some of the typical features observed in translation might be generalizable to other kinds of bilingual or contact language use, as observable in foreign language communication (Kranich 2014). Others have suggested that generalizations should be sought at an even higher level, that of constrained communication (Lanstyák and Heltai 2012). Thus, corpus based studies have been conducted to compare translated language to non-native (Kruger and Van Rooy 2016) and edited language (Kruger 2012), in an attempt to single out shared features as well as specificities of bilingual and monolingual constrained communication.

On the other hand, researchers in CTS have also begun to focus on the distinctive behaviour of individual translators. This development was triggered by an article by Baker (2000) where she proposes a methodology for investigating translator style. Several types of corpora have been used to zoom in on such a slippery research object. Ji (2010), Marco (2004), Winters (2007, 2009) and Wang and Li (2012) adopt a multi-target parallel structure consisting of a source text and two target texts. A more complex corpus is that used by Dirdal (2014), which consists of the fiction component of the *English-Norwegian Multiple Translation Corpus* (Johansson 2007). Dirdal investigates variation in the use of clause building and clause reduction in ten different commissioned translations of the same short story. Analysis techniques range from the basic statistics initially applied by Baker (2000) and others (Walder 2013; Huang and Chu 2014; Kajzer-Wietrzny 2013), to more complex stylometric techniques such as Delta, applied for example by Rybicki (2012) and Rybicki and Heydel (2013). Rybicki and Heydel (2013) look at the Polish translation of Virginia Woolf's *Night and Day*, which was carried out partly by one translator and partly by another, and the technique they adopt is able to identify the point where the second translator took over. Stylistic studies also combine quantitative methods with discourse analysis (Munday 2008) and other qualitative methods (Saldanha 2011a, 2011b, 2011c), particularly when attempting to contextualize findings beyond the text. As argued by Saldanha (2014), the choice of corpora and analysis techniques reflect different understandings of style in translation and, given the complexity of the object of study, a rigorous investigation requires triangulation based on a composite set of corpus resources, both monolingual and parallel (Huang 2015). Combining corpus resources is advocated also beyond stylistics research; it is the main principle underlying the methodology of corpus triangulation proposed by Malamatidou (2018), who demonstrates its application by examining the relationship between translation and language change.

While corpus design and exploration techniques have become increasingly sophisticated (Delaere et al 2012), corpus linguists are also aware that focusing on the corpus alone and adopting a purely descriptive, linguistic approach is not enough to account for the complexities of translation, and that increased contextualization and integration of analytical tools from other areas are needed (Olohan 2004). The integration of corpus and translation process research methods—such as keylogging, screen recording and eye tracking—has been particularly fruitful, and is exemplified in several of the chapters in Lykke Jakobsen and Mesa-Lao (2017). It is worth noting, however, that in some such cases, corpus data are elicited by researchers, and the approach thus departs somewhat from that adopted by corpus linguists who work only with authentic data, as described earlier. As well as

integrating multi-method designs, some recent research has also expanded the scope of corpus-based translation studies by integrating data from previously neglected sources, and produced under relatively new conditions. Jiménez-Crespo (2013), for example, compares the crowdsourced translation of a social media website with similar native websites in Spanish from the point of view of quality. Jiménez-Crespo (2015) uses data from this earlier study in a subsequent mixed-methods exploration of explicitation given two different production models (translation from scratch vs. translation based on pre-existing options such as those offered by a translation memory tool) and “a translation modality (web localization) and genre (social networking sites) that did not exist when scholars in TS set off to systematically research general tendencies of translation” (Jiménez-Crespo 2015:260). The Genealogies of Knowledge project based at the University of Manchester in the UK has further extended the boundaries of CTS by developing a methodology to address the relationship between translation and the construction, negotiation and dissemination of knowledge, enriching the corpus-linguistic perspective with a socio-political angle (Jones, in press; Baker and Jones, forthcoming (a), forthcoming (b); Baker, forthcoming).

The use of corpora in applied translation studies is another dynamic research area. Studies such as López-Rodríguez (2016), Kübler and Volanschi (2012), Tagnin and Teixeira (2012) and Ferraresi et al. (2010) are among many that demonstrate how corpora can be valuable aids in translation pedagogy, specialized translation, applied terminology and bilingual lexicography.

Going beyond written translation, corpus linguistics has been applied in audiovisual and interpreting research. Both types of text present considerable technical difficulties in terms of corpus compilation and analysis. Compiling requires transcription and often audio or video alignment, as well as manual annotation of contextual, linguistic and paralinguistic features such as pauses and dysfluencies. Bernardini et al. (2018) provide an operational, step-by-step account of corpus building procedures adopted to build interpreting and intermodal corpora, tapping into European Parliament data. However, even when accessing existing resources, corpus-based research in audiovisual translation raises methodological issues that have to do with “the complex semiotic fabric of audiovisual texts, their hybrid nature and multiple codes” (Baños et al. 2013a:483).

Despite technical and methodological challenges, corpus-based audiovisual and interpreting research has been productive. Substantial work has been carried out on interpreting between English and Chinese (Wang 2012; Hu and Tao 2013), confirming the strong interest in corpus studies of Chinese spurred by the pioneering work of Richard Xiao (Xiao 2010; Xiao and Hu 2015). Instances of dialogue interpreting, both authentic and simulated, are included in the corpora described by Pérez González (2006), Baraldi and Gavioli (2012) and Angermeyer et al. (2012). The articles collected in Baños et al. (2013b) attest to the richness of corpus-based AVT studies, covering audio-description (Jiménez Hurtado and Soler Gallego 2013) and multimodal analysis of humour in dubbed TV programmes (Balirano 2013) among other topics.

Finally, efforts to create sign language corpora began in Australia in 2004 (Fenlon et al. 2015). These corpora consist of spontaneous or, more usually, elicited video recordings that are enriched with various kinds of linguistic annotation and are accompanied by a translation into the local written language, giving them the status of bilingual, inter-modal, parallel corpora. Fenlon et al. (2015) discuss methodological issues in creating sign language corpora,

and Meurant et al. (2016) outline the process of creating the French Belgian Sign Language (LSFB) Corpus, and the policy governing its translation into written French.

As the variety of corpora continues to grow, as do the number of languages and language pairs covered, and given increased ease of access to corpora, in particular access to open-source software such as that provided by the Genealogies of Knowledge project (Jones, in press; Baker and Jones, forthcoming (b)), CTS looks set to continue offering diverse and rich contributions to translation studies. Consensus in the field, in corpus linguistics in general, and in corpus-based cognitive linguistics in particular (Arppe et al. 2010; Biber 2012; De Sutter et al. 2017; Halverson 2018), suggests that the way forward lies in the continued pursuit of rich contextualization of corpus data, the creation of closer bonds between product and process, the design of statistically sophisticated studies, and the integration of empirical and theoretical perspectives.

### **Further reading**

Olohan, M. (2004) *Introducing Corpora in Translation Studies*, London & New York: Routledge.

Sums up much of the earlier corpus-based research into features of translation and presents several case studies which engage with different sources of quantitative and qualitative data and approaches.

Zanettin, F. (2012) *Translation-driven Corpora*, Manchester: St. Jerome.

A practical guide for those who want to undertake corpus research in translation studies, with substantial exemplification of how to build and consult monolingual and multilingual corpora.

De Sutter, G., M.-A. Lefer and I. Delaere (eds) (2017) *Empirical Translation Studies*, Berlin: De Gruyter Mouton.

Illustrates the current trend towards greater methodological sophistication in the empirical study of translation, including the use of ingenious corpus designs and multivariate statistical methods.

Malamatidou, S. (2018) *Corpus Triangulation: Combining Data and Methods in Corpus-Based Translation Studies*. London & New York: Routledge.

Makes the case for systematic use of triangulation in corpus-based translation studies and presents an alternative typology of corpora in translation studies.